frontiers | Research Topics

New perspectives on the role of sensory feedback in speech production

Edited by
John Houde, Xing Tian, Jeffery A. Jones, Douglas M. Shiller and Lucie Menard

Published in
Frontiers in Human Neuroscience

FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-8325-2515-9
DOI 10.3389/978-2-8325-2515-9

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

New perspectives on the role of sensory feedback in speech production

Topic editors

John Houde — University of California, San Francisco, United States
Xing Tian — New York University Shanghai, China
Jeffery A. Jones — Wilfrid Laurier University, Canada
Douglas M. Shiller — Montreal University, Canada
Lucie Menard — Université du Québec à Montréal, Canada

Citation

Houde, J., Tian, X., Jones, J. A., Shiller, D. M., Menard, L., eds. (2023). *New perspectives on the role of sensory feedback in speech production*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-2515-9

Table of contents

frontiers | Frontiers in Human Neuroscience

TYPE Editorial
PUBLISHED 09 May 2023
DOI 10.3389/fnhum.2023.1189751

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Ludo Max,
University of Washington, United States

*CORRESPONDENCE
John F. Houde
✉ jfhoude@ucsf.edu

RECEIVED 20 March 2023
ACCEPTED 31 March 2023
PUBLISHED 09 May 2023

CITATION
Houde JF, Ménard L, Jones JA, Shiller DM and Tian X (2023) Editorial: New perspectives on the role of sensory feedback in speech production. *Front. Hum. Neurosci.* 17:1189751. doi: 10.3389/fnhum.2023.1189751

COPYRIGHT
© 2023 Houde, Ménard, Jones, Shiller and Tian. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Editorial: New perspectives on the role of sensory feedback in speech production

John F. Houde[1]*, Lucie Ménard[2], Jeffery A. Jones[3], Douglas M. Shiller[4] and Xing Tian[5,6,7]

[1]Department of Otolaryngology – Head and Neck Surgery, University of California San Francisco, San Francisco, CA, United States, [2]Département de Linguistique, Université du Québec à Montréal, Montréal, QC, Canada, [3]Department of Psychology, Wilfrid Laurier University, Waterloo, ON, Canada, [4]École d'orthophonie et d'audiologie, Faculté de médecine, Université de Montréal, Montréal, QC, Canada, [5]Department of Neural and Cognitive Sciences, NYU Shanghai, Shanghai, China, [6]NYU-ECNU Institute of Brain and Cognitive Science at NYU Shanghai, Shanghai, China, [7]Shanghai Key Laboratory of Brain Functional Genomics (Ministry of Education), School of Psychology and Cognitive Science, East China Normal University, Shanghai, China

KEYWORDS

speech production, speech perception, speech motor control, speech feedback monitoring, speech development, sensory feedback

Editorial on the Research Topic
New perspectives on the role of sensory feedback in speech production

Studies on the role of sensory feedback in speech production have revealed much about sensorimotor integration mechanisms in speech-motor control. These studies have a rich history dating back over a century, starting with Lombard's (1911) work on the impact of noise on speech loudness. Recent advancements in technology and techniques have greatly accelerated the progress of this field. In this Special Topic, our aim was to bring together a collection of cutting-edge studies that reflect the exciting new directions and breakthroughs in this area of research, particularly over the past few years.

The study by Oschkinat et al. adds greatly to our understanding of the role of sensory feedback in the timing of speech production. They used focal distortions of the duration of consonant-vowel-consonant syllables in speakers' auditory feedback and showed that speakers adapted to distortions of vowel duration but only adapted to distortions in consonant duration when the consonant was in the coda position. Additionally, Oschkinat et al. found that high sensitivity in rhythm and interval perception, along with high variability in rhythm and interval production, was correlated with the degree of adaptation observed in speakers. These findings offer valuable insights into the mechanisms used by the auditory system to monitor and adjust speech timing, which may have implications for the development of speech rehabilitation techniques.

The role of feedback in speech timing is also addressed in a new synthesis by Tilsen. Tilsen proposes a framework consisting of a palette of "time responders" (TiRs) that represent the ways in which feedback (both internal and external) could control the timing of utterance production. TiRs can be combined to govern gestural timing within utterances and utterance sequencing. They also form the basis of the hypothesis that speakers change their speech rate by changing how they attend to sensory feedback as they speak.

Speech scientists have long worked to understand speech variability and stability. The study by Wang and Max demonstrated that speakers actively control their speech variability by exposing them to auditory feedback alterations that either magnified or attenuated their perceived errors in producing vowels. Attenuation caused speakers to gradually increase their variability over repeated productions. Nault et al. investigated the effect of feedback variability on speech stability, revealing that speakers adapt only to consistent changes in their auditory feedback. Their work suggests that the consistency of feedback facilitates the stability of speech sensorimotor control.

Advances in neuroimaging have also greatly facilitated our understanding of how sensory feedback is processed during speaking. Recent research has demonstrated how this process is compromised in dysfunctional conditions, such as stuttering. The study by Garnett et al. is a noteworthy example, offering further evidence of the relationship between stuttering and abnormal auditory feedback processing. Additionally, this study suggests that stuttering may be linked to disruptions in speech sensorimotor function by the default mode network.

Some of the studies included in this Research Topic focus on speech perception, which sensory feedback mechanisms likely depend on. Goldenberg et al. provide further support for the findings of Gick and Derrick (2009), showing that air puffs, even on the hands, can influence the perception of ambiguous consonant sounds toward voiceless consonants. Johnson et al. found a correlation between the right-hemisphere auditory cortical speech responses and the likelihood of study participants experiencing auditory hallucinations.

One key question about sensory feedback is how its role in speaking evolves during development. To address this question, the article by Coughler et al. provides a comprehensive review of pediatric responses to altered auditory feedback. The studies they review show that while children have prolonged response times to auditory feedback perturbations, by the age of four they display sensorimotor adaptation that is qualitatively similar to adults. However, it is noted that the limited number of studies on this subject makes it difficult to draw definitive conclusions, underlining the need to explore more fully the plasticity of sensory feedback control of speaking across the lifespan.

Recently, researchers have developed various new models that help to explain the role of feedback in the development of speech production. One such model, proposed by Kröger et al., provides a comprehensive account of speech production by postulating an evolving role for sensory feedback during development. In this model, sensory feedback initially plays a crucial role in an undirected babbling process, creating internalized sensory-motor relationships. These relationships are then used when children attempt to imitate words produced by others. During this process, they initially select motor states that were previously associated with the sounds of the target utterance and then vary them until they receive feedback that their speech has been understood.

Another model, proposed by Davis and Redford, describes a dual-lexicon model of speech-motor planning that evolves continuously with experience from childhood through adulthood. According to their model, words have perceptual representations (exemplars) that evolve as the speaker hears the speech of others as well as auditory feedback of their own word productions. In addition, words have motor representations (silhouettes) that evolve as the speaker plans word productions. This process balances matching the target perceptual exemplar with articulatory ease and prior motor habits.

The final theme covered in this Research Topic is determining how sensory feedback processing varies across speakers. Kearney et al. propose a unique approach in which they fit the timecourse of a speaker's response to auditory pitch feedback perturbations to a simplified version of the DIVA model. The authors find that pitch perturbation responses vary across speakers but remain consistent within each individual, creating a distinct "fingerprint" of their speech motor system. If such fingerprints can be expressed in interpretable parameters, the authors suggest that the effects of disease states on the pitch perturbation reflex can be similarly expressed as meaningful changes in these interpretable parameters.

This marks the end of a brief overview of the papers on this Research Topic. It offers a general idea of the topics covered but may generate further questions. We encourage you to delve deeper by reading the individual papers, which offer a more comprehensive examination of this fascinating area of research.

Author contributions

JH wrote the initial draft. JH, LM, JJ, DS, and XT contributed to the writing/editing of manuscript. All authors contributed to the article and approved the submitted version.

Funding

JH was supported by NIH grants P50DC019900, R01NS100440, R01DC017091, and R01DC019167. Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery grants support LM (RGPIN-2020-05439) and DS (RGPIN-2019-05080). XT was supported by grants NSFC 32071099 and 32271101, NSF of Shanghai 20ZR1472100, and by the Program of Introducing Talents of Discipline to Universities, Base B16018.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Gick, B., and Derrick, D. (2009). Aero-tactile integration in speech perception. *Nature* 462, 502–504. doi: 10.1038/nature08572

Lombard, E. (1911). Le signe de l'elevation de la voix, Annals Maladiers Oreille. *Larynx, Nez Pharynx.* 37, 101–119.

frontiers | Frontiers in Human Neuroscience

HYPOTHESIS AND THEORY
published: 12 May 2022
doi: 10.3389/fnhum.2022.844529

On the Emergence of Phonological Knowledge and on Motor Planning and Motor Programming in a Developmental Model of Speech Production

Bernd J. Kröger[1], Trevor Bekolay[2] and Mengxue Cao[3]*

[1] Department of Phoniatrics, Pedaudiology, and Communication Disorders, Medical Faculty, RWTH Aachen University, Aachen, Germany, [2] Applied Brain Research, Waterloo, ON, Canada, [3] School of Chinese Language and Literature, Beijing Normal University, Beijing, China

OPEN ACCESS

Edited by:
John Houde,
University of California,
San Francisco, United States

Reviewed by:
Sam Tilsen,
Cornell University, United States
Jason W. Bohland,
University of Pittsburgh, United States

****Correspondence:***
Bernd J. Kröger
bkroeger@ukaachen.de

Specialty section:
This article was submitted to Speech and Language, a section of the journal Frontiers in Human Neuroscience

Received: *28 December 2021*
Accepted: *12 April 2022*
Published: *12 May 2022*

Citation:
Kröger BJ, Bekolay T and Cao M (2022) On the Emergence of Phonological Knowledge and on Motor Planning and Motor Programming in a Developmental Model of Speech Production. Front. Hum. Neurosci. 16:844529. doi: 10.3389/fnhum.2022.844529

A broad sketch for a model of speech production is outlined which describes developmental aspects of its cognitive-linguistic and sensorimotor components. A description of the emergence of phonological knowledge is a central point in our model sketch. It will be shown that the phonological form level emerges during speech acquisition and becomes an important representation at the interface between cognitive-linguistic and sensorimotor processes. Motor planning as well as motor programming are defined as separate processes in our model sketch and it will be shown that both processes revert to the phonological information. Two computational simulation experiments based on quantitative implementations (simulation models) are undertaken to show proof of principle of key ideas of the model sketch: (i) the emergence of phonological information over developmental stages, (ii) the adaptation process for generating new motor programs, and (iii) the importance of various forms of phonological representation in that process. Based on the ideas developed within our sketch of a production model and its quantitative spell-out within the simulation models, motor planning can be defined here as the process of identifying a succession of executable chunks from a currently activated phoneme sequence and of coding them as raw gesture scores. Motor programming can be defined as the process of building up the complete set of motor commands by specifying all gestures in detail (fully specified gesture score including temporal relations). This full specification of gesture scores is achieved in our model by adapting motor information from phonologically similar syllables (adapting approach) or by assembling motor programs from sub-syllabic units (assembling approach).

Keywords: motor planning, motor programming, speech production, developmental model, phonological knowledge, sensorimotor system, cognitive-linguistic system

THEORETICAL BACKGROUND

The Models of Speech Production and Speech Perception Influencing Our Model Sketch

The process of speech production can be subdivided in concept preparation, lexical selection, morphological and phonological encoding, phonetic encoding, and articulation (Levelt et al., 1999). In a word production task, concept preparation is the activation of a lexical concept, followed by selecting its lemma and subsequently by retrieving its phonological form. It is emphasized by Levelt et al. (1999) that morphemes and not syllables are stored in the mental lexicon. Thus, lexical processing is followed by syllabification. Subsequently, syllables are encoded phonetically by specifying a gestural score (ibid., see Browman and Goldstein, 1992 for defining gestural scores for lexical units like monosyllabic words) and thus by specifying basic control units for the articulatory execution of the syllable under production. In parallel to the mental lexicon as the central higher-level knowledge repository, Levelt and Wheeldon (1994) and Levelt et al. (1999) postulate a mental syllabary as a storage for highly overlearned gestural patterns. These "ready-made" gestural scores or patterns are assumed to be stored within the mental syllabary of a speaker for all frequently used syllables of the speaker, and it is assumed that these patterns can be directly accessed and executed by the articulatory system.

The mental syllabary as introduced by Levelt and Wheeldon (1994), Levelt et al. (1999), and Cholin et al. (2006) is a repository for motor programs. While the motor programs of low-frequency syllables of a language are assumed to be calculated or constructed "on-line," the mental syllabary is hypothesized to provide motor programs as "pre-compiled gestural scores" for high-frequency syllables. Moreover, it is assumed that the storage of motor programs does not overload the mental or neural capacity of the brain because only about 500 syllables can be labeled as high-frequency syllables for example in English, Dutch, or German. In these languages, 500 syllables make up only 5% of the entire syllable inventory, but these 500 syllables are sufficient for producing about 80% of all utterances in these languages (Schiller et al., 1996).

Beside storing execution-related neural representations like "motor programs" it can be assumed that auditory as well as somatosensory forms are stored in the mental syllabary as well (Kröger et al., 2019). This assumption is in accordance with the DIVA model of speech production introduced by Guenther (2006) and Guenther et al. (2006). Here, motor representations (motor commands) are stored for speech items in parallel to their sensory target representations (auditory and somatosensory states) in order to allow a sensory driven control (feedback control) during feedforward execution of a speech item. Thus, Guenther's DIVA model (Directions Into Velocities of Articulators; Guenther, 2006; Guenther et al., 2006; Guenther and Vladusich, 2012; Kearney and Guenther, 2019) differentiates a feedforward and a feedback control subsystem. Production starts with the activation of a speech item in the "speech sound map," which subsequently activates a set of motor commands passing the feedforward control system, which then activates a target in the motor map, here called "articulatory velocity and position map." Activation patterns in this map directly result in articulator movements. In parallel the activation of a speech item in the speech sound map leads to a co-activation of an auditory and somatosensory target state for that speech item. During the production process, the activated sensory target states are compared with its sensory feedback states. In case of divergence, feedback commands (i) for on-line correcting the current production or (ii) for a later offline correction are generated and forwarded to the motor map for modifying execution. Thus, motor commands and the associated sensory target states can be updated with each production trial if necessary. Bohland et al. (2010, p. 1508) interpret the speech sound map as compatible with Levelt's mental syllabary. It should be noted that the DIVA model undergoes (i) a babbling training process which provides continuous mappings between sensory and motor states and later (ii) an imitation training process in order to acquire motor representations for specific speech items like words or short phrases which are stored in the speech sound map. Imitation learning depends on knowledge concerning sensory-to-motor relations in order to generate first motor representations (first motor commands) for the speech item under imitation as well as for calculating the direction of further alterations of the motor representation of a speech item in order to approximate its acoustic target.

In parallel to the syllabification process as described by Levelt et al. (1999), Bohland et al. (2010), Guenther (2016), and Miller and Guenther (2021) propose a process for the division of the phonological sound sequence in executable speech items (chunks), for which sensorimotor programs already exist. This process is implemented in the GODIVA-model (Gradient Order DIVA model, Bohland et al., 2010) which differentiates a planning loop and a motor loop. The planning loop comprises a phonological content buffer and a sequential structure or structure frame buffer. The motor loop comprises the (speech) initiation map and the speech sound map. While the motor loop directly initiates the chain of sensorimotor programs (executable gesture scores) at the level of the speech sound map, the planning loop parses the incoming phonological sound sequence with respect to these executable chunks and selects chunks for later initiation by the motor loop. By activating potential syllabic chunks, which fit parts of the current sound chain, chunks of phonological sound sequences are selected and executed. Bohland et al. (2010) describe this process as an interaction or interfacing of selected phonological codes with "an elaborated speech sound map" to select best matching sensorimotor programs for execution (ibid., 1509). Here the speech sound map is interpreted as a neural buffer from which sensorimotor programs for high-frequency syllables can be initiated directly in full, whereas the sensorimotor programs of infrequent syllables must be assembled from smaller, e.g., phoneme-sized units (ibid., p. 1509 and see dual route approach, Varley and Whiteside, 2001)

before they can be initiated and executed. The assembly process is later concretized by Bohland et al. (2010) by stating, that a phonological word to be produced can be effectively "spelled out" during production using motor programs for the individual phonemes (ibid., p. 1512). Thus, motor plans are available for whole syllables on the one hand but on the other hand motor plans of (new) syllables can be generated "using a sequence of smaller stored programs corresponding to the syllables' individual phonemes" (ibid 1521). Thus, GODIVA stores motor plans of frequent syllables as well as motor plans for sub-syllabic phoneme-sized units within the speech sound map.

The DIVA model already stresses the importance of somatosensory and auditory feedback in speech production. While somatosensory feedback always stems from self-perception, auditory perception is self-perception as well as perception of other's speech (auditory input from communication partners). The process of auditory speech perception can be subdivided in two routes, an auditory-conceptual (ventral) and an auditory-motor (dorsal) route (Hickok and Poeppel, 2007, 2016). The dorsal route activates appropriate motor representations and somatosensory representations if an auditory speech signal is processed (cf. sensorimotor integration; Hickok et al., 2011). The functional processing steps in the speech perception and speech processing model introduced by Hickok and Poeppel (2007, 2016) are spectro-temporal acoustic signal analysis followed by phonological processing. Subsequently the perceptual pathway separates in the dorsal stream which activates the motor network via a sensorimotor interface and in the ventral stream activating the lexical and combinatorial (conceptual) network.

One of the goals of this paper is to differentiate motor planning and motor programming as well as to define functional aspects of motor planning and motor programming. Our approach is based on already published concepts. (i) In the GODIVA model a phonological chain processing or selection process is separated from motor program initiation and execution (Bohland et al., 2010, p. 1512). (ii) Riecker et al. (2005) separate a cerebral motor preparation and a motor execution loop for speech production based on fMRI experiments. Because the task here was a simple syllable repetition task, preparation here comprises activation of motor programs but not motor planning processes. (iii) A four-level model focusing on the differentiation of planning and programming is introduced by van der Merwe (2021). Here a differentiation of linguistic symbolic planning, motor planning, motor programming and execution is postulated. While linguistic planning activates a phonemic representation (lexical and grammatical processing and syllabification), the motor planning module takes phonological code as input and "assigned properties amenable to a motor code" (ibid., p. 404). A set of motor commands is activated as output of the motor planning module, mainly specifying phonological-phonetic segmental features (ibid., p. 409). The motor programming module now uses motor plan information as input and outputs fully specified spatiotemporal articulatory movement information in form of muscle-specific motor programs. Motor programs here can be defined for whole syllables but as well for sub-syllabic units like segments or gestures.

Early Phases of Speech Acquisition and Models of Speech Learning

The newborn starts to produce speech-like vocalic sounds, also called proto-vowels, at the age of about 3 months. It produces first canonical babbling patterns, also called proto-syllables or proto-CV patterns comprising proto-consonants (proto-C) and proto-vowels (proto-V), at the age of about 7 months. Language specific syllable productions start at about 10 months and first words are produced at about 12 months (Kuhl, 2004). The well-known fact that perception precedes production is underpinned by the fact that speech-specific phonetic contrasts can already be discriminated directly after birth and language specific perception of vowels already starts with 6 months. Recognition of language specific sound combination starts with 9 months (ibid.). By 18 months of age, 75% of typically developing children understand about 150 words and can successfully produce 50 words in case of American English (Kuhl, 2004, p. 834, citing Fenson et al., 1993). Moreover, the role of social interaction as occurring for example in the case of joint attention to an object is an important vehicle for word learning (e.g., Lytle and Kuhl, 2017).

Thus, the transition from newborn's first vocalizations like crying, like production of vegetative sounds, and like first non-cry phonations toward the production of speech-like vowels including speech-like phonation (i.e., proto-vowels) and the transition from gooing and marginal babbling, both consisting of primitive tongue and lip movements toward canonical babbling occurs within the first 6–9 months of lifetime (Oller, 2000; Buder et al., 2013). Canonical babbling comprises the production of proto-syllables consisting of already well-formed consonantal closures and vocalic openings accompanied by speech-like phonation. It has been shown by means of computer simulations how canonical babbling emerges from earlier babbling stages and from pre-speech vocalizations by using reinforcement learning (reward-modulated learning, see Warlaumont and Finnegan, 2016). Here a reward is given if a new vocalization produced by the infant (by the model) is acoustically more salient than vocalizations produced earlier and productions which are accompanied by a caretaker's reward are stored and reproduced more frequently. These simulations indicated that pure vocalic sounds are auditorily less salient than speech sounds which include vocal tract closures and releases of these closures, here labeled as "syllabic sounds." The simulation experiments indicate that the frequency of canonical babbling (i.e., the frequency of auditory salient events) increases during ongoing reinforcement learning.

A further model of speech learning comprising the babbling and imitation phase is introduced by Kröger et al. (2009), Kröger and Cao (2015), and Kröger et al. (2019). Here, two self-organizing neural maps, i.e., a phonetic and a semantic map form the center of the speech processing neural network. The semantic map realizes the center of the cognitive-linguistic model part and the phonetic map realizes the center of the sensorimotor or

phonetic model part. Babbling starts with a set of proto-syllables (pre-linguistic items) and proceeds toward learning of language specific sets of V-, CV-, VC-, and CCV-syllables. This babbling training leads to the development of the phonetic self-organizing map (SOM) which contains basic auditory-to-motor knowledge in order to enable imitation (Kröger et al., 2009). Imitation training leads to an advancement of this map. After imitation training the phonetic map is able to activate motor and sensory states for all syllables, trained so far. In parallel, imitation training leads to a buildup of the semantic SOM in the cognitive-linguistic part of the model (Cao et al., 2014; Kröger and Cao, 2015). Simulation experiments were carried out for learning or training a model-language comprising of about 70 monosyllabic words. After learning, word production can be simulated by activating a word node (a model neuron) within the semantic map which co-activates sensorimotor nodes within the mental syllabary and thus co-activates motor and sensory states for each selected word.

In this approach, the main result of babbling training is the association of auditory, somatosensory, and motor states of proto-syllables within the self-organizing phonetic map. In addition, an ordering of proto-syllables appears with respect to phonetic features like vocalic high-low, front-back or consonantal manner and place of articulation. The main result of imitation training is that these proto-syllabic motor and sensory states represented in the phonetic map during babbling training now are more and more shaped with respect to specific syllable realizations of the target language. Moreover, imitation training leads to an association of words with those syllables which are already represented by the phonetic map. This allows the extraction of phonological features and of phonological knowledge from the ordering of syllables within the phonetic map because this ordering which has already been established during babbling will remain and will be expanded during imitation training (Kröger and Cao, 2015; Kröger et al., 2019).

A further simulation approach for speech learning using SOMs has been proposed by Li et al. (2004). In contrast to the models described above this approach does not include acoustic or motor information. Here, a segmental feature description of speech items is used as phonological input information and two different semantic feature descriptions are used as semantic input representations. This approach models the early lexical development up to a lexicon size of about 500 words. The model starts with imitation of speech items. In this approach, the learner (the model) already has available phonological knowledge including the phoneme repertoire of the target language. On this basis the model is capable to simulate learning effects occurring during lexical development like lexical confusion effects occurring in early vocabulary learning as well as age-of-acquisition effects.

The Emergence of Phonological Representations

The models described so far differ in introducing a level of phonological representation. Because a phonological representation is language-specific this representation emerges during speech acquisition. During the imitation phase first phonetic features and broad categorizations like labial, apical vs. dorsal place of articulation, like voiced vs. voiceless and like nasal vs. oral sound production result from differentiating babbling items. Moreover, proto-vocalic productions with palatal, velar and pharyngeal narrow passages lead to phonetic vowel categories like [i], [a], and [u], and thus to phonetic features like high-low front-back. These broad categorizations and its resulting phonetic features can be interpreted as precursors of language-specific phoneme sets and phonological features. These initial processes are followed by a complex process of tuning the perceptual categories and the articulation of speech sounds in a language specific direction up to an age of 6 years (Gervain and Mehler, 2010; Redford, 2019). As an example, in case of English and Dutch, most language specific vowels are learned at about 3 years of age, and most consonants already at about 4 years of age, except some fricatives. Complex consonant clusters develop between 4 and 6 years of age (Priester et al., 2011). But typical patterns of articulatory alterations or simplifications like gliding, stopping, epenthesis, cluster simplification can still be observed until school-age years even in normally developing children (Redford, 2019, p. 2952; citing Stoel-Gammon and Dunn, 1985, pp. 43–46). Thus, it can be assumed that phonological knowledge like the notion of phonemes as well as of distinctive features emerges over the entire time span of speech acquisition (emergentist model, e.g., Menn and Vihman, 2011, continuity hypothesis, e.g., Fikkert, 2007).

Segmental Versus Gestural Approaches

Beside developmental approaches supporting segmental concepts and introducing a phonological level of representations, Redford (2019) suggests a developmental approach based on holistic motoric representations or action schemas for the representation of words. Here, four major developmental milestones are postulated: (1) A perceptual-motor map for associating perceptual and motor forms of syllable-sized speech items already develops during the pre-speech period and continuously develops during speech learning. (2) During imitation, perceptual word forms (referential adult productions) are the starting point for word learning. Action schemas are now influenced and refined by language-specific imitation of syllables. At about 12 months of age a stable perceptual lexicon of about 100 words is established. Motor routines or action schemas now are associated with first words using the already existing perceptual-motor map. (3) Perceptually based control becomes more and more important at about 18 months of age. While productions are motorically constrained during the babbling phase, perception now forces articulation to widen and to refine the movement repertoire dramatically. (4) While the third phase marks the onset of perceptual control and while speech learning is mainly communication-driven in this third developmental stage the fourth stage emphasizes self-perception. Redford (2019) states that "speech production does not become adultlike until children begin to externally monitor their own speech and consciously recognize its divergence from (chosen) adult norms" (ibid. p. 2956). Thus, the reward in reinforcement learning during imitation now switches from external reward

given by communication partners toward self-judgment of the phonetic quality of word production.

Moreover Redford (2019) separates information processing approaches and ecological dynamics approaches. In the first category phonological representations mediate between perception and production. Here the sequencing of discrete elements like phonemes plays a central role and discrete steps are needed to translate discrete symbolic representations into action plans (e.g., Levelt et al., 1999). The second category represents the non-segmental concepts like that of Articulatory Phonology (Browman and Goldstein, 1992; Goldstein et al., 2006). Here, the segmental or phonemic level is avoided by introducing gestures as an action unit on the one hand and as a distinctive phonological unit on the other hand. Moreover, this approach allows a direct linking of lexical forms to action forms (for a definition of "action units" see the task dynamics concept as introduced by Saltzman and Munhall, 1989). Gestures (or actions) are dynamically defined target-directed movement units, and the temporal coordination of gestures is quantified by using a concept of phasing which is based on intrinsic time scales (Goldstein et al., 2006). The minimal unit of speech production (molecule) described in the framework of Articulatory Phonology is the syllable or the one-syllabic word while gestures are seen here as minimal production units (atoms).

The model described in this paper assumes the neurobiological reality of gestures as well as of phonemes and distinctive features as units of speech processing (production and perception). While gestures appear to be the adequate units for describing speech during early phases of speech learning (during babbling and early phases of imitation) as well as later during adult speech production, it is assumed in our approach that an intuitive awareness of distinctive features, of phonemes and of syllable structures like CV, CVC, or CCV establishes during the time span of speech acquisition (Grunwell and Yavas, 1988; Levelt and van de Vijver, 2004). Thus, we use the concept of gestures, gesture scores and of intrinsic timing of gestures mainly as a concept for describing proto-syllables as well as language-specific syllables. But during imitation training gestures can be defined more and more by distinctive features. Thus, a glottal opening/closing gesture for example represents the feature unvoiced/voiced; a labial/apical/dorsal closing gesture represents the feature "place of articulation." A closing/near-closing gesture represents different values for the feature "manner of articulation" etc. (Kröger and Birkholz, 2007). Beside this phonological aspect of gestures, the motor aspect of gestures and gesture scores can be implemented by introducing syllabic neural oscillators for defining the temporal coordination of gestures and by introducing gesture neural oscillators for defining the spatio-temporal aspects for the realization of each gesture within a gesture score (Kröger et al., 2021).

Goals of This Paper

It is the goal of this paper to formulate a sketch for a model of speech production which comprises the cognitive-linguistic as well as the sensorimotor part of speech production, which includes developmental aspects of speech production, and which emphasizes the emergence of segmental or gestural phonological representations as an important part of developmental processes (i.e., of speech acquisition). Our model sketch can be interpreted as a theory of speech production and speech acquisition and parts of our model sketch are underpinned by quantitative computer simulations. (i) A conventional connectionist model (model 1, Kröger et al., 2019) is used for illustrating the buildup of the mental syllabary during early processes of speech acquisition, i.e., babbling and imitation. (ii) A spiking neuron approach including a detailed modeling of time-dependent neural processes (model 2, Kröger et al., 2020) is used to illustrate different processes of motor programming. Thus, two different computer-implemented models are used here in order to illustrate different aspects of speech acquisition and speech processing. While conventional connectionist approaches are able to highlight processes of increasing self-organization in neural networks, which are based on learning as they appear during speech acquisition (see e.g., the SOMs approaches of Li et al., 2004; Kröger et al., 2014), contemporary spiking neuron approaches are able to combine cognitive discrete neural processes (here mainly lexical processes) with sensorimotor processes and these models are able to model temporal aspects of neural and peripheral processing in a straight forward way (see e.g., the large scale neural model of Eliasmith et al., 2012).

THE SKETCH FOR A MODEL OF SPEECH PRODUCTION

Our model of speech processing separates modules or sub-networks for processing (production or perception) and for the storage of knowledge and skills (neural repositories). Linguistic knowledge is stored in the *mental lexicon* (repository for words, lemmas, and phonological word forms) and in a grammatical rule component (not implemented thus far). Phonetic knowledge and sensorimotor skills are stored in the *mental syllabary* (repository of motor and sensory forms of already learned syllables).

While a level of phonological representations is of central importance in many production and perception models, this level emerges in our model during the entire process of speech acquisition. For production the *phonological form* represents the output level for the cognitive-linguistic part of the model (e.g., Levelt et al., 1999) and it represents the input level for the phonetic-sensorimotor part of the model (e.g., Guenther, 2006). For perception the phonological form represents an intermediate level arising between the module of spectro-temporal analysis and the module of lexical processing in the ventral stream of speech perception as well as between the module of spectro-temporal analysis and the sensorimotor interface in the dorsal stream of speech perception (e.g., Hickok and Poeppel, 2007, 2016). Three developmental phases can be separated in our modeling approach. (i) *babbling* for processing of pre-linguistic proto-speech items (starts at an age of 3 months) and for developing an early version of the mental syllabary, i.e., a phonetic map; (ii) *imitation* as an early stage of language-specific speech *processing* (starts at an age of 6 months and overlaps with babbling) for further developing the mental syllabary and for developing the mental lexicon as well as phonological knowledge;

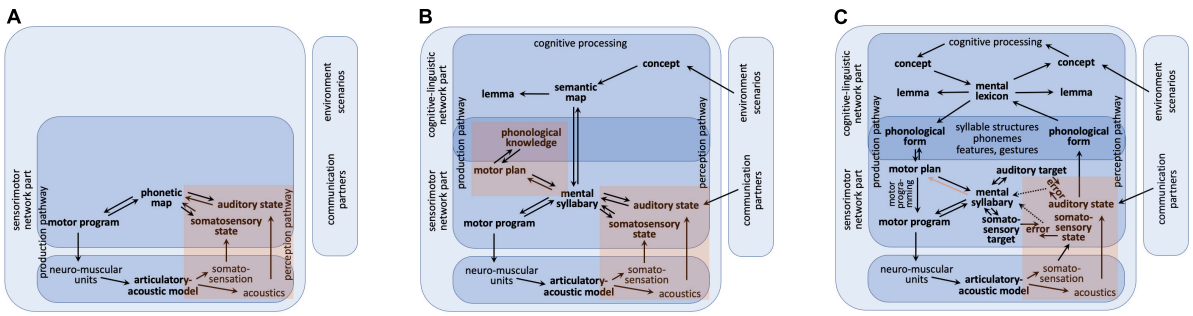


FIGURE 1 | Sketch for a model of speech production within three different developmental phases. **(A)** Babbling, **(B)** imitation, **(C)** complete production-perception network; red regions: sensory feedback pathways leading to somatosensory and auditory states **(A–C)** and leading to phonological knowledge mainly during imitation phase **(B)**; red arrow [in part **(C)** only]: pathway for forwarding information concerning existing motor programs from mental syllabary toward the motor planning level mainly during adult speech production. This leads to a choice between direct route (activation of an existing motor program via mental syllabary) and programming route (full specification of gesture score at motor plan level followed by generating a motor program for a syllable; see text).

and (iii) *adult speech* processing as a processing stage occurring after speech acquisition (starts at about 6 years of age) using mental syllabary and mental lexicon.

Babbling Stage of the Model Sketch

Babbling allows the model to learn auditory-to-motor relations from pre-linguistic proto-speech items and allows the model to build up a preliminary sensorimotor skill repository (called phonetic map) for storing the motor states, the somatosensory states, and the auditory states of already trained proto-speech items. The sensory and motor states are associated with each other for each trained proto-speech item. In our model sketch (as well as in Kröger et al., 2009) neural buffers are defined for hosting motor forms (*motor states*), and sensory forms (*auditory and somatosensory states*) of speech-like items. These buffers are connected to a neural SOM, called phonetic map, which is capable to activate each proto-speech item by activating its motor state and by co-activating its sensory states within the appropriate state buffers. Each proto-speech item is represented within the phonetic map by a specific neural activation pattern, which can be represented in a simple connectionist approach – in which the phonetic map is represented by a SOM – by the activation of a single node within the phonetic map (ibid.). Training is done here by babbling proto-V and proto-CV items over the whole range of vocalic vocal tract states and by combining these vocalic states with labial, apical, and dorsal closing gestures. An analysis of the resulting topology of the trained phonetic map reveals that these trained proto-speech items are ordered with respect to auditory as well as to somatosensory and motor features.

In our modeling approach, a babbling trial starts with the activation of a motor program for a proto-speech item (motor program in **Figure 1A**). The subsequent neuromuscular activation pattern leads to specific movements and displacements of speech articulators and this resulting articulatory pattern leads to an acoustic speech signal which is generated from the articulatory-acoustic vocal tract apparatus (**Figure 1**). The somatosensory (tactile and proprioceptive) feedback signals stemming from the articulatory movement pattern as well as the auditory feedback signal leads to neural activations in the appropriate sensory state buffers and to an activation at the level of the phonetic map (**Figure 1**). This temporally overlapping activation of a motor state and its resulting feedback sensory states for each trained proto-speech item leads to an association of sensory and motor states at the level of the phonetic map. If a proto-speech item has been produced several times (about 10 times per item, see Kröger et al., 2009, p. 802: 5000 training steps for 465 CV-training items and 5000 training steps for 500 V-training items) its motor and sensory states are associated and this item is stored or represented within the phonetic map.

Babbling ends with *a set of learned sensory-motor relations* (sensory comprises auditory and somatosensory) by storing auditory, somatosensory, and motor patterns for a variety of babbled proto-speech items. These auditory-to-motor relations are needed for later imitation training.

The somatosensory representation can be interpreted in our model as a simplified representation of motor states. While a motor program includes a detailed pattern of neural activations over time for all neuromuscular units of all articulators, the somatosensory state directly refers to articulation and thus allows a more direct and probably simplified description of an articulatory pattern.

The auditory state map is quantified in our approach by specifying the formant patterns of a syllable, which are the F1-, F2-, and F3-trajectories within the frequency-time space (Kröger et al., 2009; Cao et al., 2014; Kröger et al., 2014). The motor state map is quantified by listing the activation patterns of all gestures representing a syllable (Kröger et al., 2021). The somatosensory state map is quantified by specifying the movement patterns for the degree of lips opening, tongue tip, tongue body, and lower jaw elevation (Kröger et al., 2019).

An advantage of using articulatory gestures as basic production units is that proto-syllables can be interpreted as being composed of discrete units (i.e., raw gestures). These raw gestures already exist at very early stages of speech acquisition (i.e., the beginning of babbling) and the set of raw gestures can be used to define a set of distinctive features: (i) a proto-syllable contains at least a vocal tract opening gesture and/or contains a closing gesture (feature proto-V vs. proto-C). This allows a

separation of proto-V and proto-CV syllables; (ii) the articulator of a closing action separates labial, apical, or dorsal proto-consonants (feature: place of articulation = labial/apical/dorsal); (iii) the absence vs. presence of a glottal opening gesture separates voiced vs. voiceless proto-consonants as part of the proto-syllable (it can be assumed that this feature voiced/voiceless develops later during babbling and is refined during imitation phase); (iv) the absence vs. presence of a velopharyngeal opening gesture separates nasal vs. oral consonants (it can be assumed that the feature nasality as well develops later during babbling and imitation phase). It should be noted that the timing of all gestures as well as their targets are still raw (i.e., proto-gestures) and not fine-tuned with respect to any target language at this stage of speech learning.

Imitation Stage of the Model Sketch

The model is now capable for imitation of language-specific speech items picked up from external speakers (caretakers or communication partners, see **Figure 1B**) because a preliminary knowledge base for auditory-to-motor state mappings has been established during babbling as part of the phonetic map. An incoming auditory pattern, for example a word, which is tried to be imitated by the child, activates an auditorily similar babbling item available in the phonetic map. Because the activated babbling pattern only approximates the incoming auditory patterns the motor program of a babbling pattern is systematically varied during imitation until a word production is rewarded (i.e., understood) by the communication partner. This allows the model to adapt link weights between phonetic map and state maps in order to be able to reproduce this new or refined motor state and its appertaining feedback sensory states in the phonetic map as a preliminary word realization.

Here we assume that imitation of a word – which activates a node in the self-organizing phonetic map – always co-activates a node in the self-organizing semantic map and thus leads to an activation of the word within the semantic map as well as to an activation of its phonetic realization within the phonetic map. Therefore, we presume *communication scenarios* in which the child points or focuses on an object like a ball, then looks at the caretaker and thus forces the caretaker to produce that word. Thus, during the period of actively imitating a specific word, the cognitive-linguistic as well as the sensorimotor part of the model is involved which leads to a bilateral activation and association of a specific neural state within the self-organizing semantic and within the self-organizing phonetic map (**Figure 1B**; and see Kröger et al., 2011).

Imitation of a word may occur many times during the imitation phase which leads to an increase in approximating the correct phonetic realization of the word. This process is called *refining, tuning, and differentiating of motor patterns* (cf. Nittrouer, 1995). In our modeling approach this process expands the set of already stored pre-linguistic sensorimotor items toward *a set of language-specific syllable realizations*. The phonetic map can now be relabeled as mental syllabary (**Figure 1B**). The nodes of the mental syllabary represent language-specific frequent syllables (Kröger et al., 2009; Kröger and Cao, 2015; Kröger et al., 2019).

As a result of learning during the babbling phase basic proto-vocalic and proto-consonantal gestures appear within raw motor programs (i.e., within raw gesture scores). Later during imitation training gesture scores and the appropriate motor programs can be differentiated not only with respect to basic types of gestures like closing and opening gestures or with respect to different gesture-executing articulators like lips, tongue tip and tongue dorsum but in addition with respect to *segmental features* like voicing and nasality because now the language-specific temporal location of proto-vocalic, proto-consonantal, velopharyngeal and glottal opening and closing gestures is learned. In our model sketch this type of motor representation is called *motor plan* or *raw gesture score*. Motor plans are available at the end of the babbling phase and thus during the entire imitation phase (motor plan level, **Figure 1B**). The process of refining, tuning and differentiation of motor plans and motor programs during the imitation phase leads to a set of language-specific gestures and features. This can be interpreted as emergence of phonological knowledge.

Thus, learned items (motor plans and motor programs and their sensory correlates) at the end of imitation can already be ordered with respect to phonological categories of the target language and thus can be interpreted as realizations of (language-specific) syllables (Kröger et al., 2019). Realizations of syllables belonging to the same phonemic state appear to build "phoneme regions" within the SOM (ibid.). Specific regions appearing within the SOM of the mental syllabary can now be labeled as phonological distinctive regions, because the syllable realizations stored here are linked with words and thus with meanings. The model develops *phonological knowledge* concerning (i) syllable structures, (ii) sound types (e.g., vowels vs. consonants) and (iii) sound features (e.g., place and manner of articulation). The syllable can now be specified by a bundle of features for the articulatory closing and opening portions occurring within the syllable and thus different types or categories of consonants and vowels can be distinguished and it can be assumed that the speaker (the model) now is aware of a sequence of different segmental categories (which can be labeled as a sequence of phonemes at the motor plan level). The corresponding motor plan state is labeled as "raw gesture score".

The step from imitation phase (**Figure 1B**) toward the adult production-perception model (**Figure 1C**) is done now by including a level of phonological representations (based on the phonological knowledge acquired during imitation) as a concrete neural state level within our model. It can be assumed that the neural structure for this neural state level is already defined within the developing neural network laid out for (later) speech processing and this structure starts growing during the imitation phase of speech acquisition (Zhang and Wang, 2007).

This phonological level is part of the top–down processing of speech production (from lexical output toward motor plan specification) and of the bottom-up-processing in speech perception (from auditory form to lexical processing) in the adult speech processing model. Moreover, the adult production-perception model includes additional processing steps at the

cognitive level based on knowledge developed during imitation training as described in the following section.

Adult Speech Processing Within Our Model Sketch

The adult model of speech processing (production and perception) can be separated in a linguistic-cognitive part and in a sensorimotor part. Moreover, the speech processing model comprises a *production pathway* and a *perception pathway*, but both pathways access the same knowledge repositories, i.e., the mental lexicon and the mental syllabary. The cognitive-linguistic part of the speech production network starts with cognitive processing on the concept level (thinking, decision making, forming intentions, etc.) followed by concept, lemma, and phonological form activation. The associated neural activation patterns appear within the concept, lemma, and phonological form buffers which are closely connected to the mental lexicon. Thus, the cognitive-linguistic part of the speech processing network transforms an intended utterance (or just a word) into a phonological representation or phonological form (**Figure 1C**). This level is comparable to the phonological form level following phonological encoding and preceding syllabification and phonetic encoding in the Levelt approach (Levelt et al., 1999) and this level is comparable with the phonological content buffer exemplified in the GODIVA model (Bohland et al., 2010).

On the perception side the cognitive-linguistic part of the speech processing network allows comprehension, i.e., concept activation based on the activation of a phonological form. The activation of the phonological form results for each acoustic input, if this input has been processed (or perceived) auditorily, i.e., after passing the sensorimotor part of the network (**Figure 1C**). In the context of the dual route approach of speech processing (Hickok and Poeppel, 2007, 2016) the level of phonological representation of perceived speech items follows the spectro-temporal signal processing module and precedes processing within the lexical and combinatorial network part of the ventral path.

On the production pathway side, the processing within the sensorimotor part of the production network starts with syllabification, i.e., with a fragmentation of the phonological sound sequence in chunks, which potentially can be directly executed as motor programs. Syllabification leads to an activation of motor plans, i.e., by activating *raw* gesture score for syllable-sized chunks as part of the incoming phoneme sequence (phonological form in **Figure 1C**). These raw gesture scores or discrete motor plan specifications are carrying not more information than a (segmental) phonological description, i.e., the phoneme sequence of the syllable itself (see below: concept of speech gestures). If a motor program exists for the syllable under production, this information is forwarded from the mental syllabary to the motor plan level (red arrow in **Figure 1C**) and the motor program of the syllable can be activated directly and subsequently the syllable can be executed. A motor program exists if that syllable has been trained during the imitation phase. If the syllable does not exist, it needs to be programmed in detail which starts with a *full specification* of the gesture score. In our model we have implemented two routes for realizing that process. (i) *Adapting approach*: The motor plan of a phonologically similar syllable can be activated, for which a motor program exists, and many quantitative parameters of the gesture score can be copied for a first version of the fully specified motor plan of the new syllable. This full specification affects quantitative parameters like duration of gestures and exact temporal coordination of beginning and ending of gestures while qualitative discrete (or phonological) gesture parameters are already set within the raw gesture score. (ii) *Assembling approach*: If no phonologically similar syllable exists, e.g., in case of the production of a CCV-syllable if only CV-syllables are acquired so far, the syllable can be fragmented in sub-syllabic parts like single consonants or CV-units like C@ (@ is SAMPA notation for schwa-sound) are activated and need to be assembled in order to build up a first fully specified motor plan which subsequently allows the generation of a first version of a motor program for the new syllable. An example is the generation of a motor plan for /pla/, which may be assembled from CV-syllables like /pa/ and /la/ or like /p@/ and /la/. This complex process is already established during the imitation phase of speech acquisition if more complex syllables need to be learned.

The task of fragmentation of the phonological sound chain of the utterance to be produced is called *motor planning*. Following Levelt et al. (1999) as well as Bohland et al. (2010), syllables are assumed to be to be basic units for *motor programming* and thus the phonological phase of motor planning is syllabification. Thus, the major task of motor planning is to identify syllabic units within the flow or sequence of phonological sounds already activated by the cognitive-linguistic part of the model. If the motor program exists for a syllable, the step of motor programming is just to activate and execute the motor program. If the motor plan does not exist, the planning needs to be extended by selecting sub-syllabic units and the subsequent process of motor planning is a complex procedure of combining sub-syllabic units.

In our model sketch (**Figure 1C**) the motor programs of already learned syllables are stored in combination with their appropriate sensory states (auditory and somatosensory states) in the mental syllabary. This is comparable to the fact that in GODIVA (Bohland et al., 2010) already existing (prelearned) motor programs are stored in the speech sound map.

A *bottom–up process for forwarding motor information* is introduced in our approach, i.e., forwarding the information, whether a motor program for a syllable exists or not from the level of the mental syllabary to the motor plan level (red arrow in **Figure 1C**) in order to allow the choice between direct motor program activation and motor planning.

A concrete neurobiologically inspired realization of specific parts of our sketch of a production model introduced here is given in section "Experiments" of this paper by introducing two different quantitative computer-implemented model approaches, which were used for the simulation of speech acquisition and adult speech production.

Phonological Knowledge and Structural Specifications of Syllables

Phonological and phonotactic knowledge is important for successful motor planning. It is needed for dividing the phonological sound sequence in syllables as well as for selecting phonologically similar syllables in the case of motor programming of new syllables. Thus, the typical phonological representation of a syllable is its phoneme sequence, e.g., /ba/, /da/, /dat/, /bla/, /blat/, /pa/, /ta/ etc. As already reported above it can be assumed that adult speakers have knowledge concerning different types of syllables, i.e., concerning basic syllable structures like CV, CVC, CCV, CCVC, etc. With respect to phonological features the type of syllable can be specified in more detail, e.g., as BV, PV, NV, LV, BLV, BNV, etc. Here CV syllables are separated concerning its initial consonant, i.e., voiced vs. unvoiced plosive, and nasal vs. lateral (B, voiced plosives; P, voiceless plosives; N, nasals; L, laterals). Consequently, CCV syllables can be separated with respect to initial voiced plosive-lateral-consonant clusters, initial voiced plosive-nasal clusters and so forth. In the next section it will be shown that a phonological representation of a syllable is comparable with a raw specification of a gesture score. A concrete example for the realization (or implementation) of phonological knowledge is given in **Supplementary Appendix A** for the computer-implemented neural simulation model 2.

The Concept of Speech Gestures and Gesture Scores

Gestures are target-directed dynamically defined movement units of speech articulation (Saltzman and Munhall, 1989; Browman and Goldstein, 1992; Goldstein et al., 2006; Kröger and Bekolay, 2019). Gesture scores define the temporal organization of gestures of a speech item like a word or a syllable. In the strict interpretation of Articulatory Phonology gestures and their temporal coordination are already defined at the lexical level for words. In our approach two levels of gestural representation are introduced. At the phonological level gestures are specified *discretely* (as feature bundles: *raw gesture score; discrete phonological specification of a motor plan*). At the sensorimotor level gestures are parameterized *quantitatively* by specifying the exact beginning and ending of each gesture activation within a gesture score, by specifying the (relative) articulatory velocity for reaching a target, and by specifying the exact target location. This results in a *phonetic or full specification of a motor plan*. This quantitative description of all gestures within a gesture score serves as basis for the generation of a detailed and complete neural activation pattern of all neuromuscular units controlling all articulators during the production of a speech item (*motor program*).

If a gesture is activated, it aims to reach a certain articulatory target in a certain time interval. Consonantal targets are places of articulation or location of constriction, as defined by features like labial, apical, or dorsal. Vocalic targets are specific tongue positions or specific vocal tract shapes, as defined by features like high, low, front, back, rounded, and unrounded. In the case of consonantal gestures, the definition of the gesture target also includes the definition of degree and type of constriction like full closure (plosives and nasals), near closure (fricatives), lateral closure (laterals), etc. The differentiation of plosives and nasals is achieved by introducing two further gestures, which are the velopharyngeal closing or opening gestures. Moreover, glottal opening and closing gestures appear for differentiating voiceless and voiced speech sounds. Thus, in the case of the velum and of the glottis, the goal of the gesture is the formation of a closure or of an opening of the glottal or velopharyngeal passage. Beside closing for phonation in case of the glottis (glottal closing gesture), a glottal tight closing gesture exists if a glottal stop sound needs to be produced. Beside closing for producing oral sounds in the case of the velum (velopharyngeal closing gesture), a velopharyngeal tight closing gesture needs to be activated simultaneously with the oral closure or near closure in case of plosives and fricatives. That guarantees an air-tight closure of the velopharyngeal port in case of obstruents (fricatives and plosives) for building up an oral pressure which is needed for producing frication noise in case of fricatives, or for producing a noise burst in case of plosives.

Gestures can be described as bundles of features, where the features mainly describe the gesture targets. It is shown below how single speech sounds (phonemes) can be built-up by one, two, or more gestures (see **Tables 1**, **2**), even if the gesture is seen as a non-segmental unit in the framework of Articulatory Phonology (Browman and Goldstein, 1992). It should be mentioned that some gestures may only represent one single distinctive feature, e.g., velopharyngeal opening/closing gesture for nasal/oral or glottal opening/closing gesture for voiced/unvoiced, while other gestures determine more than one feature, e.g., vocal tract shaping gestures determine the features high-low and front-back; consonantal constriction forming gestures generally determine place and manner of articulation.

In our adult production model (**Figure 1C**) the motor plan level is realized as raw gesture score. This specification directly results from the phoneme sequence (phonological form level in **Figure 1C**) but it simplifies the transition from a segmental-linguistic toward a motor description of each syllable. At the

TABLE 1 | On the relationship between phonemes and gestures.

Segment (phoneme)	Gestures, building up a segment (realizing that phoneme)
vowels (a, i, u, . . .)	vocal tract form gesture + labial form gesture + velopharyngeal closing gesture + glottal closing gesture
plosives, voiced	full closing gesture + velopharyngeal tight closing gesture + glottal closing gesture
plosives, unvoiced	full closing gesture + velopharyngeal tight closing gesture + glottal opening gesture
fricatives, voiced	near closing gesture + velopharyngeal tight closing gesture + glottal closing gesture
fricatives, unvoiced	near closing gesture + velopharyngeal tight closing gesture + glottal opening gesture
nasals	full closing gesture + velopharyngeal opening gesture + glottal closing gesture
lateral	lateral constriction gesture + velopharyngeal closing gesture + glottal closing gesture

TABLE 2 | On the relationship between gestures and features.

Gesture	Features, determined by the gesture
vocal tract shaping gesture (vocalic)	high-low, front-back
labial shaping gesture (vocalic)	rounded-unrounded
velopharyngeal closing vs. tight closing gesture	sonorant vs. obstruent
velopharyngeal (tight) closing vs. opening gesture	oral (non-nasal) vs. nasal
full vs. near closing gesture (consonantal)	plosive vs. fricative (or nasal vs. fricative)
full closing gesture (consonantal)	labial, apical, dorsal
labial constriction or closing gesture (consonantal)	bilabial, labiodental
apical constriction or closing gesture (consonantal)	dental, alveolar, postalveolar
dorsal constriction or closing gesture (consonantal)	palatal, velar
lateral constriction gesture (consonantal)	lateral, alveolar
phonation vs. glottal opening gesture	voiced vs. voiceless

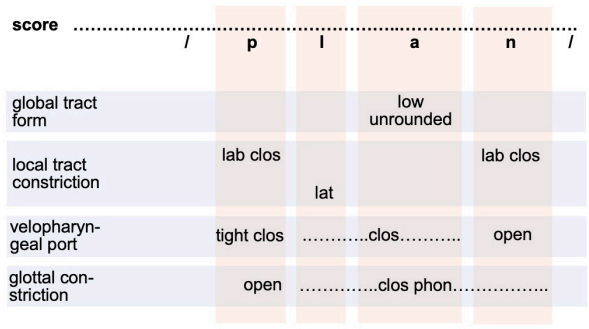


FIGURE 2 | Raw gesture score (or motor plan specification in terms of the sketch model in section "The sketch for a model of speech production") for syllable /plan/. Gestures are ordered on four articulatory tiers (rows as light blue boxes). Columns (light red boxes) indicate segment-gesture relations. The global tract forming gestures comprise two sub-gestures acting on different articulators: tongue and lips (see **Table 2**). The local constriction forming gestures act on three sub-tiers with respect to the gesture performing articulator: labial, apical, dorsal. Velopharyngeal and glottal gestures are listed in separate tiers. lab clos, labial full closing; lat, lateral; clos phon, closing for phonation.

motor plan level, all gestures are specified and arranged in four basic tiers (light blue rows in **Figure 2**). These tiers represent primary articulators, i.e., the main organs with which gestures are performed. A bundle of gestures appears for each sound (see **Figure 2** and **Table 1**). However, gestures assimilate if neighboring sounds show the same gestures on an articulatory tier (for example three neighboring sounds /lan/ are voiced in /plan/; see **Figure 2**). The vertical light red columns in **Figure 2** indicate all gestures which are related to one sound.

In order to generate a motor program from a motor plan (raw gesture score) all parameters of all gestures and the temporal coordination of all gestures need to be specified (fully specified gesture score). Thus, the exact points in time describing the beginning and the end of the neural activation of each gesture as well as describing the reaching and leaving of the spatial target region must be specified for each gesture. A full temporal specification of all gestures for a realization of /plan/ is shown

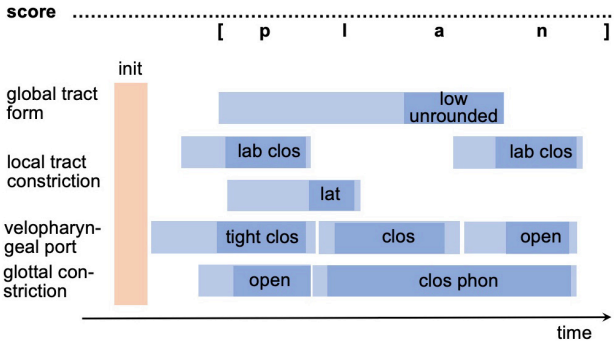


FIGURE 3 | Exact temporal specification of gestures [or motor program specification in terms of the sketch model in section "The sketch for a model of speech production"] for syllable /plan/ in a gesture score (fully specified gesture score). For abbreviations see figure caption of **Figure 2**; init, time instant for motor program activation.

in **Figure 3**. Light blue intervals in **Figure 3** mark the time interval defining beginning and ending of neural activation for each gesture while the dark blue intervals mark the beginning and ending of the target phase of each gesture. Thus, the initial light blue time interval marks the target-directed movement phase and the final light blue time interval the release phase.

Beside the exact specification of temporal parameters and target parameters, motor programming needs information concerning the extent to which secondary articulators need to be involved in the execution of a gesture. Primary articulators are those mainly defining the gestures target (lips, tongue dorsum, tongue tip, velum, and glottis). A typical secondary articulator is the lower jaw in case of vocalic and consonantal gestures. For example, to implement the motor program of the syllable /ba/, it must be clear how much the lower jaw should be raised during the formation of the lip closure within the production interval of /b/ in order not to endanger the subsequent production of the /a/, because during the production of the /a/ the lower jaw must be lowered to a certain degree. Thus, conflicting requirements for the secondary articulators involved in gesture realizations of temporally neighboring gestures must be brought into harmony with one another, and consequently the displacement of the primary articulators relative to the secondary articulators must be adapted accordingly.

EXPERIMENTS

In this section, we describe two sets of simulation experiments (using two different neural modeling approaches, i.e., model 1 and model "The sketch for a model of speech production," see below) that demonstrate key ideas described in the sketch of our overall model in Section "The sketch for a model of speech production." A comprehensive implementation of the model sketch is reserved for future work. In simulation experiment 1 (using the computer-implemented model 1) babbling and imitation training is simulated for small vocabularies using a connectionist network approach including

growing SOMs (Kröger and Cao, 2015; Kröger et al., 2019). In simulation experiment 2 (using the computer-implemented model 2) adult production for already learned as well as for new syllables is simulated using a spiking neuron network approach (Kröger et al., 2020). The computer-implemented babbling, imitation and adult production models used in these experiments are realizations of parts of the model sketch described above. Both computer-implemented models use different neuro-computational approaches. Model 1 implements nodes representing neuron ensembles and edges representing neural connections between nodes. Neural activity is averaged over defined time intervals as well as over neuron ensembles (see Kröger and Bekolay, 2019, p.133ff). This approach can be labeled as spatio-temporal activity averaging connectionist approach, while the spiking neuron approach used in model 2 (Eliasmith, 2013; Stewart and Eliasmith, 2014; Bekolay et al., 2014) includes the modeling of spatial and temporal details of neural processes and the modeling of neural control and decision processes.

Experiment 1

The simulation of babbling and imitation training was done for ten virtual learners or virtual speakers, modeled by ten instances of our connectionist neural model of speech learning (simulation model 1, see Kröger et al., 2019). The main goal of this simulation experiment is to show how phonological knowledge can be gained during early phases of speech acquisition based on motor information and sensory information resulting from processing of sensory feedback. The architecture of simulation model 1 represents the babbling and imitation stage of the model sketch (see **Figure 1A** for babbling and **Figure 1B** for imitation). The phonetic map and the semantic map are implemented here in form of growing self-organizing maps (G-SOMs; Cao et al., 2014).

Self-organizing maps are able to represent the main features of a set of training items (Kröger and Bekolay, 2019). The network in which a SOM is included always comprises one or more state maps in which cognitive, motor or sensory states of training items can be activated, while the SOM itself represents the learned knowledge in a structured way. All neurons of each state map are connected with all neurons of the SOM. The state maps can be seen as an input–output interface within the neural network. The learning algorithm of SOM is shaped in a way that with increasing training by applying each training stimulus several times, each stimulus is represented in a specific local area (i.e., by a specific set of neurons) of the SOM. Different regions within a SOM represent different types of training items, or in other words, specific regions of a self-organizing neural map represent different features of items. Thus, SOM are often also labeled as feature maps.

A typical disadvantage of SOMs is the fact that the number of neurons building up this map needs to be defined in advance. In order to model the learning procedure of SOMs in a more natural way, a self-organizing neural map should grow during learning (i.e., should capture neighboring neurons so far not part of the network). Our G-SOM approach includes this demand by starting with a basic set of just 4 nodes, which allows a representation of one or two training stimuli, but in the case of applying more stimuli to the network, a driving force can be defined which leads to a recruitment of more nodes (neuron ensembles) in order to be able to represent the whole set of incoming stimuli within this growing SOM (GSOM, see Cao et al., 2014). After training, the growing self-organizing network including the growing SOM, all state maps, and all edges between the nodes of these maps can be driven in a way that an activation of a neuron within this network leads to an activation of each specific (generalized) state which is included in the training set. In the case of our model the activation of a node within the growing SOM leads to an activation of the motor and sensory states of all (generalized) speech items represented by the training set.

Babbling training starts with a set of 70 items which combine proto-consonantal labial, apical, or dorsal closing gestures with proto-vocalic gestures. At the beginning of babbling training, gesture targets varied freely with respect to degree and location of constriction. During babbling training, bidirectional neural connections are established between the phonetic map and the motor and sensory state maps in order to associate motor and sensory states (see section "Method").

During imitation training bidirectional neural connections are established in addition between phonetic map (mental syllabary) and semantic map in order to associate motor and sensory states with concept states (meanings). The specification of neural connections as well as the adding of new nodes to both GSOMs is described in detail by Cao et al. (2014). The training corpus comprises 70 syllables (CV- an CCV-syllables). Each of these syllables are associated with a meaning thus establishing a word (Kröger and Cao, 2015). Five different vowels [V = /i, e, a, o, u/ and three different types of consonants; six plosives C = /b, d, g, p, t, k/, one glottal stop C = /?/(see the V-syllables in Kröger and Cao, 2015), two nasals C = /m, n/and one lateral C = /l/; 10 consonants in total] were allowed to be combined with each other resulting in 5 vowels × 10 consonants = 50 CV-syllables. Furthermore, four CC-clusters /bl, gl, pl, kl/ were allowed to be combined with all vowels resulting in 5 vowels × 4 CC-clusters = 20 CCV-syllables.

Method

In this connectionist model, concepts (**Figure 1B**) were represented by semantic feature bundles comprising 470 features (Cao et al., 2014). Thus, the neural representation of concept states comprises a neural state map of 470 neurons representing semantic features like "is living," "can bark," etc. The auditory state map comprises 24 × 64 neurons (nodes), where 24 neurons represent the frequency scale (bark scaled center frequencies) and 64 neurons represent a time scale (time steps of 10 ms; Kröger et al., 2019). The somatosensory state map comprises 4 × 64 neurons, where 4 neurons represent relative articulator to articulator distance for lips, articulator to vocal tract wall distance for tongue tip and tongue dorsum and a relative displacement value for the lower jaw (ibid.). The motor plan state map comprises 10 neurons for each gesture representing all gesture parameters (four points in time, two target values, one parameter naming the articulator) and the motor program state

comprises 2 × 64 neurons representing the agonist/antagonist neuromuscular activation of each of the 10 model muscle groups (six model muscle groups for controlling lips, tongue tip, and tongue body; three model muscle groups for controlling velum, glottis, and lower jaw). The articulatory-acoustic model used was developed by Birkholz et al. (2011).

All syllables or words (concept, sensory, and motor states) are coded by *distributed neural representations* within the state maps. Here many neurons of each state map can be activated in parallel for representing a specific state. All syllables or words are represented locally by one neuron in each GSOMs (*local neural representation*). Here each neuron or node represents a learned word or syllable. The link weights of the neural connections between a specific GSOM node and all nodes of a state map directly represent the neural activation pattern for that state for a specific word or syllable.

Babbling trainings was carried out using an early version of our GSOM model and a set of proto-V and proto-CV training items as introduced by Kröger et al. (2009). A later imitation training was carried out using 210 training items item as introduced by Kröger and Cao (2015). Here, each of the 70 syllables was imitated or resynthesized three times. The resynthesis procedure was done manually (Bauer et al., 2009). Ten runs of imitation training were executed leading to 10 different training results, representing 10 instances of the model (10 virtual learners). Each run comprised 50 imitation training cycles with 1470 training steps per cycle, i.e., 7 training steps for each of the 210 training items per training cycle (Kröger et al., 2019).

Results

Babbling training results in an association of auditory and motor states with an error rate of less than 5% after 10 training cycles per babbling item (10 cycles × 1470 training steps). During later imitation training syllable-to-meaning associations were established after 50 training cycles (50 cycles × 1470 training steps) for 66 ± 3 words (whole corpus is 70 words). This leads to a mean error rate of about 5.7% for word production (in this case a node of the phonetic map represents two different words, Kröger et al., 2019). Production errors occur here because the model represents the state of speech experience of children of one to one and a half year. For all correct syllable-to-meaning associations the phonetic representation is reliable because all phonetic realizations of a syllable (all nodes representing a syllable in the mental syllabary) are coded by nodes which are in a direct neighborhood with a maxim distance of 2 intermediate nodes. This reflects the fact that after training phonetic realizations of the same word vary only in a small range.

An evaluation of the number of feature regions per feature are summarized for each of 10 trained instances of the model (**Table 3**). A feature region is defined as a space within the G-SOM of the mental syllabary which includes all syllable nodes which represent syllables, which share at least one identical segmental feature value for type of vowel, for voicing, for place of articulation, and for manner of articulation (see **Figure 4**). Respecting the fact that our syllable corpus comprises two types of syllables (CV and CCV) this leads to a separation of (i) three vocalic features (vowel V within CV or CCV is a front vowel /i/ or /e/, vowel is a back vowel /o/ or /u/, vowel is the low vowel /a/), (ii) four glottal features (initial C of a CV syllable is voiced or voiceless, initial C of CV is voiceless, initial C of a CV is a glottal stop, both CC's in the CC-cluster within a CCV are voiced, or CC-cluster within CCV is a voiceless consonant followed by a voiced consonant), (iii) four consonantal features specifying manner of articulation (initial C in CV is a plosive, a lateral, or a nasal; the initial CC-cluster in CCV is a plosive followed by a lateral) and (iv) six consonantal features specifying the place of articulation (place of articulation of initial C in CV is labial, alveolar, velar, or glottal, place of articulation in the CC-cluster of CCV-syllables is labial for the first and alveolar for the second consonant or velar for the first and alveolar for the second consonant).

Feature regions (regions bordered by solid lines in **Figure 4**) were extracted manually here by applying the following rules: (1) A (sub-)feature region only includes nodes which are associated with syllables which are related to this feature. Moreover, all nodes need to be in direct neighborhood. (2) Two sub-feature regions are labeled as one feature region if they are in a relative neighborhood. That is the case if all three conditions stated below are fulfilled: (a) nearness: an interconnection of a length of less than 10 (free) network nodes can be found; (b) coherence: the interconnecting pathway does not cross more than one other feature region; (c) neutrality: network nodes representing a speech item (i.e., occupied network nodes) are not allowed to appear within this pathway. Interconnections between subregions representing one feature are indicated by dashed lines in **Figure 4**. (3) Outlier region: A subregion is not included in our evaluation if it appears with only one node.

The median of the sum of feature regions per feature is calculated for all single features as well as for all features belonging to a feature group for each of the ten model instances. "Type of syllable" is listed in **Table 3** as well because it reflects an important phonotactic feature (CV vs. CCV). The median over all model instances for all features and all feature groups is low ($\leq$3). This reflects a high degree of ordering of items with respect to all features at the level of the mental syllabary.

It can be concluded that our G-SOM realizations for the mental syllabary are capable to separate and thus to represent all features for all feature groups in an organized manner that is reflective of basic neural topography and map formation observed across multiple model instances. This can be seen as an indicator for the fact that the model is able to abstract these features and feature groups for describing phonological contrast if babbling training is done. It can be assumed that the model now is capable to establish a level of phonological representation as indicated in **Figure 1C**.

Experiment 2

In Experiment 2 the adult production of monosyllabic words is simulated for already learned and for new words respectively new syllables. It is demonstrated that single word production can be successfully simulated using a spiking neuron model (simulation model 2). This holds for already learned words as well as for new words and their corresponding syllables. In case of new syllables, the process of activating phonologically similar

TABLE 3 | Number of feature regions for each of 10 trained model instances (tr01, . . ., tr10) and median for all single features as well as for feature groups.

		tr01	tr02	tr03	tr04	tr05	tr06	tr07	tr08	tr09	tr10	Median: single features	Median: feature group
vocalic features	front	1	1	2	1	1	1	2	1	1	1	1	1
	back	1	1	1	1	2	1	1	1	2	1	1	
	low	1	1	1	1	1	1	1	2	1	1	1	
glottal features	voiced	1	2	1	1	1	1	1	2	3	1	1	2
	v.less	2	3	2	2	2	3	2	1	3	2	2	
	glott.stop	3	3	1	2	1	2	3	3	3	1	2,5	
	v.less-voi	2	1	1	2	1	3	3	3	2	2	2	
consonantal feature: manner	plosive	2	1	1	1	2	1	1	1	1	1	1	2
	lateral	2	2	3	4	3	3	3	2	3	3	3	
	nasal	2	3	2	3	1	3	4	2	1	2	2	
	plos-lat	1	1	1	2	1	2	3	2	2	1	1,5	
consonantal feature: place	labial	3	4	2	3	3	3	4	1	2	2	3	2
	alveolar	1	2	3	1	2	1	4	4	3	2	2	
	velar	3	3	3	3	2	3	1	2	3	2	3	
	glottal	3	3	1	3	1	2	2	3	2	1	2	
	lab-alveo	2	2	3	3	1	2	4	2	2	2	2	
	vel-alveo	2	2	3	1	1	3	3	2	3	3	2,5	
type of syllable	CV	1	2	1	1	1	1	1	1	1	1	1	1
	CCV	1	1	1	2	1	2	3	2	2	1	1,5	

Sub-regions are counted as one feature region and outlier-regions are not included (see text).

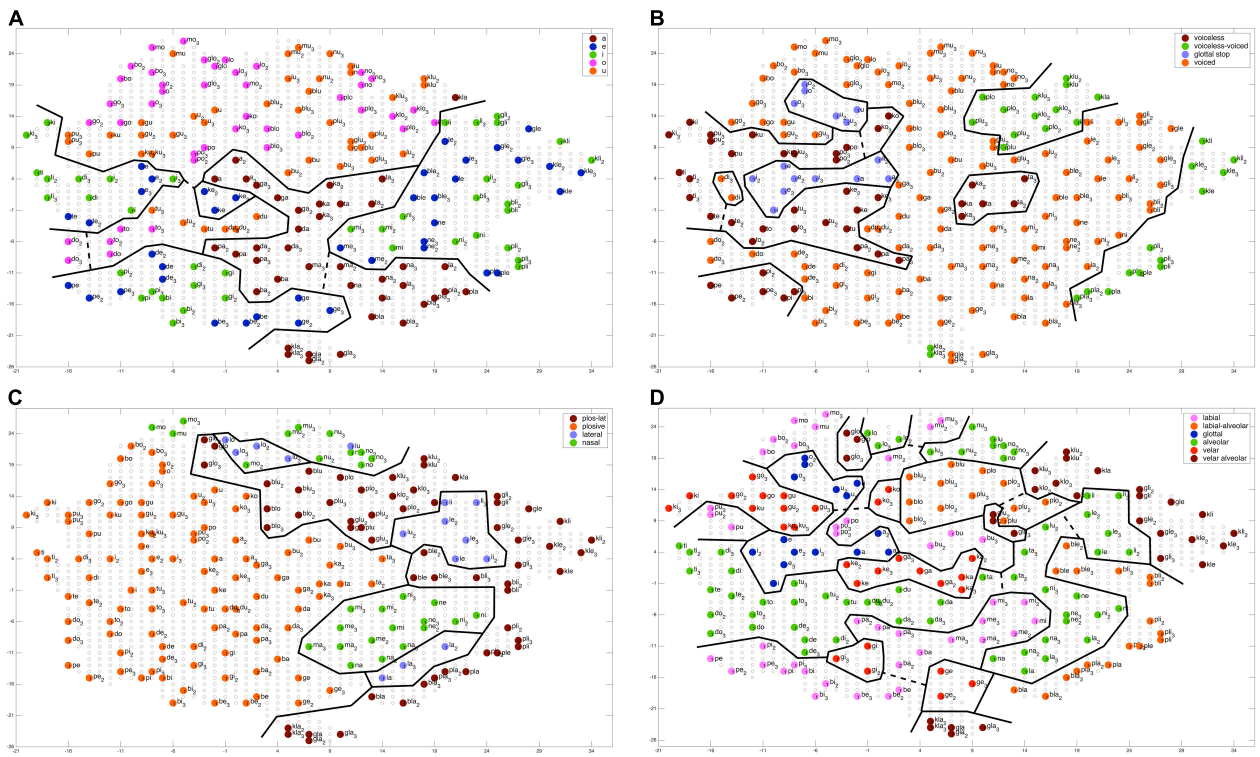


FIGURE 4 | Topology of a self-organizing phonetic feature map (G-SOM) after 50 training cycles for training one of 10 training runs. Network nodes are labeled with respect to different features: **(A)** vocalic features; **(B)** voicing features; **(C)** manner of articulation for initial consonants; **(D)** place of articulation for initial consonants. Solid black lines indicate the borders of feature regions, dashed black lines indicate interconnecting pathways for sub-regions representing the same feature (see text). Outliers are not marked here. They appear within a specific feature region as nodes with a different color, i.e., with a color that differs from the main color of a feature region. The main color of a feature region indicates the feature represented by that region.

syllables for adapting motor program information (see "adapting approach," section "Adult speech processing within our model sketch") is described in detail here.

The model used here (simulation model 2) is based on a spiking neuron approach including a detailed modeling of time-dependent neural processes by using the Neural Engineering Framework including the Semantic Pointer Architecture (Eliasmith et al., 2012; Eliasmith, 2013; Stewart and Eliasmith, 2014). The architecture of simulation model 2 represents the adult speech production which is part of the model sketch (see section "The sketch for a model of speech production" and see **Figure 1C**). The cognitive linguistic component comprises concept, lemma, and phonological form level (**Figure 1C** and see also **Figure 5**; Kröger et al., 2020). Semantic similarities of concepts as well as phonological similarities of syllables were modeled here using semantic pointer networks (Crawford et al., 2015; Kröger et al., 2016, 2020). In this approach semantic pointers represent meaningful neural activity patterns of words, lemmas, phonological forms of syllables (segmental phonological description of syllables, e.g., /plan/), or motor plan forms (cf. **Figure 2**) which can be activated in neural buffers. Modeling gestures and their temporal coordination (cf. **Figure 3**) is realized here using neural oscillators which are implemented as neuron ensembles (Kröger et al., 2021). A model language of 45 monosyllabic words (CV- and CCV-syllables, see **Supplementary Appendix B**) including an arbitrary mapping of word meanings to the phonological representation of theses syllables has already been learned and coded as sets of semantic pointers (lexical and phonological knowledge repository, see **Figure 5**).

Neural activation patterns of phonological forms appear within the phonological buffers P_prod or P_perc (for abbreviations of neural buffers see legend of **Figure 5** and **Supplementary Appendix D**). The related semantic pointer network for phonological forms comprises all four layers of phonological representations (see last paragraph of this section) which are implemented as deep layers in the S-pointer network of phonological forms (for S-pointer networks and deep layers see Kröger et al., 2020 and **Supplementary Appendix A**).

Figure 6 shows neural activation patterns for different neural buffers and neuron ensembles as function of time for the simulation of word production for the word "eat," represented arbitrarily by the syllable /ta/(already learned; motor program available) and for the simulation of the word "done," represented arbitrarily by syllable/du/ (not yet learned; motor program does not exist; here a phonologically similar syllable needs to be activated and adapted as new motor program; for the arbitrary linking of meaning and phonological form of syllables see **Supplementary Appendix B**). The top two buffers in **Figure 6** (in_con, out_con) represent incoming and outgoing control activity. The selection process for control actions results from comparing utility values (**Figure 6**, third row labeled utility_val) which are associated with different potential control actions. The control action used here is the selection of the direct or indirect production pathway (labeled as "d" or "i" in **Figure 5**). In case of an existing motor program for an already activated motor plan, the direct route is used (see **Figure 5**; the concept of direct and indirect pathways in motor planning/programming has been described as dual route theory by Whiteside and Varley, 1998; see also Miller and Guenther, 2021). In case of a non-existing motor program this program will be assembled or adapted from motor plan information of a similar syllable (indirect route).

The whole production activity, modeled in this Experiment 2 works as follows: The model (the speaker) starts with listening to auditory input and activates the target word which should be produced (control action LISTEN, **Figure 6**). The listening process uses the perception pathway of the model, i.e., subsequent activation of the target word within the buffers A_perc (auditory input level), P_perc (phonological level), and L_perc (Lemma level) and C_perc (concept level; see also **Figure 5**). The word is stored for a short time interval at the concept level (C_cog_in) but no other cognitive processing is done than forwarding the word toward the production pathway (via C_cog_out toward C_prod). Now the word passes the production pathway within the cognitive-linguistic part of the model via C_prod (concept level), L_prod (lemma level) toward P_prod (phonological level).

Subsequently, the production process activates the motor plan of the syllable (buffer M_plan; see **Figures 5**, **6**). In case of the syllable/ta/(word "eat," **Figure 6A**) a motor program exists and can be activated. In case of syllable/du/(word "done") no motor program exists (no activity occurs in buffer M_prog; see **Figure 6B**) and the activation of a similar syllable (/da/) is further processed (from buffer P_plan_sim to buffer M_score and from buffer M_plan to M_score_targ; see **Figures 5**, **6B**). In case of syllable /ta/ (**Figure 6A**) the motor program is directly executed via activation of the syllable oscillator M_score and subsequently the gestures associated with this syllable are activated (M_gest; here gest_tdn represent a consonantal gesture and gest_a represents the vocalic gesture; score_end informs the control component of the model that the next syllable can be activated; see also **Supplementary Appendix C**). In case of the syllable /du/ (**Figure 6B**) the motor program of a phonologically similar syllable is activated in buffer P_plan_sim (see activation of /da/in that buffer in **Figure 6B**) which triggers the activation of the syllable oscillator of the target syllable /da/. The semantic pointer which is now activated within the M_score_targ buffer gives the information, which gestures needs to emerge in the motor program for /du/. Thus, the new syllable /du/ is programmed by using the temporal information from the fully specified gesture score (motor program) of /da/, and by substituting the target of the vocalic gesture from the /a/ -target to the /u/ -target.

The control actions DIRECT_CALL_MOTOR vs. ADAPT_MOTOR (control module, buffer con_out, see **Figure 5**) determine whether the information of buffer M_plan or of buffer P_plan_sim is used for selecting or adapting a motor program. No activity in buffer M_prog (i.e., motor plan does not exist) leads to activation of ADAPT_MOTOR (indirect route, see **Figures 5**, **6B**) and subsequently leads to activation of M_score_targ (based on P_prod). Activity in buffer M_prog indicates the existence of the motor program for that syllable and subsequently leads to a direct activation of M_prog and M_score. Moreover, the control actions mentioned above determine whether the current motor program information (buffer M_prog) needs to be modified by taking into account

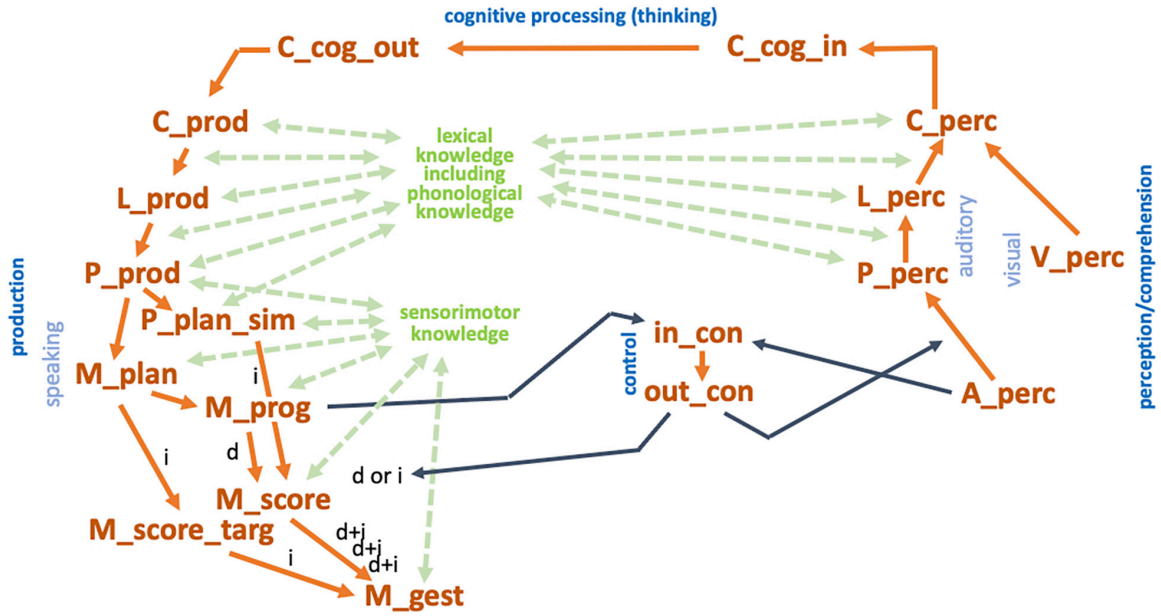


FIGURE 5 | Architecture of the computer-implemented model 2 used for simulation Experiment 2. Yellow-ocher: Neural buffers of perception pathway, production pathway, control module, and input and output buffers for cognitive processing (C_cog_in/C_cog_out: cognitive input/output buffer (concept level); in_con/out_con: input/output buffer of control module; . . ._prod/. . ._perc: buffers within the production/perception pathway with: C_. . ./L_. . ./P_. . . representing the concept, lemma, or phonological level and with A_. . ./V_. . ./M_. . . representing the level of auditory, visual, or motor representations; see also **Supplementary Appendix D**). In this experiment, cognitive processing is not modeled (only a direct forwarding the perceived speech input into the production pathway). Yellow-ocher arrows indicate the associative buffers which realize the neural associations between specific neural buffers (see also Kröger et al., 2020). Green: Knowledge reservoirs for lexical including phonological knowledge and for sensorimotor knowledge. Knowledge is implemented by semantic pointer networks (for S-pointer networks see Kröger et al., 2020). These networks determine similarity relations between semantic pointers representing concepts, lemmata, phonological forms, and motor plans. Moreover, these networks determine the relationships between concepts and lemmata, lemmata and phonological forms, phonological forms and motor plan, motor plans and motor programs or gesture scores, and between gesture scores and their appertaining gestures. Black arrows indicate input and output information for the control module of the neural network. The auditory input information from the perception side is used for starting the production process. The motor program input from the production side (buffer M_prog) indicates, whether a motor program is already stored in the mental syllabary or not. This information allows the control module (i) to start the production of the word or syllable (from buffer A_perc) and (ii) do decide whether the direct or indirect route (d or i) needs to be taken for syllable production. Direct route (d): If the motor program exists, an activation occurs in buffer M_prog) which leads to a further direct activation of the syllable oscillator (buffer M_score) and to the activation of the gestures associated with that syllable (buffer M_gest). Indirect route (i) also called programming route (assembling or adapting syllables): If the control model gets the information that for a currently activated motor plan (buffer M_plan) the appertaining motor program does not exist (no neural activity in buffer M_prog) and the indirect route is activated by the control module: activation of motor plan (buffer M_plan) leads to activation of a similar motor plan in buffer P_plan_sim. This leads to a co-activation of the syllable oscillator for that similar syllable (from buffer P_plan_sim to buffer M_score) and to an activation of the correct gestures (from buffer M_plan to buffer M_score_targ). This leads to a first estimation of the gesture score for the planned syllable which is activated subsequently in M_gest. This gesture score is based on gesture timing from the similar syllable, taken from M_score, but it includes the correct gestures as demanded in the motor plan, which are adapted from M_score_targ.

the information from buffer M_score_targ (case of adaptation: ADAPT_MOTOR) or whether the current motor program of the currently activated syllable can be used directly (case direct route: DIRECT_CALL_MOTOR, see **Figures 5**, **6A**).

The current version of simulation model 2 is capable of processing already learned words in the cognitive-linguistic part of the model. It is assumed that syllable learning by adapting an already learned similar syllable starts at the motor plan level using the processes within the sensorimotor component mentioned above in this section, because the phonological to concept relations for that new syllable are learned later, i.e., if the syllable can be produced already at the sensorimotor level. Furthermore, it can be assumed that this sensorimotor production process needs to be repeated a few times for that new syllable before the syllable becomes part of the mental syllabary and thus is available for direct motor plan execution. The current version of simulation model 2 is not able to model this learning process as well. Moreover, the association of the phonological form to the concept needs to be learned and stored as a new word entry in the mental lexicon as well. Thus, simulation model 2 only gives us a first impression how a *first production (or imitation) trial for a syllable adaptation process* can be spelled out in this spiking neuron-based modeling approach before this syllable is consolidated in the mental syllabary and the associated word is consolidated in the mental lexicon. In our current version of model 2 a temporary meaning-to-phonological form association is already available in the mental lexicon, but it would perhaps be more realistic to start the motor program adaptation process on the phonological form level of the syllable within the production pathway.

The vocabulary used in this simulation experiment covers 45 monosyllabic words and their associated syllables (see

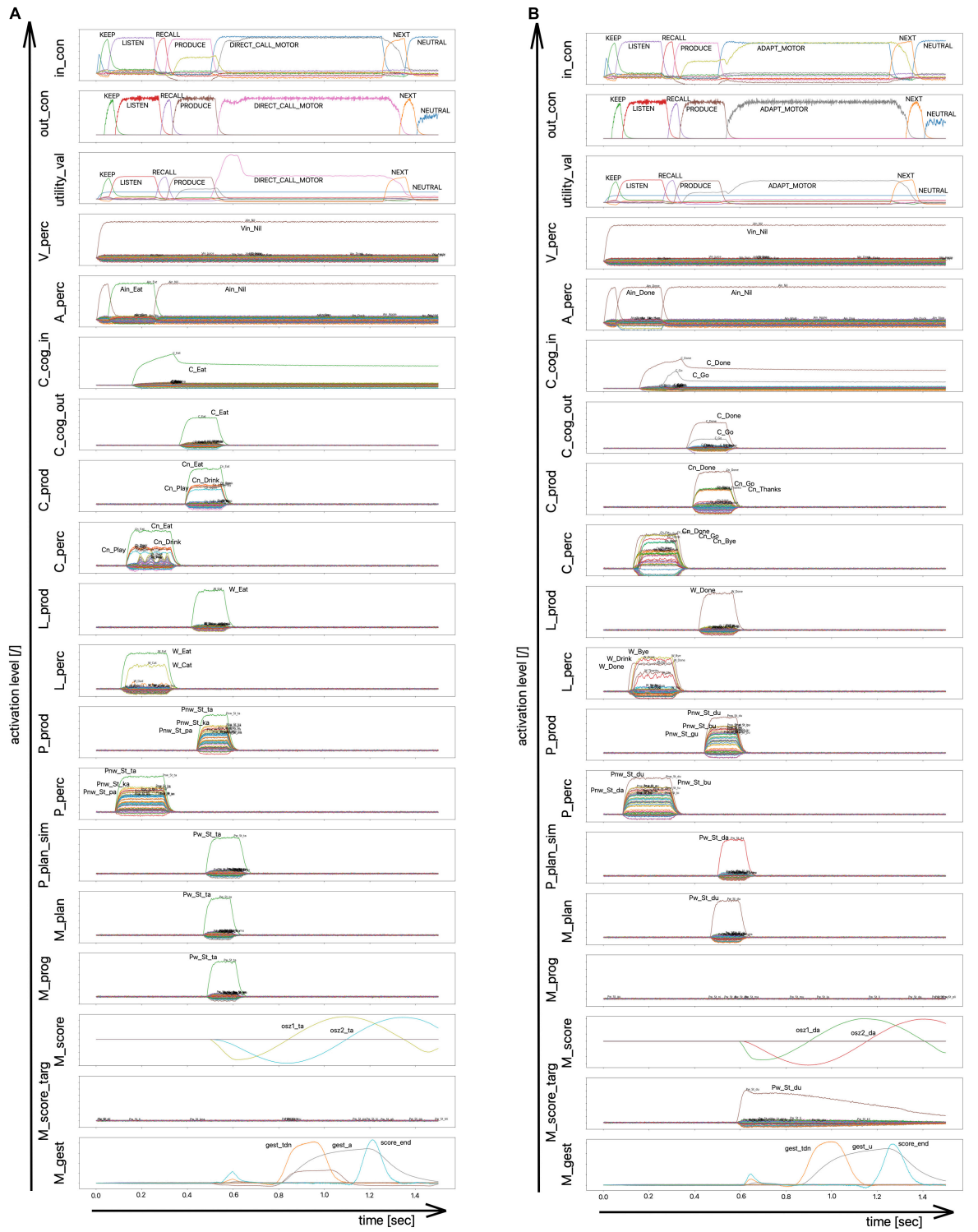


FIGURE 6 | Neural activation patterns as function of time for different buffers of simulation model 2 for two different word (syllable) productions. (i) Word production task for the word "eat" [left side: **(A)** the word has been trained] and (ii) for the word "done" [right side; **(B)** the word has not been learned yet]. The architecture of simulation model 2 is given in **Figure 5** and all acronyms for buffers are explained in **Supplementary Appendix D**. The ordering of buffers in this figure reflects the hierarchy i.e., the ordering of levels of the simulation model (cf. **Figure 5**): control module (con_in, con_out, utility_val) and input signals (A_perc, V_perc), cognitive processing module (C_cog_in, C_cog_out), perceptual and productive access buffers to the mental lexicon's cognitive level (C_perc, C_prod), to its lemma level (W_perc, W_prod), and to its phonological form level (P_perc, P_prod), followed by planning buffers (P_plan_sim, M_plan), programming buffers (M_prog, M_score, M_score_targ), and execution buffer (M_gest).

Supplementary Appendix B). The syllable corpus comprises 27 CV-syllables and 18 CCV-syllables. CV-syllables include all combinations of nine consonants, i.e., six plosives /b/, /d/, /g/(three voiced plosives) and /p/, /t/, /k/(three voiceless plosives), two nasals /m/ and /n/, one lateral sound /l/, and three vowels, /i/, /a/, and /u/. CCV-syllables include all combinations of six consonant clusters, /bl/, /gl/, /pl/, /kl/, /gn/, and/kn/, and of the three vowels.

Four different types of *phonological structure features* were differentiated in our model (see also **Supplementary Appendix A**): (i) type of *syllable* (values: CV, and CCV); (ii) type of *gesture score* (values: BV, PV, NV, and LV, for CV-syllables and PLV, BLV, PNV, and BNV for CCV-syllables; with B, voiced plosives; P, voiceless plosives; N, nasals; L, lateral); these types of gesture score can be seen as subtypes of syllables like BV vs. PV and are forming groups of nearly identical gesture scores with the same ordering and same types of gestures at a specific temporal position; (iii) type of *segments* within the syllable (values: /Ca/, /Ci/, /Cu/, /bV/, /dV/, /gV/, /pV/, /tV/, /kV/, /mV/, /nV/, and /lV/ for CV-syllables and /CCa/, /CCi/, /CCu/, /bCV/, /gCV/, /pCV/, /kCV/, /ClV/, and /CnV/ for CCV-syllables); (iv) type of a *feature* of a segment within the syllable (values: V_high, V_low, V_front, V_back, C_full, C_lat, C_lab, C_api, C_dors, C_nas, C_nonas, C_voice, and C_vless for CV-syllables, C1_lab, C1_dors, C2_lat, C2_nas, CC_nonas, C1_voice, and C1_vless for CCV-syllables, and V_high, V_low, V_front, and V_back for CV- and CCV-syllables). For understanding the meaning of each phonological structure feature and of its values, the values of all four types of phonological structure features are compared with each other in **Table 4**. These four different types of phonological structure features defining *four layers of phonological representations* are used below for defining *five different levels of phonological knowledge* (see section "Method").

Simulation experiments (see sections "Method," "Results for experiment 2a: CV-syllable learning," and "Results for experiment 2b: CCV-syllable learning") were performed using different levels of phonological knowledge in model 2 in order to evaluate (i) how much phonological knowledge is needed in order to adapt new syllabic motor programs from the motor plan and motor program information of similar syllables, (ii) which layers of phonological representations are most relevant for detecting similar syllables in order to perform a successful adaptation process for new syllables, and (iii) how the amount of phonological similarity between a detected similar syllable and the intended new syllable (i.e., the amount of gesture targets which need to be adapted) depends on the different levels of phonological knowledge.

TABLE 4 | Comparison of values of phonological structure features for all types of syllables occurring within the vocabulary.

Type of syllable	Features	Segments	Scores
CV	full closure	bV, dV, gV, pV, tV, kV, nV, mV	BV, PV, NV
CV	lateral	lV	LV
CV	labial	bV, pV, mV	–
CV	apical	dV, tV, nV, lV	–
CV	dorsal	gV, kV	–
CV	nasal	mV, nV	NV
CV	oral (non-nasal)	bV, dV, gV, pV, tV, kV, lV	BV, PV, LV
CV	voiced	bV, dV, gV, mV, nV, lV	BV, NV, LV
CV	voiceless	pV, tV, kV	PV
CV	high	Ci, Cu	–
CV	low	Ca	–
CV	front	Ci	–
CV	back	Cu	–
CCV	labial C1	blV, plV	–
CCV	dorsal C1	glV, klV, gnV, knV	–
CCV	lateral C2	blV, glV, plV, klV	BLV, PLV
CCV	nasal C2	gnV, knV	NV
CCV	oral (non-nasal)	blV, glV, plV, klV	BLV, PLV
CCV	voiced C1	blV, glV, gnV,	BLV, BNV
CCV	voiceless C1	plV, klV, knV	PLV, PNV
CCV	high	CCi, CCu	–
CCV	low	CCa	–
CCV	front	CCi	–
CCV	back	CCu	–

An empty field in the case of scores indicates that this set of syllables is not represented by score values. This holds only for specifications of different places of articulation or types of vowels.

Method

Ten different versions or variants of the production model (10 different "virtual speakers") were trained with respect to a variation in two different categories. Category 1 are five different levels of phonological knowledge. These levels are (i) all types of phonological structure features are available, (ii) all types minus scores, (iii) all types minus segments, (iv) all types minus the segment features, and (v) all types minus scores and minus segment features are available. Category 2 are two different levels of the model concerning the state of speech acquisition, i.e., concerning the level of already learned syllables. These levels are: (i) CV-learning stage: all CV-syllables with V = /a/ are already learned: CV-syllables with V = /i, u/ and all CCV-syllables (V = /i, a, u/) have yet to be learned; (ii) CCV-learning stage: all CV-syllables are learned (V = /i, a, u/), and all CCV-syllables with V = /a/ are learned, but CCV-syllables with V = /i, u/ have yet to be learned.

In part one of the simulation experiment (simulation experiment 2a) only CV-syllables were trained based on the acquisition level (i). All five different levels of phonological knowledge were simulated for producing each CV-syllable three times. In this experiment (2a) 5 levels × 3 trials × 27 CV-syllables = 405 simulation trials were carried out. The simulation trials can be differentiated according to whether a word can be produced directly (motor program of corresponding syllable exists; this is the case for 9 of 27 syllables, i.e., for 135 simulations) or whether a word (respectively a syllable) has not yet been trained (motor program needs to be programmed; 18 of 27 syllables, i.e., 270 simulations).

In part two of the simulation experiment (simulation experiment 2b) only CCV-syllables were trained based on the acquisition level (ii). All five different levels of phonological knowledge were simulated. In this experiment (2b) 5 levels × 3 trials × 18 CCV-syllables = 270 simulation trials were carried out.

The simulation trials can be differentiated according to whether a word can be produced (this is the case for 6 of 18 syllables, i.e., for 90 simulations) or whether a word (respectively syllable) has not yet been trained (12 of 18 syllables, i.e., for 180 simulations).

Results for Experiment 2a: CV-Syllable Learning

In the case of the 135 simulations of producing already learned words (CV-syllables with V = /a/), no errors occurred (see **Figure 7** top). Thus, already learned syllables can be easily produced in our model because the motor program of the corresponding syllable already exists. In the case of the remaining 270 simulations, depending on the type of phonological knowledge (five levels, see above), a phonologically similar motor program cannot be activated directly in several cases. For each model instance representing a specific type of phonological knowledge 3 trials × 18 CV-syllables = 54 simulations were performed for those CV-syllables for which no motor program exists (syllable has not yet been learned), i.e., for the CV-syllables with V = /i/ and V = /u/.

If all phonological structure features are available (case "all"; full phonological knowledge), the most similar syllable is activated directly in all cases of simulated production attempts. This means that the chosen most similar syllable always shows the same type of gesture score as is needed for adapting a specific new syllable and thus allows a successful adaptation process. The selected phonologically similar syllable differs only concerning the vowel target in this case.

In the case of the phonological knowledge level "all minus scores" no phonologically similar syllable can be activated at the first production attempt for 4 out of 54 cases, but correct similar syllables are activated in 92.6% of all attempts. In case of "all minus segments" that holds for only 1 out of 54 runs, leading to 98.1% correct productions, in case of "all minus scores and minus segments" that holds for 13 out of 54 runs (75.9% correct productions) but in case "all minus features" that holds for 51 out of 54 runs (only 5.6% correct productions; see **Figure 7** top, left columns per knowledge level).

These results describe cases in which the difference between the chosen phonologically similar syllables and the syllable for which the motor program needs to be generated (new syllable) is up to two consonantal features beside the vocalic feature. If the degree of phonological similarity is thus high, that only the vocalic feature is different, i.e., all consonantal features are correct and thus no gesture needs to be adjusted but the vocalic gesture, the percentage of productions decreases by about 7.4% (4 runs) in case of "all minus segments" and in case of "all minus scores and minus segments," and for about 1.8% (one run) in case of "all minus features" [see **Figure 7** by comparing the left (dark blue) and right (light blue) of each pair of bars; the right represents productions without any adjustment of consonantal gestures; the left represents productions, in which up to two consonantal features need to be adjusted]. The percentage does not decrease in the case "all" and in case "minus scores."

The most important result of Experiment 2a is that phonologically similar syllables for generating motor programs can be detected and activated in our model easily if the full phonological knowledge is available. Moreover, the strongest decrease for activating similar syllables occurs in case of reduction of phonological knowledge by "features." In this case only 5.6% of all productions are correct, i.e., enable the activation of a phonologically similar syllable to start motor programming. The amount of similarity between similar syllable and new syllable (syllable under production) does not depend strongly in case of all different levels of phonological knowledge. Thus, in most cases of syllable adaptation the similar syllable differs only in one gesture parameter (as demonstrated in the example given in **Figure 6B**).

Results for Experiment 2b: CCV-Syllable Learning

In the case of the 90 simulations of producing already learned words (CCV-syllables with V = /a/), no errors occurred. Learned syllables can therefore be easily produced in our model. In the case of the remaining 180 simulations, errors occurred (no phonologically similar syllable can be activated) in different quantity depending on the type of phonological knowledge available. For each type of phonological knowledge 3 trials × 12 CCV-syllables = 36 simulations were done for CCV-syllables with no motor programs available (syllables that had not yet been learned), i.e., for the CCV-syllables with V = /i/ and V = /u/.

If all phonological structure features are available (case "all"; full phonological knowledge), the most similar syllable for realizing the production of the new syllable is found directly for 33 out of 36 runs of simulated production attempts. Thus, in case of 91.7% of all productions a phonologically similar syllable is already found directly in the first run (**Figure 7** bottom). Simulation results indicate, that in the remaining four runs, a phonologically similar syllable is found in the second production attempt, so that motor program generation here as well is possible without problems.

In the case of the phonological knowledge level "all minus scores" phonologically similar syllables can be activated directly in 32 out of 36 runs (88.9%), in case of "all minus segments" in 33 out of 36 runs (91.7%), and in case "all minus scores minus segments" in 35 out of 36 runs (97.2%). But in case "all minus features" the direct activation of a phonologically similar syllables occurs only in 10 out of 36 runs (27.8%; see **Figure 7** bottom, left columns per knowledge level).

These results describe cases in which the difference between phonologically similar syllables and the syllable for which the motor program needs to be generated is up to two consonantal features beside the vocalic feature (mainly place or place and voice). If the degree of phonological similarity should be thus high, that only the vocalic feature is different, i.e., all consonantal features are correct, the percentage of productions decreases by about 8.3% (three runs) in case of "all minus scores," decreases by about 19.4% (7 runs) in case of "all minus segments," and decreases by about 30.6% (11 runs) in case of "all minus scores and minus segments."

The most important result of simulation Experiment 2b is that in case of programming a new CCV-syllable a phonologically similar syllable can be detected and activated in 88.9% up

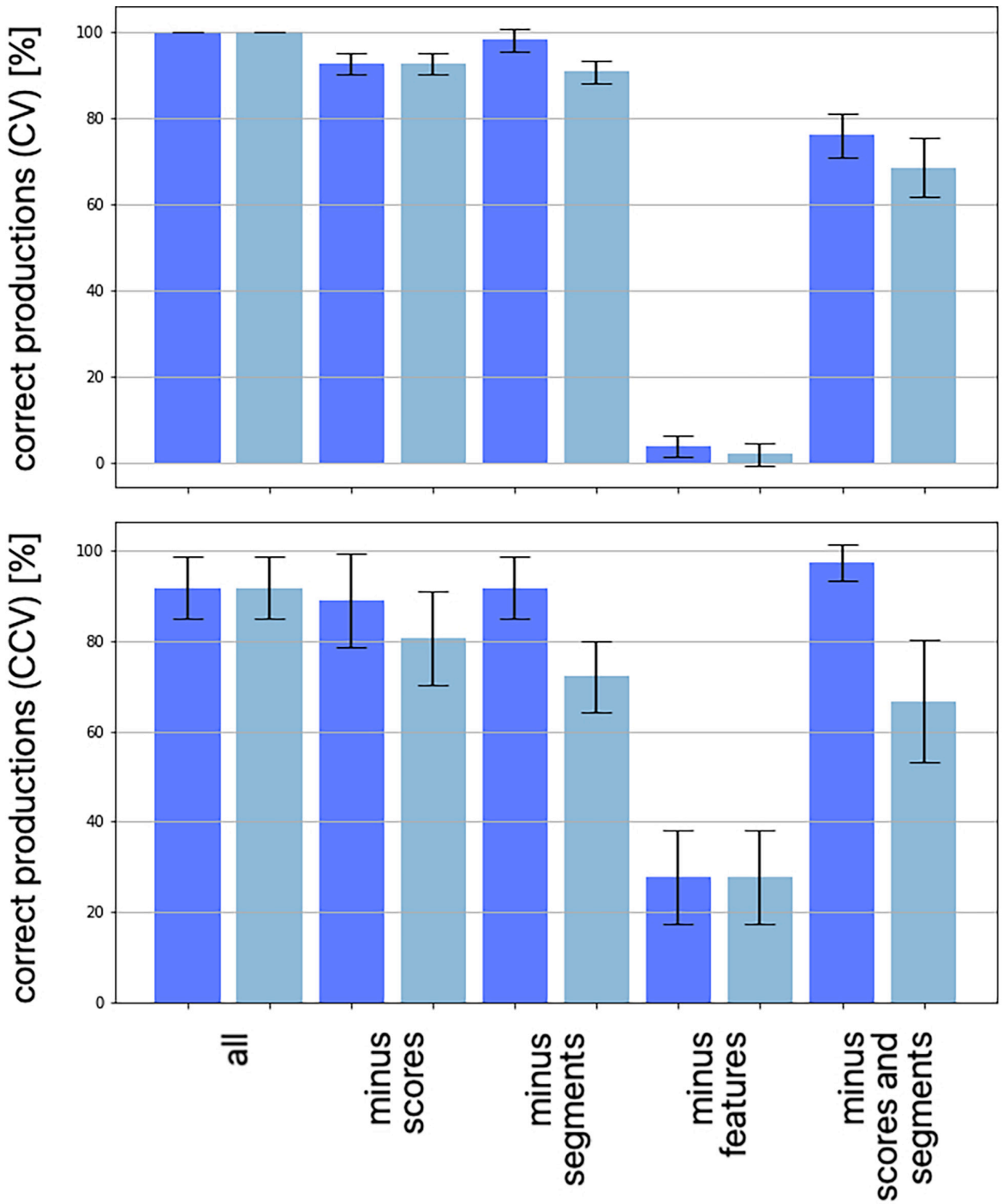


FIGURE 7 | Percentage of correct productions for CV syllables and CCV syllables in case of five different degrees of phonological knowledge available during motor planning. These levels are (i) **all** four types of phonological structure features are available (type of *syllable*, type of *gesture score*, type of *segments* within the syllable, and type of a *feature* of a segment within the syllable; see section "Experiment 2" for a detailed description of all types of features and possible feature values), (ii) all types of structure features **minus scores** are available (only the phonological information concerning the structure feature "type of gesture score" is not available), (iii) all types **minus segments** are available (only the phonological information concerning the structure feature "type of segments within the syllable" is not available), (iv) all types **minus features** are available (only the phonological information concerning the structure feature "type of feature of a segment within the syllable" is not available), and (v) all types **minus scores and minus segments** are available (only the phonological information concerning the structure feature "type of gesture score" and concerning the structure feature "type of segment within the syllable" is not available). Furthermore, we separated the types of similar syllables with respect to the amount of similarity. Left side, dark blue: high degree of similarity: only the target of the vocalic gesture needs to be adapted. Right side, light blue: lower degree of similarity: target of the vocalic gesture and the target of up to two gestures affecting consonants within the syllable needs to be adapted.

to 97.2% of all trials depending on the level of phonological knowledge. Only in case of "all minus features" phonological knowledge is so small that a phonologically similar syllable is activated only in 27.8% of all production attempts. Thus, the results concerning the three points listed above [i.e., (i) how much phonological knowledge is needed in order to adapt new syllabic motor programs from the motor plan and motor program information of similar syllables, (ii) which layers of phonological representations are most relevant for detecting similar syllables in order to perform a successful adaptation process for new syllables, and (iii) how the amount of phonological similarity between detected similar syllable and intended new syllable depends on the different levels of phonological knowledge] are comparable for CV and CCV syllables.

DISCUSSION

A sketch for a model of speech production has been proposed including developmental aspects like the buildup of skills and speech knowledge during early phases of speech acquisition. While other models mainly concentrate on modeling of cognitive-linguistic aspects of speech production (e.g., Levelt et al., 1999) or mainly concentrate on modeling the sensorimotor aspects of speech production (e.g., Guenther, 2006; Bohland et al., 2010) it is the goal of our model sketch to give the complete view on speech production, i.e., linguistic as well as sensorimotor aspects. While the phonological level can be used for interfacing cognitive-linguistic and sensorimotor model parts of a speech production model in case of adult speech production the situation is more complex in early phases of speech acquisition. Thus, a comprehensive model of speech production needs to include the developmental processes occurring in speech processing. Our model sketch takes this into account by including early phases of speech acquisition, i.e., the babbling and the imitation phase.

While babbling constitutes a first realization of the sensorimotor part of the speech processing model, imitation establishes the cognitive-linguistic part and in addition further develops the sensorimotor part of the model. Imitation needs specific communication scenarios like triangulation (i.e., focusing an object and learning its meaning and its pronunciation by imitating the productions of the communication partner) and leads to the buildup of a mental lexicon as repository for concepts and lemmas as well as of the mental syllabary as a repository of sensory and motor forms of syllables. Here, imitation training tunes and differentiates already stored pre-linguistic babbling speech items (stored in a proto-syllabary, called phonetic map in our approach) into the direction of target-language specific speech items, mainly syllables. These assumptions play a central role in our model sketch and are based on literature (e.g., Levelt and Wheeldon, 1994; Oller, 2000; Cholin et al., 2006; Hickok et al., 2011; Buder et al., 2013; Lytle and Kuhl, 2017; Redford, 2019).

Furthermore, our sketch of a production model postulates that during the imitation phase the mapping between the items represented in the mental syllabary and in the mental lexicon introduces distinctiveness at the interface level between both repositories and thus converts phonetic into phonological features. This hypothesis is underlined by the emergence of phoneme regions at the level of the mental syllabary if the mental syllabary is modeled using a SOMs approach (e.g., Kröger and Cao, 2015). At the beginning of the imitation phase, phonological forms are not available which could be stored in the mental lexicon, but neural connections are established now between both repositories which associate words with syllables. Because the word-to-syllable association is established in a bidirectional way during the imitation phase (e.g., Kröger and Cao, 2015) firstly speech production can be simulated now by activating words to syllables from the mental lexicon toward the mental syllabary and secondly the dorsal stream of speech perception can now be simulated by using syllable-to-word associations from mental syllabary toward mental lexicon. Moreover, a successful word-to-syllable and syllable-to-word association allows the phonetic features to become categorical. Now, different feature values allow a separation of syllables which represent words of different meaning. In our model sketch a phonological level is established now, which on the side of speech production appears as interface between cognitive-lexical and sensorimotor processing and which on the side of speech perception now allows to establish the ventral stream of speech perception, which forwards speech items from the auditory processing via the phonological processing toward a lexical processing (cf. Hickok and Poeppel, 2007, 2016).

At the end of the imitation phase the adult speech processing model is established which comprises a cognitive-linguistic component as already introduced by Levelt et al. (1999) and a sensorimotor component which separates motor and sensory states and thus forward motor and feedback sensory processing (as introduced by Guenther, 2006; Bohland et al., 2010) and which separates motor planning and motor programming. In our model sketch, gesture scores are introduced as a vehicle for transforming segmental phonological syllable specifications into motor forms by specifying raw or categorical gesture scores followed by fully specified or quantitative gesture scores.

Two simulation experiments were carried out in this paper to substantiate distinct aspects of our sketch for a model of a speech production. In a first simulation experiment the model components are realized by implementing growing self-organizing networks for the sensorimotor as well as for the cognitive-lexical part of the model. A main result of this modeling is the ability of topographically organizing and later of differentiating speech items with respect to phonetic and later with respect to phonological features. Thus, the simulation of babbling and imitation by using growing SOMs exemplifies the emergence of phonological features based on knowledge gained from motor representations and sensory representations resulting from sensory feedback information.

In a second simulation experiment which is carried out by using a spiking neuron approach including an explicit modeling of time-dependent neural processes (Eliasmith, 2013) it is demonstrated how a new syllable is learned if motor programs for phonologically similar syllables are available. Here, the gesture timing parameters are copied from the already existing motor program of the similar syllable and only some gesture targets

need to be exchanged to generate a first version of a motor program for the new syllable (adapting process). Further fine-tuning of gesture parameters may occur in further production attempts of this syllable. In two experiments it is shown that the phonological information concerning features like vocalic high-low front-back or consonantal place of articulation is important for allowing to select syllables exhibiting similar gesture scores. Moreover, it should be stated that phonological information can be used to specify or characterize segments as well as gestures at the motor plan level. What remains to be solved is the question how new *types* of syllables like first CCV-syllables can be learned if only CV-syllable motor plans are available (assembling process).

It is not the goal of the model sketch developed in this paper to combine segmental and gesture-phonological descriptions in one approach at each level of the model. As stated by Goldstein et al. (2006) segmental approaches in comparison to a gestural approach "appear to present problems ... when they attempt to account for the temporal structure of speech - like regularities in relative timing between units, stochastic variability in that timing, and systematic variability in timing due to rate, speaking style, and prosodic context" (ibid., p. 222, footnote 6). Moreover, "temporal sliding of some (but not all) production units with respect to one another ... " (ibid.) is not possible on the segmental level but increasing gestural overlap together with temporal reduction of duration of some gestures for example leads to significant effects at the segmental phonetic surface like assimilations and elisions as they appear in casual of fast speech. This has been demonstrated by Suprenant and Goldstein (1998) as well as by perceptual studies in early versions of our own gesture-phonological approach (Kröger, 1993). These results indicate that a gestural control approach cannot be replaced or mixed with a segmental control approach at a quantitative phonetic level where time and temporal relations between phonetic articulatory events come into play. And these facts are consistent with the sketch of a production model introduced in this paper. Within the sensorimotor part of our production model, we start with a raw gesture score description followed by a full quantitative specification of the gesture score for controlling articulation. A segmental phonological description of lexical units down to the syllable is introduced in our approach exclusively within the cognitive-linguistic model part.

Moreover, it is stated above that in our approach a raw gesture score which specifies gestures purely in a phonological manner as distinctive units can be converted into a segmental phonological description using phonemes as distinctive units and vice versa. Thus, our approach allows a description of lexical units by using a segmental or a raw gestural description comparable to that given by coupling graphs in the concept of Articulatory Phonology (Goldstein et al., 2006). But proto-syllables occurring in early phases of speech acquisition are described in our approach exclusively as gesture scores. Segmental phonological descriptions in our model appear later during speech acquisition and appear in our approach in the adult production model as a result of language-specific learning which occurs during the imitation phase. This is consistent with Goldstein et al. (2006, paragraph 7.2.3, p. 226): "What is the "glue" that allows articulatory gestures to be coordinated appropriately within a word form and that permits the establishment of lexically distinct coordination patterns across word forms? One possibility would be to hypothesize that gestures are organized into hierarchical segment and syllable structures that could serve as the scaffolding that holds the gestures in place through time. However, these relatively complex linguistic structures could only exist as part of an already developed phonology and could not be available pre-phonologically as part of an account of how articulatory gestures begin to be combined into larger structures."

Our sketch of a model is based on well-known neurobiologically inspired approaches of speech production and speech perception (e.g., Levelt et al., 1999; Guenther, 2006; Hickok and Poeppel, 2007, 2016; Bohland et al., 2010; Guenther and Vladusich, 2012) and is consistent with these approaches. One further main goal of this paper was to highlight the importance of sensory feedback and thus to emphasize the importance of motor *and* sensory syllable representations at the level of mentally syllabary for establishing phonological knowledge and a phonological level during speech acquisition as interface between the cognitive-linguistic and the sensorimotor part of a production-perception model.

Moreover, a bottom–up pathway for motor information concerning already existing motor programs is introduced to enable the selection of separate processing routes for producing already learned syllables (direct route) versus producing syllables which are not learned so far and thus having no ready-made motor programs available (programming route). As part of the programming route the adapting process is implemented successfully and works satisfactorily if enough phonological information is available. Further work is needed for implementing the assembling process in order to generate motor programs for new types of syllables. This assembling process is not only an important process for adult speech production but also an important sub-process already occurring during the imitation phase of speech acquisition if new types of syllables must be acquired.

A limitation of our current modeling approach could be that the production of pseudowords is not included. But this reflects the fact that pseudoword production primarily appears in scenarios like logopedic diagnosis in case of suspicion on specific speech and language disorders or in case of suspicion of hearing loss. The main task in speech acquisition is that the child tries to communicate information (i.e., meanings in form of lexical items). Even if first production trials of words are relatively degraded it is the goal of the child to be understood by its caretaker or communication partner. The production of pseudowords differs from this goal but can be easily incorporated in our model sketch if a neural perception-production shortcut is included at the phonological form level as it has already been realized in our spiking neuron modeling approach for the simulation of phonological retrieval aids in case of an logopedic diagnostic word retrieval scenario (Kröger et al., 2020).

Both simulation experiments outlined in this paper can be seen as a proof of principle (i) for the idea how phonetic features – which appear in the sensorimotor representations of syllables at the level of the mental syllabary – become phonologically relevant

by linking syllables with word meanings, (ii) how the emergence of knowledge and skill repositories (i.e., mental lexicon and mental syllabary) can be specified at the neural level as growth of neural maps and as an adjustment of neural connections between all neurons of these maps, (iii) how in case of speech perception and production of a word the flow and processing of information can be simulated in detail at concrete neural levels using a spiking neuron approach, and (iv) how specific processes of speech production like motor programming of a new syllable can be implemented in detail by adapting motor program features from phonologically similar and already learned syllables.

But in our current work we still must use two different neural modeling approaches in order to highlight distinct aspects of the model sketch. Model 1 (simulation experiment 1) is a comparably simple connectionist approach which is not capable of modeling spatial and temporal details like the generation of spike patterns (i.e., specific neural activation patterns for single neurons) but which allows the quantification of mean activation rates over specific time intervals (like activation interval for selecting a lexical item) and over a set of neurons (like neuron ensembles or neuron buffers representing a specific cognitive, lexical, sensory, or motor item). Model 2 (simulation Experiment 2) is a more detailed spiking neuron approach capable of modeling the spiking behavior of cortical neurons, which subsequently allows a detailed and straight forward modeling of the temporal aspects of the flow and of the processing of neural activation patterns within the speech production-perception network. It is a main goal or our future work to unify this modeling approaches into one (probably spiking neuron) approach capable of instantiating all developmental aspects and all processing aspects of the production-perception network. Currently one of the main difficulties is to model developmental aspects in a spiking neuron approach because of the immense computational loads appearing in learning scenarios.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

BK developed the raw architecture of both simulation models (GSOM-model = model 1; Nengo-model = NEF-SPA model = model 2), conducted the experiments, and wrote the manuscript. TB developed main routines for simulation model 2 while MC developed main routines for simulation model 1. TB, MC, and BK together developed the detailed architecture of both simulation models.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnhum.2022.844529/full#supplementary-material

REFERENCES

Bauer, D., Kannampuzha, J., and Kröger, B. J. (2009). "Articulatory speech re-synthesis: profiting from natural acoustic speech data," in *Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions, LNAI 5641*, eds A. Esposito and R. Vich (Berlin: Springer), 344–355. doi: 10.1007/978-3-642-03320-9_32

Bekolay, T., Bergstra, J., Hunsberger, E., DeWolf, T., Stewart, T. C., Rasmussen, D., et al. (2014). Nengo: a python tool for building large-scale functional brain models. *Front. Neuroinform.* 7:48. doi: 10.3389/fninf.2013.00048

Birkholz, P., Kröger, B. J., and Neuschaefer-Rube, C. (2011). Model-based reproduction of articulatory trajectories for consonant-vowel sequences. *IEEE Trans. Audio Speech Lang. Process.* 19, 1422–1433. doi: 10.1109/tasl.2010.2091632

Bohland, J. W., Bullock, D., and Guenther, F. H. (2010). Neural representations and mechanisms for the performance of simple speech sequences. *J. Cogn. Neurosci.* 22, 1504–1529. doi: 10.1162/jocn.2009.21306

Browman, C. P., and Goldstein, L. (1992). Articulatory phonology: an overview. *Phonetica* 49, 155–180. doi: 10.1159/000261913

Buder, E. H., Warlaumont, A. S., and Oller, D. K. (2013). "An acoustic phonetic catalog of prespeech vocalizations from a developmental perspective," in *Comprehensive Perspectives on Speech Sound Development and Disorders: Pathways from Linguistic Theory to Clinical Practice*, eds B. Peter and A. A. N. MacLeod (Hauppauge, NY: Nova Science Publishers, Inc).

Cao, M., Li, A., Fang, Q., Kaufmann, E., and Kröger, B. J. (2014). Interconnected growing self-organizing maps for auditory and semantic acquisition modeling. *Front. Psychol.* 5:236. doi: 10.3389/fpsyg.2014.00236

Cholin, J., Levelt, W. J. M., and Schiller, N. (2006). Effects of syllable frequency in speech production. *Cognition* 99, 205–235. doi: 10.1016/j.cognition.2005.01.009

Crawford, E., Gingerich, M., and Eliasmith, C. (2015). Biologically plausible, human-scale knowledge representation. *Cogn. Sci.* 40, 782–821. doi: 10.1111/cogs.12261

Eliasmith, C. (2013). *How to Build a Brain: A Neural Architecture for Biological Cognition*. New York, NY: Oxford University Press.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., and Tan, Y. (2012). A large-scale model of the functioning brain. *Science* 338, 1202–1205. doi: 10.1126/science.1225266

Fenson, L., Dale, P. S., Reznick, J. S., Thal, D., Bates, E., Hartung, J. P., et al. (1993). *The MacArthur Communicative Development Inventories: User's Guide and Technical Manual*. San Diego, CA: Singular Publishing Group.

Fikkert, P. (2007). "Acquiring phonology," in *Handbook of Phonological Theory*, ed. J. A. Goldsmith (Blackwell Reference).

Gervain, J., and Mehler, J. (2010). Speech perception and language acquisition in the first year of life. *Annu. Rev. Psychol.* 61, 191–218. doi: 10.1146/annurev.psych.093008.100408

Goldstein, L., Byrd, D., and Saltzman, E. (2006). "The role of vocal tract gestural action units in understanding the evolution of phonology," in *Action to Language Via the Mirror Neuron System*, ed. M. A. Arbib (Cambridge, MA: Cambridge University Press), 215–249. doi: 10.1017/cbo9780511541599.008

Grunwell, P., and Yavas, M. (1988). Phonotactic restrictions in disordered child phonology: a case study. *Clin. Ling. Phonetics* 2, 1–16. doi: 10.3109/02699208808985240

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *J. Commun. Disord.* 39, 350–365. doi: 10.1016/j.jcomdis.2006.06.013

Guenther, F. H. (2016). *Neural Control of Speech*. Cambridge, MA:MIT Press.

Guenther, F. H., and Vladusich, T. (2012). A neural theory of speech acquisition and production. *J. Neuroling.* 25, 408–422. doi: 10.1016/j.jneuroling.2009.08.006

Guenther, F. H., Ghosh, S. S., and Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain Lang.* 96, 280–301. doi: 10.1016/j.bandl.2005.06.001

Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. *Nat. Rev. Neurosci.* 8:393402.

Hickok, G., and Poeppel, D. (2016). "Neural basis of speech perception," in *Neurobiology of Language*, eds G. Hickok and S. L. Small (Cambridge, MA: Academic Press), 299–310. doi: 10.1016/b978-0-12-407794-2.00025-0

Hickok, G., Houde, J., and Rong, F. (2011). Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron* 69, 407–422. doi: 10.1016/j.neuron.2011.01.019

Kearney, E., and Guenther, F. H. (2019). Articulating: the neural mechanisms of speech production. *Lang. Cogn. Neurosci.* 34, 1214–1229. doi: 10.1080/23273798.2019.1589541

Kröger, B. J. (1993). A gestural production model and its application to reduction in German. *Phonetica* 50, 213–233. doi: 10.1159/000261943

Kröger, B. J., and Bekolay, T. (2019). *Neural Modeling of Speech Processing and Speech Learning. An Introduction*. Berlin: Springer International Publishing.

Kröger, B. J., and Birkholz, P. (2007). "A gesture-based concept for speech movement control in articulatory speech synthesis," in *Verbal and Nonverbal Communication Behaviours, LNAI 4775*, eds A. Esposito, M. Faundez-Zanuy, E. Keller, and M. Marinaro (Berlin: Springer Verlag), 174–189. doi: 10.1007/978-3-540-76442-7_16

Kröger, B. J., and Cao, M. (2015). The emergence of phonetic-phonological features in a biologically inspired model of speech processing. *J. Phonetics* 53, 88–100.

Kröger, B. J., Bafna, T., and Cao, M. (2019). Emergence of an action repository as part of a biologically inspired model of speech processing: the role of somatosensory information in learning phonetic-phonological sound features. *Front. Psychol.* 10:1462. doi: 10.3389/fpsyg.2019.01462

Kröger, B. J., Bekolay, T., Blouw, P., and Stewart, T. C. (2021). "Developing a model of speech production using the Neural Engineering Framework (NEF) and the Semantic Pointer Architecture (SPA). Proceedings of the International Seminar on Speech Production ISSP2020," in *Proceedings on the 12th International Seminar on Speech Production (ISSP2020)*, eds M. Tiede, D. H. Whalen, and V. Gracco (New Haven, CT: Haskins Press), 186–189.

Kröger, B. J., Birkholz, P., Kannampuzha, J., Kaufmann, E., and Neuschaefer-Rube, C. (2011). "Towards the acquisition of a sensorimotor vocal tract action repository within a neural model of speech processing," in *Analysis of Verbal and Nonverbal Communication and Enactment: The Processing Issues. LNCS 6800*, eds A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud, and A. Nijholt (Berlin: Springer), 287–293. doi: 10.1007/978-3-642-25775-9_27

Kröger, B. J., Crawford, E., Bekolay, T., and Eliasmith, C. (2016). Modeling interactions between speech production and perception: speech error detection at semantic and phonological levels and the inner speech loop. *Front. Comput. Neurosci.* 10:51. doi: 10.3389/fncom.2016.00051

Kröger, B. J., Kannampuzha, J., and Kaufmann, E. (2014). Associative learning and self-organization as basic principles for simulating speech acquisition, speech production, and speech perception. *EPJ Nonlinear Biomed. Phys.* 2:2.

Kröger, B. J., Kannampuzha, J., and Neuschaefer-Rube, C. (2009). Towards a neurocomputational model of speech production and perception. *Speech Commun.* 51, 793–809. doi: 10.1016/j.specom.2008.08.002

Kröger, B. J., Stille, C., Blouw, P., Bekolay, T., and Stewart, T. C. (2020). Hierarchical sequencing and feedforward and feedback control mechanisms in speech production: a preliminary approach for modeling normal and disordered speech. *Front. Comput. Neurosci.* 14:99. doi: 10.3389/fncom.2020.573554

Kuhl, P. (2004). Early language acquisition: cracking the speech code. *Nat. Rev. Neurosci.* 5, 831–843. doi: 10.1038/nrn1533

Levelt, C. C., and van de Vijver, R. (2004). "Syllable types in cross-linguistic and developmental grammars," in *Constraints in Phonological Acquisition*, eds R. Kager, J. Pater, and W. Zonneveld (Cambridge: Cambridge University Press), 204–218. doi: 10.1017/cbo9780511486418.007

Levelt, W. J. M., and Wheeldon, L. (1994). Do speakers have access to a mental syllabary? *Cognition* 50, 239–269. doi: 10.1016/0010-0277(94)90030-2

Levelt, W. J. M., Roelofs, A., and Meyer, A. S. (1999). A theory of lexical access in speech production. *Behav. Brain Sci.* 22, 1–75.

Li, P., Farkas, I., and MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *Neural Netw.* 17, 1345–1362. doi: 10.1016/j.neunet.2004.07.004

Lytle, S. R., and Kuhl, P. K. (2017). "Social interaction and language acquisition," in *Handbook of Psycholinguistics*, eds E. M. Fernández and H. S. Cairns (Hoboken, NJ: Wiley Online).

Menn, L., and Vihman, M. (2011). "Features in child phonology," in *Where do Phonological Features Come From? Cognitive, Physical and Developmental Basos if Distinctive Speech Categrories*, eds N. Clements and R. Ridouane (Amsterdam: John Benjamins Publishing Company).

Miller, H. E., and Guenther, F. H. (2021). Modelling speech motor programming and apraxia of speech in the DIVA/GODIVA neurocomputational framework. *Aphasiology* 35, 424–441. doi: 10.1080/02687038.2020.1765307

Nittrouer, S. (1995). Children learn separate aspects of speech production at different rates: evidence from spectral moments. *J. Acoust. Soc. Am.* 97, 520–530. doi: 10.1121/1.412278

Oller, D. K. (2000). *The Emergence of the Speech Capacity*. Mahwah, N. J: Lawrence Erlbaum Associates.

Priester, G. H., Post, W. J., and Goorhuis-Brouwer, S. M. (2011). Phonetic and phonemic acquisition: normative data in English and Dutch speech sound development. *Int. J. Pediatr. Otorhinolaryngol.* 75, 592–596. doi: 10.1016/j.ijporl.2011.01.027

Redford, M. A. (2019). Speech production from a developmental perspective. *J. Speech Lang. Hear. Res.* 62, 2946–2962. doi: 10.1044/2019_JSLHR-S-CSMC7-18-0130

Riecker, A., Mathiak, K., Wildgruber, D., Erb, M., Hertrich, I., Grodd, W., et al. (2005). fMRI reveals two distinct cerebral networks subserving speech motor control. *Neurology* 64, 700–706. doi: 10.1212/01.WNL.0000152156.90779.89

Saltzman, E., and Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecol. Psychol.* 1, 333–382. doi: 10.1207/s15326969eco0104_2

Schiller, N. O., Meyer, A. S., Baayen, R. H., and Levelt, W. J. M. (1996). A comparison of lexeme and speech syllables in Dutch. *J. Quant. Ling.* 3, 8–28. doi: 10.1080/09296179608590060

Stewart, T. C., and Eliasmith, C. (2014). Large-scale synthesis of functional spiking neural circuits. *Proc. IEEE* 102, 881–898. doi: 10.1109/JPROC.2014.2306061

Stoel-Gammon, C., and Dunn, C. (1985). *Normal and Disordered Phonology in Children*. Austin, TX: University Park Press.

Suprenant, A., and Goldstein, L. (1998). The perception of speech gestures. *J. Acoust. Soc. Am.* 104, 518–529.

van der Merwe, A. (2021). New perspectives on speech motor planning and programming in the context of the four- level model and its implications for understanding the pathophysiology underlying apraxia of speech and other motor speech disorders. *Aphasiology* 35, 397–423. doi: 10.1080/02687038.2020.1765306

Varley, R., and Whiteside, S. P. (2001). What is the underlying impairment in acquired apraxia of speech? *Aphasiology* 15, 39–49.

Warlaumont, A. S., and Finnegan, M. K. (2016). Learning to produce syllabic speech sounds via reward-modulated neural plasticity. *PLoS One* 11:e0145096. doi: 10.1371/journal.pone.0145096

Whiteside, S. P., and Varley, R. A. (1998). A reconceptualisation of apraxia of speech: a synthesis of evidence. *Cortex* 34, 221–231. doi: 10.1016/s0010-9452(08)70749-4

Zhang, Y., and Wang, Y. (2007). Neural plasticity in speech acquisition and learning. *Biling. Lang. Cogn.* 10, 147–160. doi: 10.1017/s1366728907002908

Conflict of Interest: TB is employed by Applied Brain Research.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Kröger, Bekolay and Cao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Human Neuroscience

SYSTEMATIC REVIEW
published: 17 May 2022
doi: 10.3389/fnhum.2022.858863

Pediatric Responses to Fundamental and Formant Frequency Altered Auditory Feedback: A Scoping Review

Caitlin Coughler[1], Keelia L. Quinn de Launay[2,3], David W. Purcell[4,5], Janis Oram Cardy[4,5] and Deryk S. Beal[2,3,6]*

[1] Graduate Program in Health and Rehabilitation Sciences, Faculty of Health Sciences, The University of Western Ontario, London, ON, Canada, [2] Bloorview Research Institute, Holland Bloorview Kids Rehabilitation Hospital, Toronto, ON, Canada, [3] Rehabilitation Sciences Institute, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada, [4] School of Communication Sciences and Disorders, Faculty of Health Sciences, The University of Western Ontario, London, ON, Canada, [5] National Centre for Audiology, Faculty of Health Sciences, The University of Western Ontario, London, ON, Canada, [6] Department of Speech-Language Pathology, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada

OPEN ACCESS

Edited by:
Xing Tian,
New York University Shanghai, China

Reviewed by:
Nicole Eva Neef,
University Medical Center Göttingen, Germany
Ewen MacDonald,
University of Waterloo, Canada

***Correspondence:**
Caitlin Coughler
ccoughle@uwo.ca

Specialty section:
This article was submitted to Speech and Language, a section of the journal Frontiers in Human Neuroscience

Received: *20 January 2022*
Accepted: *12 April 2022*
Published: *17 May 2022*

Citation:
Coughler C, Quinn de Launay KL, Purcell DW, Oram Cardy J and Beal DS (2022) Pediatric Responses to Fundamental and Formant Frequency Altered Auditory Feedback: A Scoping Review. Front. Hum. Neurosci. 16:858863. doi: 10.3389/fnhum.2022.858863

Purpose: The ability to hear ourselves speak has been shown to play an important role in the development and maintenance of fluent and coherent speech. Despite this, little is known about the developing speech motor control system throughout childhood, in particular if and how vocal and articulatory control may differ throughout development. A scoping review was undertaken to identify and describe the full range of studies investigating responses to frequency altered auditory feedback in pediatric populations and their contributions to our understanding of the development of auditory feedback control and sensorimotor learning in childhood and adolescence.

Method: Relevant studies were identified through a comprehensive search strategy of six academic databases for studies that included (a) real-time perturbation of frequency in auditory input, (b) an analysis of immediate effects on speech, and (c) participants aged 18 years or younger.

Results: Twenty-three articles met inclusion criteria. Across studies, there was a wide variety of designs, outcomes and measures used. Manipulations included fundamental frequency (9 studies), formant frequency (12), frequency centroid of fricatives (1), and both fundamental and formant frequencies (1). Study designs included contrasts across childhood, between children and adults, and between typical, pediatric clinical and adult populations. Measures primarily explored acoustic properties of speech responses (latency, magnitude, and variability). Some studies additionally examined the association of these acoustic responses with clinical measures (e.g., stuttering severity and reading ability), and neural measures using electrophysiology and magnetic resonance imaging.

Conclusion: Findings indicated that children above 4 years generally compensated in the opposite direction of the manipulation, however, in several cases not as effectively as adults. Overall, results varied greatly due to the broad range of manipulations and designs used, making generalization challenging. Differences found between age groups in the features of the compensatory vocal responses, latency of responses, vocal variability and perceptual abilities, suggest that maturational changes may be occurring in the speech motor control system, affecting the extent to which auditory feedback is used to modify internal sensorimotor representations. Varied findings suggest vocal control develops prior to articulatory control. Future studies with multiple outcome measures, manipulations, and more expansive age ranges are needed to elucidate findings.

Keywords: altered auditory feedback, speech motor control, sensorimotor learning, speech development, fundamental frequency manipulation, formant frequency manipulation

INTRODUCTION

The ability to produce speech begins shortly after birth as infants begin mapping speech sounds onto the position and movement of articulators during the babbling stage (Siegel et al., 1976; de Boysson-Bardies, 2001; Civier et al., 2010). By 3 years of age, children can speak fluently, mastering a variety of consonant and vowel sounds to form words (Coplan and Gleason, 1988). During this early development, dramatic anatomical changes occur to the shape, size, and muscles of the structures involved in speech production (Guenther, 1994; Kent and Vorperian, 1995; Kent, 1999; Callan et al., 2000). Despite these changes, children's speech remains relatively fluent through the support of speech motor control (Guenther, 1994; Callan et al., 2000). Motor actions involved in speech production are monitored, and execution errors are detected and subsequently corrected, through feedback and feedforward mechanisms (Guenther, 2006; Alsius et al., 2013). Feedback controllers use sensory information (i.e., auditory and somatosensory feedback) to monitor and adjust motor commands sent to speech production articulators (i.e., vocal tract and larynx). Feedforward controllers guide the production of motor commands by reading out previously learned motor programs, without using incoming sensory information. Speech production requires both feedback and feedforward control, with auditory feedback playing a key role.

Auditory feedback, that is, our ability to hear ourselves speak, has been shown to play an important role in the development and maintenance of intelligible speech *via* studies showing how speech acquisition is negatively impacted when hearing is impaired at birth (Oller and Eilers, 1988), as well as how speech deteriorates following loss of hearing later in life (Cowie and Douglas-Cowie, 1992). As auditory feedback informs our correct production of speech, analyzing children's speech production under altered auditory feedback can provide important information about how auditory feedback is involved in the maturing speech motor control system. In the present scoping review, the use of frequency altered auditory feedback, specifically fundamental and formant frequency manipulations, in speech production research with pediatric populations was examined. Responses to these manipulations provide key information about articulatory and vocal motor control.

Altered Auditory Feedback Paradigms

Altered auditory feedback paradigms have been used to study auditory processing, sensorimotor control, and auditory-motor integration, independent of factors such as memory, complexity, or attentional control. This paradigm has been used in adults to expand our understanding of auditory feedback. Auditory feedback plays an important role in two primary functions: (a) accommodating *vocal settings* of respiratory, laryngeal, and supraglottal systems, and (b) maintaining *articulatory settings* to preserve phonemic distinctions and intelligibility (Perkell et al., 1997; Möbius and Dogil, 2002). Fundamental frequency (f_0), whose perceptual correlate is vocal pitch, is associated with vocal control. Fundamental frequency relates to the positioning and frequency of vocal fold vibrations and is determined by the length and tension of the vocal folds (Stemple et al., 2000; Zhang, 2016). Shifted fundamental frequency results in participants hearing their own voice sound higher or lower in pitch than anticipated. Formant frequencies relate to the positioning of the lip, tongue, and jaw, or our articulation, with changes in formant frequencies resulting in different sounds (and words) being produced (Anstis and Cavanagh, 1979; Elman, 1981; Larson, 1998; Purcell and Munhall, 2006b). The first formant (F1) is inversely related to tongue height, where sounds with a higher tongue position have a lower F1. The second formant (F2) is related to tongue front or backness, where sounds closer to the front of the mouth (e.g., lips) have a higher F2. Formant frequency manipulation studies aim to shift F1 and/or F2 and measure the participants' responses. For example, if the F1 in the vowel/ɛ/is raised by approximately 200 Hz in the word "head," the auditory feedback provided to the talker would be closer to that of the word "had" with the vowel/ae/. Manipulations of speech sounds characteristics other than vowel formants have also been used to examine articulatory control. For example, a change in the first spectral moment, or frequency centroid, of sibilant fricatives (e.g., /s/) results in participants perceiving a shifted version of the fricative (e.g., closer to/ʃ/or "sh"). The effects of altering

auditory feedback of fundamental and formant frequencies has been extensively studied in adults using altered auditory feedback paradigms where these acoustic parameters are shifted in real-time and the magnitude, direction, timing and variability of compensatory responses to these shifts are studied (Burnett et al., 1998, 1997; Houde and Jordan, 1998). These responses have been examined in paradigms of *unexpected* trial shifts and predictable *sustained* shifts.

Unexpected Shift

Altered auditory feedback studies using sudden, unexpected shifts have explored how participants respond when their auditory feedback is shifted multiple times during a sustained vocalization (Behroozmand et al., 2009), at a random point during sustained vocalizations (Larson et al., 2001; Franken et al., 2018), or during a random trial (Elman, 1981; Burnett et al., 1997; Purcell and Munhall, 2006a; Tourville et al., 2008). Participants in f_0 and formant manipulated auditory feedback studies have been found to typically produce a *reflexive compensatory* response in the opposite direction of the manipulation (Burnett et al., 1997; Hain et al., 2000). These responses are usually only partial compensations to the shift (Purcell and Munhall, 2006a; Chen et al., 2007). Although manipulations in f_0 studies typically range from ±25 to 600 cents (100 cents = 1 semitone), participants on average produce response magnitudes of less than 60 cents (Burnett et al., 1997; Larson et al., 2000; Natke and Kalveram, 2001; Burnett and Larson, 2002; Donath et al., 2002; Bauer and Larson, 2003; Natke et al., 2003; Sivasankar et al., 2005; Liu and Larson, 2007). In formant manipulation studies, participants produce compensatory responses that are on average less than 30% of the total shift (Purcell and Munhall, 2006a; Tourville et al., 2008; Mitsuya et al., 2015). A second response type, where vocal productions follow the same direction as the shift, called *following* responses, has also been observed (Burnett et al., 1997, 1998; Hain et al., 2000; Larson et al., 2007). It has been suggested that these following responses occur more frequently with large magnitude shifts (i.e., in f_0 perturbations; Burnett et al., 1998). Behroozmand et al. (2012) and Franken et al. (2018) however found that most individuals who produced opposing (compensatory) responses on average, tended to also produce following responses on some trials.

Both following and compensatory responses typically show an onset latency of approximately 100–150 ms, suggesting these responses are reflexive and automatic (Tourville et al., 2008). This has been supported by findings showing that participants produce compensatory productions even when instructed to ignore any manipulations (Burnett et al., 1997; Zarate and Zatorre, 2008; Patel et al., 2014; Hu et al., 2015). Hain et al. (2000), however, found that there appear to be two responses produced in f_0 manipulations: an early automatic response that can be modulated by instruction and a later response under voluntary control. Overall, these compensatory responses are thought to provide information about an individual's auditory feedback control (Tourville et al., 2008; Cai et al., 2011). Larger response magnitudes opposing the direction of the shift are postulated to reflect greater reliance on auditory feedback, although the magnitude and direction of the applied perturbation in studies need to be taken into consideration (Heller Murray and Stepp, 2020).

Predictable Sustained Shift

In contrast, predictable, *sustained* auditory shifts are used to evaluate the updating of the feedforward system (feedforward control) through sensorimotor adaptation. In these paradigms, participants are presented with shifted auditory feedback stimuli over multiple successive trials, and gradually develop/learn an *adaptive response* to compensate for the perturbation (Houde and Jordan, 1998, 2002; Jones and Munhall, 2000, 2002; Purcell and Munhall, 2006b; Villacorta et al., 2007). In adults, these compensatory effects remain immediately following removal of the altered auditory feedback; such *adaptation* indicates a learned response in which stored motor programs have been updated (adapted) in response to the persistent compensatory productions made (Houde and Jordan, 1998, 2002; Jones and Munhall, 2000; Purcell and Munhall, 2006b). These studies typically consist of four phases: a baseline phase, where participants receive normal feedback; followed by a ramp phase, where the auditory feedback is incrementally shifted; then a hold phase, where the shifted stimuli is held at its maximum; and finally, sometimes, an end phase where the perturbations are removed. In these studies, two responses are examined: (a) how individuals' responses during the hold phase differ from their average baseline phase productions (looking for *compensation* to shifts), and (b) how individuals' productions at the end phase (when the perturbation is removed) differ from the baseline phase (looking for *adaptation*).

These adaptation paradigms provide key information about how speakers use auditory feedback, for calibration and maintenance (i.e., during hold phase) and to incorporate long-term changes in their speech production (i.e., during end phase; Houde and Jordan, 1998; Purcell and Munhall, 2006b; MacDonald et al., 2010). Similar to the reflexive responses to sudden perturbations, participants' responses are typically in the opposite direction of the manipulation, with some responses following the perturbations (Houde and Jordan, 1998, 2002; Jones and Munhall, 2000, 2005; Purcell and Munhall, 2006b; Villacorta et al., 2007). These responses also only partially compensate for the total perturbation magnitude (Jones and Munhall, 2000, 2005; Houde and Jordan, 2002; Purcell and Munhall, 2006b; Villacorta et al., 2007; MacDonald et al., 2010). Katseff et al. (2012) suggested that based on findings that individuals showed greater compensation for small F1 perturbations than for larger perturbations, auditory feedback may play a larger role in small discrepancies. These responses have also been found to be automatic, occurring when participants are instructed to ignore manipulations (Munhall et al., 2009; Keough et al., 2013).

Across studies exploring responses to sudden shifts, there is consensus that these responses describe *compensation*. However, within the sensorimotor adaptation literature, inconsistencies persist. In some studies, responses produced when the perturbation is held at its maximum (hold phase trials) are described using the term *compensation,* while in others, these trials are referred to as *adaptation.* Although responses in

these trials are thought to gradually reflect updating of motor commands and hence adaptation to the shift, within this article, the term *compensation* will be used to describe responses within the hold phase of a sensorimotor adaptation paradigm, as these productions represent both compensation and adaptation. Productions made following removal of shifts (during end phase trials) will be described as *adaptation* (also known in the literature as after-effects).

Relations Between Unexpected and Sustained Perturbation

Examining responses to unexpected and sustained perturbations provides important information about feedback and feedforward control. Contrasting participants' responses to sudden and sustained F1 manipulations, Franken et al. (2019) and Raharjo et al. (2021) found that individuals' responses in the sudden vs. sustained conditions were not correlated with each other. In contrast, Lester-Smith et al. (2020) explored reflexive compensatory and adaptive responses to F1 and f_0 perturbation. Participants showed similarities in their reflexive and adaptive responses, where individuals with larger reflexive responses to sudden F1 perturbation also showed larger adaptive responses to predictable F1 manipulations. However, reflexive and adaptive responses to f_0 manipulated auditory feedback were not related. This highlights that differences may not only be evident in the mechanisms underlying responses in sudden (reflexive) and sustained (adaptive) perturbation paradigms, but also in control of articulatory and vocal settings. Although responses to fundamental and formant frequency altered auditory feedback have extensively been studied in adults, a contrastive look at responses in children has not previously been examined. Investigating how and if these responses differ developmentally will help shed light on underlying mechanisms and improve our understanding of speech motor control.

Models of Speech Motor Control

Prominent models of speech motor control have strived to model how we regulate our speech production. The *directions into velocities of articulators* (DIVA) model uses auditory and somatosensory feedback control combined with feedforward control to maintain fluent speech (Guenther, 1995, 2016; Guenther et al., 1998; Tourville and Guenther, 2011). Mismatches in the feedback control systems between the actual and expected sensory state are used to form corrective motor commands (Tourville and Guenther, 2011; Guenther and Vladusich, 2012). In a sustained shift condition, over multiple corrective motor commands, these adjustments are used to update the feedforward command. In this way, the DIVA model postulates that similar mechanisms are employed in response to sudden and sustained perturbations. In contrast, the *state feedback control* (SFC) model postulates that responses to sudden and sustained perturbations are driven by different mechanisms. In the SFC model, sensory feedback can be used directly to update the internal model estimate of the dynamical state of the vocal tract (Houde and Nagarajan, 2011; Houde et al., 2013). Thus, unlike the DIVA model, the SFC model does not require the integration of corrective motor commands into feedforward control in order to accommodate adaptation.

Neurophysiology and Neuroimaging Association

Behavioral data from altered auditory feedback paradigms provide information about the final product of the manipulation, the vocal response to the shift. This data however, does not elucidate what may be contributing to differences in these responses. Examining neural activation and neural structure provide key information about how the brain processes stimuli leading to the final vocal production. Neurophysiology (e.g., EEG) and neuroimaging (e.g., MRI) are useful in conjunction with altered auditory feedback paradigms, however, it is important to take into consideration potential limitations of these techniques. While a comprehensive review of potential caveats that might hamper the interpretability of these techniques is out of the scope for this article, one of the biggest challenges to consider when using these techniques with altered auditory feedback is filtering out activation that occurs as a result of motor movement from spoken productions.

Neurophysiological and neuroimaging data have been instrumental in informing models of speech motor control. Using neuroimaging (i.e., functional magnetic resonance imaging [fMRI]), individual components of the DIVA model have been mapped onto brain regions based on experimental neuroimaging findings (Bohland and Guenther, 2006; Ghosh et al., 2008; Tourville et al., 2008; Golfinopoulos et al., 2011). Examining typical neural regions of activation in response to altered auditory feedback (i.e., using fMRI), as well as structural similarities (i.e., using diffusion-weighted imaging), provides important information relating to typical and atypical productions, expanding our understanding of neural correlates related to feedback and feedforward control. While MRI provides excellent spatial resolution, it has low temporal resolution as it measures changes in blood flow.

Electroencephalography (EEG) in contrast has high temporal resolution and low spatial resolution, making it an ideal methodology to examine the timing of neural responses, which is particularly important given the quick pace of speech. Neurophysiological activity measured through EEG responses to auditory stimuli provides important information about the processing of auditory input, expanding on behavioral findings (Hillyard and Picton, 1978). Common event related potentials observed in response to auditory stimuli are characterized by a positive-negative-positive sequence, the P1-N1-P2 complex. The initial positive peak (P1) is approximately 30–110 ms after stimulus onset, followed by a negative peak (N1) approximately 80–150 ms after stimulus onset, and a final positive peak (P2) 140–160 ms after stimulus onset (Ponton et al., 2000). Developmentally, latency of the P1 and N1 components has been found to negatively correlate with age in response to speech and non-speech stimuli (Polich et al., 1990; Kraus et al., 1993; Tonnquist-Uhlen et al., 1995; Cunningham et al., 2000; Ponton et al., 2000; Wunderlich and Cone-Wesson, 2006), whereas the latency of the P2 component has not been found to significantly

vary with age (Ponton et al., 2000; Fitzroy et al., 2015). In terms of amplitude, P1 has been found to decrease with age in response to speech and non-speech stimuli (Kraus et al., 1993; Sharma et al., 1997; Cunningham et al., 2000; Fitzroy et al., 2015), whereas N1 and P2 amplitudes are less consistently found to change developmentally (Sharma et al., 1997; Wunderlich and Cone-Wesson, 2006; Fitzroy et al., 2015).

During EEG altered auditory feedback studies in adults, increased activity has been found in the P1-N1-P2 complex (Heinks-Maldonado et al., 2006; Behroozmand et al., 2011). Amplitudes of the N1 and P2 components have been found to be correlated with the magnitude of perturbations, whereas the amplitude of the P1 component has been found to increase in a non-specific manner (Liu et al., 2011; Scheerer et al., 2013a). Based on these patterns of response, it has been theorized that P1 represents a general recognition of a mismatch between expected and actual auditory feedback, whereas N1 is related to the determination of whether feedback is internally or externally generated, and P2 represents processing of the size of the mismatch (Scheerer et al., 2013a). As the P1-N1-P2 complex has been associated with age-related changes during auditory processing, exploring differences in neurophysiological activity during altered auditory feedback paradigms provides an additional avenue for expanding our understanding of developmental differences in the use of auditory feedback.

Speech, Language, Auditory Feedback, and Clinical Populations

Speech and language processes are interactive and influence each other throughout development, and examining their interaction can provide key developmental information (Kent, 2004; Smith and Goffman, 2004; Terband and Maassen, 2010; Strand et al., 2013). Reading, in turn, builds on these speech and language skills (Mattingly, 1972; Liberman, 1989; Rueckl et al., 2015). As such, speech motor control impairments have been documented in children with speech sound disorders (SSD; Namasivayam et al., 2013), individuals who stutter (Bloodstein and Bernstein-Ratner, 2008), individuals with dyslexia (van den Bunt et al., 2017), and children with developmental disorders that often include co-occurring language impairments such as autism spectrum disorder (ASD; Belmonte et al., 2013). Exploring the differences in the integration of auditory information during speech in children with speech and language disorders could provide more insight into the mechanisms that typically developing children use to respond and process this feedback.

While auditory feedback is considered important for the development of speech motor control, the means by which children use this feedback to establish and refine their internal sensorimotor representations and to control online speech production remain relatively unknown. Specifically, determining children's capacity to integrate auditory information into upcoming motor commands is essential to better understanding the role of auditory feedback in the acquisition and refinement of speech production, as well as the mechanisms that govern compensation for changes in auditory feedback throughout development. The purpose of this scoping review is to explore the current use of frequency altered auditory feedback paradigms in pediatric speech research, and investigate how the integration of auditory information during speech changes across development, through an examination of behavioral responses to auditory perturbation in pediatric populations. This is essential for ultimately understanding the mechanisms underlying the acquisition of speech motor control.

METHODS

Objectives and Rationale

A scoping review of published studies was conducted to identify existing articles that have used frequency altered auditory feedback paradigms with children. The structured framework presented by Arksey and O'Malley (2005) and further developed by Levac et al. (2010) was utilized. The aim of our scoping review was to identify and analyze the current state of research for pediatric responses to frequency altered auditory feedback in order to examine how the research has been conducted, clarify key concepts, summarize the current evidence, and identify gaps in the existing research. Research questions included: (a) what were the characteristics of the children included in these studies (e.g., age ranges, clinical populations, country, language); (b) how many studies included a manipulation of fundamental frequency, formant frequency, or both; (c) are differences in responses to altered auditory feedback evident across developmental stages and clinical populations, and (d) what methodological designs have been used with children?

Search Strategy

The literature was searched for publications up until October 19th, 2021. Searches were conducted in six academic databases: CINAHL, Embase, Medline (*via* Ovid), PsycINFO (*via* Proquest), Scopus, and Web of Science. A standardized list of keywords, as well as database subject headings (MeSH) for the concepts of altered auditory feedback and pediatric were developed (see **Supplementary Material 1**). All six databases were searched using keywords. MeSH searches were combined with their respective keyword searches (altered auditory feedback or pediatric) for four databases (CINAHL, Embase, Medline and PsycInfo). **Supplementary Material 2** shows a sample of how the search was conducted in PsycInfo. Following searches of all databases, citations were uploaded to the referencing software, Covidence (Covidence Systematic Review Software, 2020).

Inclusion Criteria

A PRISMA flow chart showing the systematic selection of articles for inclusion is provided in **Figure 1**. Titles and abstracts were screened for inclusion in the full-article review using three criteria: (a) real-time perturbation of frequency auditory input was used (i.e., formant or fundamental frequency), (b) an analysis of immediate effects on speech in response to the perturbation (e.g., compensatory response) was included, and (c) typically developing (TD) and/or clinical participants between the ages of 2–18 years were included in the sample. Articles that were

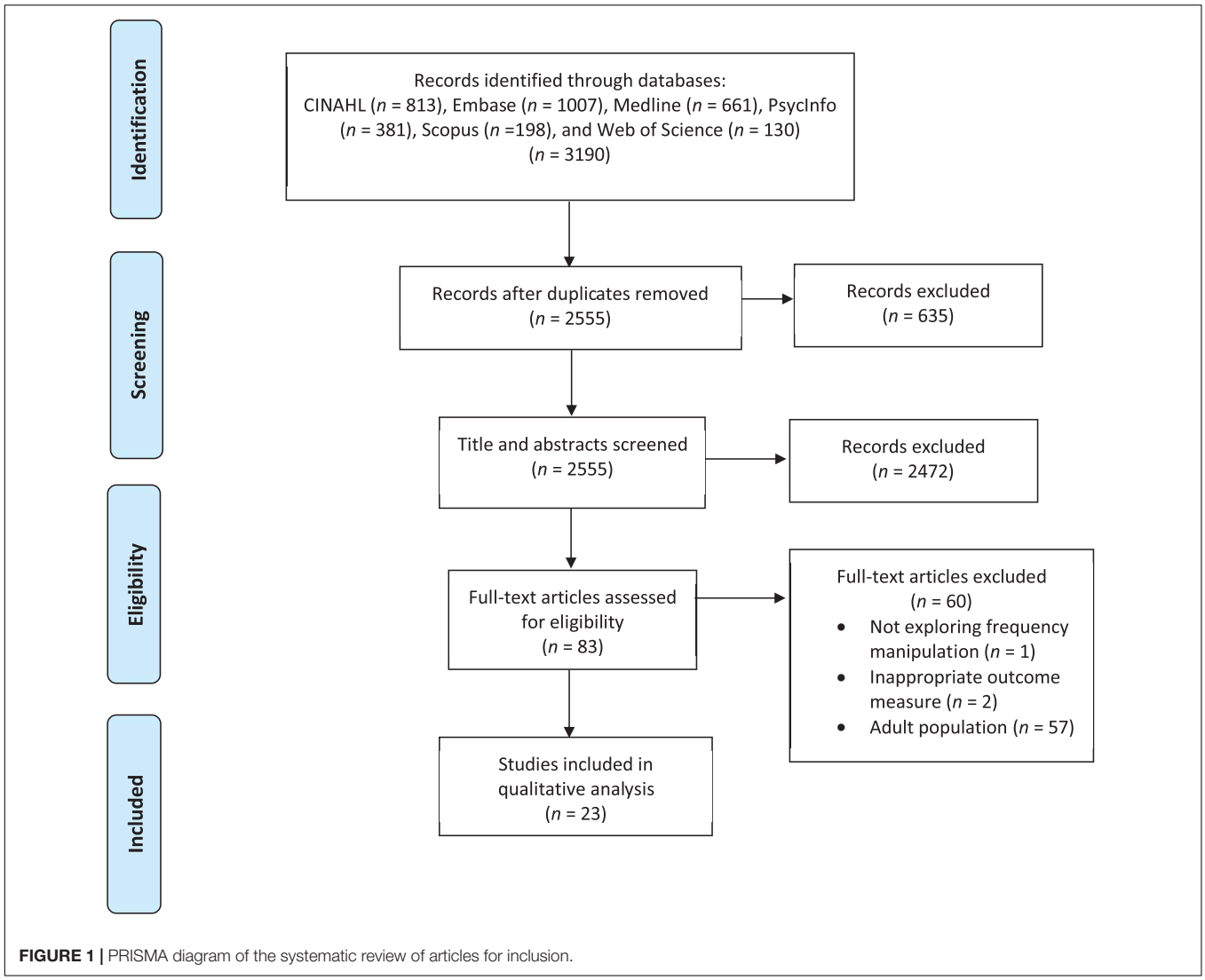


FIGURE 1 | PRISMA diagram of the systematic review of articles for inclusion.

not experimental studies (e.g., commentaries) were excluded, with the exception of one review article that was included because it introduced a relevant case study. Titles and abstracts were additionally screened by an independent graduate student to ensure no selection bias was present. Inter-rater reliability showed 97% agreement, with a Cohen's Kappa of 0.62, showing substantial agreement. During full-text review, articles that did not separate participants under 18 years old from adults in the analysis were additionally excluded (n = 57). Articles were also excluded due to not exploring frequency manipulated altered auditory feedback (n = 1) or analyzing compensatory and/or adaptation responses (n = 2).

Data Extraction

Complete records of data extracted from all articles can be found in **Supplementary Material 3**.[1] Information extracted included: (a) article information (title, authorship, year, country conducted in), (b) participant characteristics (age range, sample size, language spoken), (c) primary aim of study, (d) speech and language measures collected, (e) perturbation information (e.g., pitch or formant manipulation, and direction and magnitude), (f) methodological design (e.g., number of trials in each phase), and (g) main outcomes.

FINDINGS

Search Outcome

Twenty-three articles were identified as meeting inclusion criteria across the six databases. Of these articles, manipulations included: nine of fundamental frequency, twelve of formant frequency, one of the frequency centroid of fricatives (henceforth grouped as formant manipulation), and one of both fundamental and formant frequencies. Of these studies, nine looked at clinical populations in addition to TD children (three f_0, seven formant[2]).

[1] https://osf.io/p9jz8/

[2] One study included both fundamental and formant frequency manipulations.

All included studies examined behavioral responses to altered auditory feedback, with four also exploring neurophysiological responses (two electrophysiological, one diffusion-weighted imaging, and one resting-state functional connectivity). See **Supplementary Material 3** (see Footnote 1) for an in-depth description of our findings.

Research Context and Participant Characteristics

The majority of the studies reviewed were conducted in Canada (ten total; four f_0 manipulation, six formant manipulation), followed by the United States (seven total [see Footnote 2]; four f_0, four formant), the Netherlands (three formant), and China (two f_0), and one study which collected children in the United States and the Netherlands (formant). The majority of participants were English speakers (16 total [see Footnote 2]; eight f_0 and nine formant), followed by Dutch (three formant), Mandarin Chinese (two f_0), and French (one formant). All studies included pediatric populations under 18 years with the exception of one clinical population group that included individuals up to 18.3 years of age (as well as TD children under 18 years). Six f_0 manipulation studies and seven formant manipulation studies also included adult populations.

Methodological Designs and Findings

Methodological designs used with children were divided into those relating to vocal control (i.e., f_0 manipulation) and those relating to articulatory control (i.e., formant manipulations) in order to explore potential influences of these differing paradigms on response outcomes, as well as potentially differing developmental trajectories.

Fundamental Frequency Manipulation

All ten of the studies exploring responses to f_0 altered auditory feedback involved unexpected (within trial) perturbations. Two studies also contrasted responses to unexpected perturbations with sustained (predictable) manipulations (Scheerer et al., 2016; Heller Murray and Stepp, 2020). In terms of manipulations applied, all ten of the studies included a negative manipulation of one semitone (−100 cents). Five of the studies also included a positive manipulation of one semitone (+100 cents), and two studies included additional magnitude manipulations either in negative (−50, −200 cents; Liu et al., 2013) or negative and positive directions (±50, 200, and 500 cents; Liu P. et al., 2010). **Table 1** includes a summary of findings, sample size, age range and manipulations of f_0 manipulation studies.

Formant Frequency Manipulation

All fourteen studies exploring responses to formant manipulated auditory feedback explored responses to sustained perturbations over several trials. **Table 2** includes a summary of findings, sample size, age range, and formant manipulations, with more in-depth findings available in **Supplementary Material 3** (see Footnote 1). Generally, studies included a baseline phase (ranging from 10–50 trials, $M = 23.0$) followed by a ramp phase where the perturbation was gradually introduced (ranging from 10–60 trials, $M = 25.4$), and a hold phase where the perturbation was held at its maximum (ranging from 18–120 trials, $M = 41.0$). Ten studies also included an end phase where the perturbation was removed (ranging from 10–40 trials, $M = 20.15$). One study additionally included a ramp-down phase where the perturbation was gradually removed following the hold phase (van den Bunt et al., 2018b).

In terms of magnitude and direction of manipulations, eight studies manipulated F1 and F2 values of vowels, and five studies manipulated F1 only. F1 was manipulated in various ways, including: increased by 25% (Shiller and Rochon, 2014; van den Bunt et al., 2018a), 175 Hz (Shiller et al., 2010b), or 340 Hz, or decreased by 230 Hz (Coughler et al., 2021), or manipulated individually so the maximum perturbation represented a change from/ɛ/to/ae/ (Ohashi and Ostry, 2021). One study manipulated the frequency centroid of fricatives (decreased by 3 semitones; Shiller et al., 2010a). Manipulations of F1 and F2 were language dependent. In Dutch, this manipulation included an F1 increase of 25% and an F2 increase or decrease by 12.5% (Terband et al., 2014; van Brenk and Terband, 2020 respectively), in French F1 was increased by 27% and F2 decreased by 10% (Caudrelier et al., 2019), and in English F1 was increased by 200 Hz or 25% and F2 was decreased by 250 Hz or 12.5% (MacDonald et al., 2012; Daliri et al., 2018; Demopoulos et al., 2018). One study (van den Bunt et al., 2018b) individualized the manipulation so the maximum perturbation meant a change from/ɪ/to/ɛ/for each participant. Kim et al. (2020) manipulated F1 and F2 upward 335 cents (adult population received manipulation of ±250 cents).

Results across both fundamental and formant frequency manipulations found age-dependent developmental trajectories related to response latencies, magnitude of compensatory responses, variability and perceptual abilities, as well as relationships of compensation with literacy abilities. These findings as well as clinical and neurophysiological/neuroimaging findings are discussed below.

Response Latencies

Fundamental Frequency Manipulation

Four f_0 studies compared response latencies, the onset of the compensatory response to altered auditory feedback, across age groups and found that children consistently demonstrated longer response latencies to perturbations in auditory feedback than adult populations (Liu H. et al., 2010; Liu P. et al., 2010; Liu et al., 2013; Scheerer et al., 2013b). Two of these studies used multiple child age groups to explore potential age gradients within responses (Liu et al., 2013; Scheerer et al., 2013b). Scheerer et al. (2013b) found a main effect of age, where three of the four age groups under 18 (4–6, 7–10, and 11–13 years) independently demonstrated longer response latencies than the 18–30-year-olds (14–17-year-olds did not). Similarly, Liu et al. (2013) found their younger children (ages 10–12) had significantly longer response latencies than the adult group, however, older children (ages 13–15) did not differ from adults.

Response Latencies Summary

The finding of significantly longer response latencies for children compared to adults in the compensatory responses

TABLE 1 | Behavioral findings in response to fundamental frequency (f_0) manipulated altered auditory feedback in typically developing (TD) children and children with speech and language disorders.

Study	Child sample size	Child age range	Manipulation	Findings	Adult contrast	RL	VV	CF	NF
Scheerer et al. (2020a)	(1) *n* = 11 (2) *n* = 9	(1) 24–35 months (2) 40–46 months	"baa" Within trial ±100 cents	• Both groups of toddlers compensated to the perturbation • No main effect of age found			X		
Scheerer et al. (2016)	*n* = 25	3.0–8.0 years	/a/ (1) Within trial (unpredictable) −100 cents (2) Sustained (predictable) −100 cents	• Children showed compensation in both designs but *smaller responses* than adults • Children *more variable*	X		X		
Scheerer et al. (2020b)	(1) *n* = 45 (2) *n* = 30*	3.0–13 years	/a/ Within trial ±100 cents	• Autistic children had *shorter response latencies* • Both autistic and TD children compensated in opposite direction of shift (similar in magnitude and variability)			X	X	
Scheerer et al. (2013b)	*n* = 80 children (10 M, 10 F per group)	(1) 4–6 years (2) 7–10 years (3) 11–13 years (4) 14–17 years	/a/ Within trial −100 cents	• Younger children had *longer response latencies* • Children 4–6 years *more variable* than adults • No significant interaction of age and sex on response magnitude	X	X	X		X
Heller Murray and Stepp (2020)	*n* = 20	6.6–11.7 years	/ɑ/ (1) Within trial shift ±1 ST (2) Sustained shift ±1 ST	(1) Opposing responses only: children with less sensitive pitch discrimination (C-L; >2 *SD* from adults) showed significantly *larger* responses than adults or children with adult-like pitch discrimination (C-A) (2) C-L had *smaller* vocal response magnitudes than C-A and adults	X		X		
Liu P. et al. (2010)	*n* = 19	7.0–12.0 years	/u/ Within trial ±50, 100, 200, 500 cents	• Children showed significantly *larger* compensatory responses to adults • Children produced *longer latencies* than adults	X	X			
Liu H. et al. (2010)	*n* = 10	7.0–12.0 years	/a/ Within trial −100 cents	• Older adults produced significantly larger response magnitudes than children and young adults • Children produced significantly *longer latencies* than younger and older adult groups	X	X			
Russo et al. (2008)	(1) *n* = 19 (2) *n* = 18*	(1) 7.0–12.0 years (2) 7.0–12.0*	/a/ Within trial −100 cents	Subset of children with ASD produced larger responses than TD children			X	X	
Liu et al. (2013)	(1) *n* = 22 (2) *n* = 20	(1) 10–12 years (2) 13–15 years	/u/ Within trial −50, −100, or −200 cents	Younger children elicited *longer latency* vocal response than young adults	X	X			X
Demopoulos et al. (2018)	(1) *n* = 11 (2) *n* = 12**	(1) 10.3–15.4 years (2) 8.3–18.3** years	"ah" Within trial ±100 cents	Children with 16p11.2 deletion showed larger pitch compensation compared to controls				X	

**Refers to children with Autism Spectrum Disorder (ASD), and*
***refers to children with 16p11.2 deletion.*
RL, response latencies; VV, vocal variability; CF, clinical findings; NF, neuroimaging findings.

TABLE 2 | Behavioral findings in response to formant manipulated altered auditory feedback in typically developing (TD) children and children with speech and language disorders.

Study	Child sample size	Child age range	Manipulation	Findings	Adaptation	Adult contrast	VV	CF	NF
MacDonald et al. (2012)	(1) *n* = 20 (2) *n* = 26	(1) 23–35 months (2) 43–59 months	/ɛ/ F1 increased by 200 Hz and F2 decreased by 250 Hz	• Young children compensated in opposite direction of perturbation, but toddlers did not • No significant difference in compensation in adults and young children • Variability decreased with age		X	X		
Kim et al. (2020)	(1) *n* = 8 (2) *n* = 8 (3) *n* = 8[a] (4) *n* = 8[a]	(1) 3.75–6.83 years (2) 7.25–9.33 years (3) 3.50–6.83[a] years (4) 7.08–9.33[a] years	"buck," "bus," "puck," "pup," "cut," "cup," "gut,", "duck" Upward shift of 335 cents (gradual with ramp and without ramp conditions)	• TD children had similar compensation to TD adults • Both younger and older children who stutter did not show compensation (in either condition)		X		X	
Terband et al. (2014)	(1) *n* = 17 (2) *n* = 11[b]	(1) 4.1–8.7 years (2) 3.9–7.5[b] years	/lː/ F1 increased by 25% and F2 decreased by 12.5%	• Children with SSD followed the perturbation in F1 during hold and end phase • TD children compensated in F1 and F2 and showed trend of adaptation in F1 in end phase	X			X	
Caudrelier et al. (2019)	(1) *n* = 29 (2) *n* = 24	(1) 4–5 years (2) 7–8 years	/e/ F1 increased by 27% and F2 decreased by 10%	• Some preschoolers and school-aged children compensated for the perturbation • No significant difference between groups • Adaptation magnitude similar across age groups	X	X			
van Brenk and Terband (2020)	*n* = 23	4.0–8.7 years	/lː/ F1 increased by 25% and F2 increased by 12.5%	• In F1, children showed stronger compensation and adaptation response than adults • In F2, children showed a compensation but no adaptation response • In F1 and F2, children showed higher token-to-token variability than adults	X	X	X		
Shiller and Rochon (2014)	*n* = 22 (Exp and Sham groups)	5.0–7.0 years	/ɛ/ F1 increased by 25%	• Both Exp and Sham group compensated to perturbation • Following perceptual training, Exp group showed increased magnitude compensation (Sham group showed no change) • Change in F1 persisted after removal of manipulation	X				
van den Bunt et al. (2018a)	US: *n* = 96 NL: *n* = 148	preschool – grade 2 (∼5–8 years)	/ɛ/ F1 increased by 25%	Significantly stronger compensation in hold and end phase for literate children relative to preliterate children	X		X		
Ohashi and Ostry (2021)	*n* = 19	5–12 years	/ɛ/ F1 increased to make/ae/average 23.9 ± 1.59% (SE)	Children showed similar compensation to adults, adaptation in children remained longer than adults	X	X	X		X
Coughler et al. (2021)	(1) *n* = 16 (2) *n* = 16[c]	(1) 6.83–11.68 years (2) 7.83–13.2 years	/ɛ/ (1) F1 increased by 340 Hz (2) F1 decreased by 230 Hz	Children with DLD showed greater compensation in the positive F1 manipulation condition and compensated less than TD children in the negative shift condition	X		X	X	
Shiller et al. (2010b)	*n* = 1[b]	6.5[b] years	/ɛ/ F1 increased by 175 Hz	• Compensated to perturbation • Adaptation seen following removal of manipulation	X			X	
Daliri et al. (2018)	(1) *n* = 20 (2) *n* = 20[a]	(1) 7.08–11.42 years (2) 6.08–11.17[a] years	/ɛ/ F1 increased by 25% and F2 decreased by 12.5%	• Both children groups compensated to F1 perturbation but not F2 perturbation • No significant difference between adults and children who do not stutter for F1 or F2 perturbation • Children who stutter compensated more than adults who stutter for F1 perturbation	X	X		X	

(Continued)

TABLE 2 | (Continued)

Study	Child sample size	Child age range	Manipulation	Findings	Adaptation	Adult contrast	VV	CF	NF
Shiller et al. (2010a)	$n = 11$	9.4–11.3 years	/s/ frequency centroid decreased by 3 semitones (averaging −1222 Hz)	Children showed compensatory response of similar magnitude to adults (no significant difference)		×	×		
van den Bunt et al. (2018b)	(1) $n = 10$ (2) $n = 27$[d]	10.0–13.0 years	F2-F1 manipulation individualized: from/ɪ/to/ɛ/at maximum perturbation	• All participants compensated in opposite direction of manipulation • Children with dyslexia showed weaker return to baseline during ramp-down phase than typically reading children	×			×	×
Demopoulos et al. (2018)	(1) $n = 11$ (2) $n = 12$[e]	(1) 10.3–15.4 years (2) 8.3–18.3[d] years	/ɛ/ F1 increased by 200 Hz and F2 decreased by 250 Hz	Control children showed significantly greater compensation than children who were 16p11.2 deletion carriers				×	

[a]*Refers to children who stutter.*
[b]*Refers to children with a speech sound disorder (SSD).*
[c]*Refers to children with developmental language disorder (DLD).*
[d]*Refers to children with dyslexia.*
[e]*Refers to children with 16p11.2 deletion.*
VV, vocal variability; CF, clinical findings; NF, neuroimaging findings.

in pitch-shifted paradigms is consistent with developmental trends. Neurophysiological response latencies (i.e., event-related potential latencies) are considered an objective measure of the speed of neural integration and activity, reflecting the efficiency of information processing and the synaptic density in the auditory cortex, where shorter latencies reflect faster integration of auditory information (Eggermont, 1988; Kotecha et al., 2009). Vocal response latencies were similarly found to relate to maturational changes. Integration of rapid information in adult-like auditory processing may therefore be due to increased velocity and efficiency of neural conduction and intercortical communication in gray and white matter of the cortex.

Response Magnitudes

Fundamental Frequency Manipulation

All f_0 manipulation studies explored compensatory responses to pitch perturbations, and generally found children compensated in the opposite direction of the shift. Following responses were examined in three studies (Russo et al., 2008; Liu P. et al., 2010; Scheerer et al., 2020b), with two studies excluding participants who followed the perturbation (Scheerer et al., 2016; Demopoulos et al., 2018). Results exploring the magnitude of compensation responses to unexpected perturbations were mixed. When contrasting across age groups, children were found to have: (a) reduced magnitude responses compared to adults (Liu H. et al., 2010; Scheerer et al., 2016), (b) increased magnitude responses compared to adults (Liu P. et al., 2010; Heller Murray and Stepp, 2020), (c) increased responses that followed the manipulation compared to adults (i.e., following responses; Liu P. et al., 2010), and (d) no effect of age across childhood (Liu et al., 2013; Scheerer et al., 2013b, 2020a) or compared to adults (Scheerer et al., 2013b; Heller Murray and Stepp, 2020). Heller Murray and Stepp (2020) found when analyzing opposing responses that only children with less sensitive pitch discrimination showed significantly larger responses, compared to adults and children with adult-like pitch discrimination. Liu et al. (2013) found an effect of sex, where male speakers produced larger response magnitudes than female speakers. Findings from Russo et al. (2008), Demopoulos et al. (2018), and Scheerer et al. (2020b) are described below in the *Clinical Findings* section.

The two studies exploring sustained perturbation found children showed smaller compensatory responses compared to adults (Scheerer et al., 2016; Heller Murray and Stepp, 2020). Specifically, Heller Murray and Stepp (2020) found children with less sensitive pitch discrimination produced smaller compensatory responses compared to adults and children with adult-like pitch discrimination. The magnitude of responses produced to a sustained shift was negatively correlated with the magnitude of responses to a sudden shift (Heller Murray and Stepp, 2020). In contrast, Scheerer et al. (2016) found that children produced smaller compensatory responses compared to adults in both sudden and sustained pitch shift conditions, however, these responses were not examined for correlation. Scheerer et al. (2016) also explored adaptation in the end phase, finding magnitudes of responses following removal of pitch perturbation did not differ between children and adults.

In general, these findings provide evidence supporting the DIVA model, where responses to sudden and sustained shifts are not considered separate processes (Guenther, 2006; Guenther and Vladusich, 2012).

Formant Frequency Manipulation

All formant frequency manipulation studies explored compensatory responses, and overall found typically developing children generally showed compensation in the opposite direction of the perturbation in hold and end phases. Two studies examined following responses (Terband et al., 2014; van Brenk and Terband, 2020). Seven studies contrasted child and adult responses to formant manipulated altered auditory feedback (Shiller et al., 2010a; MacDonald et al., 2012; Daliri et al., 2018; Caudrelier et al., 2019; Kim et al., 2020; van Brenk and Terband, 2020; Ohashi and Ostry, 2021). Across the studies, children showed: (a) stronger compensation and adaptation responses in F1 than adults (van Brenk and Terband, 2020), (b) similar magnitude compensation to adults (Shiller et al., 2010a; MacDonald et al., 2012; Daliri et al., 2018; Kim et al., 2020; Ohashi and Ostry, 2021), and (c) no age effect in compensation or adaptation across childhood (Caudrelier et al., 2019). MacDonald et al. (2012) found young children showed similar compensation to adults, however, children under 4 years of age showed no compensatory response. Shiller and Rochon (2014) found after a period of perceptual training, children showed increased magnitude of compensation.

Response Magnitude Summary

Based on the underlying mechanisms being examined, response magnitudes for unexpected and sustained shifts were analyzed separately. Unexpected shifts, used to explore an individual's reliance on auditory feedback control, examined in f_0 manipulation studies, elicited mixed results, ranging from reduced magnitude to increased magnitude compensatory responses compared to adults to no age effect. One potential reason for these mixed findings could be due to proximity of shifted trials. As discussed in Cai et al. (2012), cross-trial adaptation effects have been found where participants' early productions within a trial contain compensation responses to the perturbation of the previous trial.

By contrast, sustained shifts, used to explore the updating of feedforward control, generally showed no age effect after 4 years of age in formant manipulation studies. The lack of age effect suggests that children are using feedforward control similar to adults. In contrast, in f_0 sustained shift studies, children exhibited smaller magnitude compensatory responses compared to adults. This smaller compensation response may indicate a greater reliance on sensory feedback, with reduced weighting on feedforward control.

Adaptive responses, when perturbations were removed, showed mixed results ranging from stronger adaptive responses compared to adults in formant manipulation studies, to no age effect across childhood in formant or f_0 manipulation paradigms. Although mixed, these findings of the presence of adaptation responses show that children used the altered auditory feedback to update their sensorimotor mappings for future vocalizations.

Contrasting pitch and formant manipulation paradigms, clear developmental differences are seen in the youngest ages where children appear to be using their auditory feedback to manipulate their vocal productions. Scheerer et al. (2020a) found children as young as 2 years of age compensated to f_0 shifted stimuli, whereas MacDonald et al. (2012) did not find compensatory responses in children under 4 years of age to formant shifted stimuli. This lack of compensation suggests that the ability to adaptively regulate measures of vocal control (i.e., f_0) arises before control over measures of articulatory settings (i.e., formants). However, further research is required to confirm this assumption, as the number of studies examining responses in children under 3 years of age is restricted to these two studies.

Vocal Variability and Perceptual Abilities

As variability in both the motor and perceptual system play an important role in feedback and feedforward control these correlates were examined together.

Fundamental Frequency Manipulation

Five studies explored differences in vocal variability in f_0, contrasting baseline standard deviation in vocal productions across age groups (Russo et al., 2008; Scheerer et al., 2013b, 2016, 2020a; Heller Murray and Stepp, 2020). Four of these studies contrasted children with adults, finding children consistently showed more variability than adults (Scheerer et al., 2013b, 2016, 2020a). Heller Murray and Stepp (2020) found children with less sensitive pitch discrimination had significantly higher variability at baseline than both the children with adult-like pitch discrimination and adults. Children with less sensitive pitch discrimination also showed larger response magnitudes to unexpected pitch shifts and smaller responses to sustained pitch shifts compared with adults and children with adult-like pitch discrimination. Baseline vocal variability was also found to positively correlate with the magnitude of responses to unexpected perturbations, and negatively correlate with the magnitude of responses to sustained perturbations (Heller Murray and Stepp, 2020). Through regression analyses, Scheerer et al. (2013b) found that vocal variability accounted for a significant amount of the variance in the magnitude of the compensatory responses. Scheerer et al. (2016), however, did not find vocal variability correlated with the magnitude of compensatory responses. In further exploration of electrophysiological correlates, Scheerer et al. (2013b) found that age and vocal variability were significant predictors of N1 amplitude.

Formant Frequency Manipulation

Six studies explored variability of baseline vocal productions related to articulatory control (Shiller et al., 2010a; MacDonald et al., 2012; van den Bunt et al., 2018a; van Brenk and Terband, 2020; Coughler et al., 2021; Ohashi and Ostry, 2021). Generally, children showed greater variability compared to adults (Shiller et al., 2010a; MacDonald et al., 2012; van Brenk and Terband, 2020; Ohashi and Ostry, 2021). MacDonald et al. (2012) found that variability decreased with age. Similarly, van den Bunt et al. (2018a) found literate

children who read more non-words per minute showed less variation in vowel production. Coughler et al. (2021) found increased variability negatively correlated with the amount of compensation in TD children, whereas Ohashi and Ostry (2021) did not find variability correlated with the amount of compensation in children or adults.

Five studies additionally examined perceptual abilities related to articulatory control (i.e., discrimination; Shiller et al., 2010a,b; Shiller and Rochon, 2014; Terband et al., 2014; Coughler et al., 2021). Coughler et al. (2021) found F1 discrimination thresholds did not significantly correlate with percent compensation in the positive or negative condition, or with language scores in TD children or children with a specific deficit in language. Shiller and Rochon (2014) found that productions following an experimental block of relevant (to the formant frequency shift) auditory-perceptual training resulted in enhanced compensatory responses in children. Based on results from a phoneme identification test, Shiller et al. (2010a) found children had more imprecise perceptual boundaries than adults. Additionally, while adults demonstrated a significant shift in their perceptual boundaries for the perturbed sound contrast after testing, children did not reliably change these internal perceptual boundaries.

Vocal Variability and Perceptual Abilities Summary

Studies involving f_0 and formant manipulated feedback consistently showed increased variability in child baseline productions compared to adults. Individuals with more variable vocal productions are thought to have less defined internal sensorimotor representations. These sensorimotor representations encode the relationship of stored motor commands (utilized for feedforward control) and their auditory and somatosensory consequences (utilized for feedback control). Early in development, it is hypothesized, that children must rely more on auditory feedback during vocalization to ensure that their speech is in line with their desired vocal output, resulting in unstable vocal productions (Scheerer and Jones, 2012). As exposure to speech increases, the reliability of internal sensorimotor representations increases and over-dependence on auditory feedback becomes unnecessary, with vocal output becoming more consistent, shifting their reliance to feedforward control (Scheerer and Jones, 2012). Feedback control, however, continues to be an integral part of speech motor control, as auditory and somatosensory feedback are used to inform and maintain feedforward control, updating and refining sensorimotor representations when mismatches occur between expected and actual output (Franken et al., 2019).

Perceptually, findings across f_0 and formant studies generally demonstrated reduced precision in vocal and articulatory control in children compared to adults. While baseline variability represents potentially different control mechanisms (i.e., articulatory or vocal), reduced perceptual discrimination, and increased vocal variability in younger children across paradigms aligns with the DIVA model where reliance is postulated to shift from feedback to feedforward control as sensorimotor targets are refined over multiple productions, with initial targets being larger and discrimination abilities less sensitive (Guenther, 2006; Tourville et al., 2008; Guenther and Vladusich, 2012). This was further supported by findings by MacDonald et al. (2012) where baseline F1 and F2 variability was found to decrease with age. This suggests that maturational changes occurring in the speech motor control system affect the extent to which auditory feedback is used to modify internal sensorimotor representations.

Speech, Language and Literacy

Nine studies collected additional information related to speech, language, reading, cognitive, and social competence. **Supplementary Material 4**[3] details the additional measures collected and related findings.

Speech and Language

Three studies collected articulation information, with two using the Goldman-Fristoe Test of Articulation 2 (GFTA-2; Shiller et al., 2010b; Daliri et al., 2018). Five studies collected receptive and expressive language information with the most common test used being the Clinical Evaluation of Language Fundamentals (CELF; Russo et al., 2008; Shiller et al., 2010b; Coughler et al., 2021). Five studies collected information about cognitive abilities with the most common test used being the Wechsler Abbreviated Scale of Intelligence (WASI; Russo et al., 2008; Daliri et al., 2018; Scheerer et al., 2020b; Coughler et al., 2021). In these studies results included: (a) no significant correlation found between speech and language tests and compensation (Daliri et al., 2018; Scheerer et al., 2020b; Coughler et al., 2021); (b) a significant correlation of response magnitudes with core, receptive and expressive language scores, where decreased magnitude responses were related to higher language scores (Russo et al., 2008); and (c) a significant positive correlation of compensation with performance on non-word repetition (Terband et al., 2014). Scheerer et al. (2020b) also found average response latency significantly predicted Multidimensional Social Competence Scale scores (MSCS).

Literacy

Phonological awareness measures were collected in three studies (van den Bunt et al., 2018a,b; Caudrelier et al., 2019). Reading measures were additionally collected in two of these studies (van den Bunt et al., 2018a,b). Better rapid automatized naming was found to correlate with better compensation (van den Bunt et al., 2018a), as well as correlated with weaker deviation from the baseline during the ramp-up phase and stronger de-adaptation during the ramp-down phase (van den Bunt et al., 2018b).

Average phonological awareness scores were significantly higher in children who compensated than to non-adapting children (Caudrelier et al., 2019), and was associated with stronger compensation during ramp-up and hold phase, and weaker de-adaptation in the ramp-down and end phases (van den Bunt et al., 2018b). Phonological awareness, rapid naming, and letter knowledge correlated significantly with compensation in preliterate children, whereas reading correlated with compensation in literate children (van den Bunt et al., 2018a). Overall, literacy was also found to play a role in compensatory

[3]https://osf.io/2jwt4/

response magnitude, where significantly stronger compensation in hold and end phases were found for literate children relative to preliterate children (van den Bunt et al., 2018a).

Speech, Language and Literacy Summary

In general, findings examining the relationship of speech, language, and cognitive measures in relation to compensation magnitude were limited. A few studies found no significant relationship of speech and language abilities with compensation, while one study found a significant negative relationship with language abilities. One possibility for these differences could be a result of differences in the assessment tools used to assess speech, and language. As well, although several studies collected information on speech, language and cognitive abilities, the relationship of these abilities with compensation magnitude were not examined.

A clear relationship however was evident for literacy in relation to the developmental trajectory of auditory feedback control. Reading and preliteracy skills (e.g., phonological awareness) significantly correlated with compensation ability. In particular, phonological awareness scores, a strong predictor of later reading ability, were significantly higher in children who compensated compared to those who showed no compensatory response (Caudrelier et al., 2019), and in those who showed greater compensation (van den Bunt et al., 2018b). These results suggest an interplay among auditory-integration, speech motor control, and reading, supporting theories that impaired phonological representations (essentially sensorimotor representations) may underlie reading deficiencies (Ramus and Szenkovits, 2008), although further exploration is needed to understand other potential factors that may influence this relationship.

Clinical Findings

Fundamental Frequency Manipulation

Three studies exploring clinical population responses to f_0 manipulated altered auditory feedback included children with autism spectrum disorder (ASD; Russo et al., 2008; Scheerer et al., 2020b) and children who are 16p11.2 deletion carriers (Demopoulos et al., 2018). Deletion at 16p11.2 is commonly observed in individuals with diagnoses of developmental coordination disorder, phonological processing disorder, language disorders, and ASD (Demopoulos et al., 2018). Russo et al. (2008) had mixed findings, where some children with ASD demonstrated smaller mean magnitude compensatory responses, while others created atypically large compensatory responses compared to TD children. Scheerer et al. (2020b) found children with ASD had shorter response latencies than TD children, but showed similar compensatory responses to TD peers. In contrast, Demopoulos et al. (2018) consistently found children with 16p11.2 deletion showed larger pitch compensation compared to controls.

Formant Frequency Manipulation

Seven studies explored compensatory responses to formant frequency manipulations of children with speech and language difficulties. Two of these studies examined children with SSD (Shiller et al., 2010b; Terband et al., 2014), two examined responses of children who stutter (Daliri et al., 2018; Kim et al., 2020), one of children with dyslexia (van den Bunt et al., 2018b), one with children who are 16p11.2 deletion carriers (Demopoulos et al., 2018), and one with children with developmental language disorders (DLD; Coughler et al., 2021). No consistent findings were seen across these clinical populations. Children who stutter were found in one study (Daliri et al., 2018) to show greater compensation in F1 than adults who stutter, however they did not differ from children who do not stutter. However, Kim et al. (2020) found children who stutter showed no compensatory response. Conflicting results were similarly found for children with SSD. Terband et al. (2014) found children with SSD followed the perturbation in F1 during hold and end phases, whereas Shiller et al. (2010b) found compensation as well as an adaptation response. It is important to note that Shiller et al. (2010b) only included one participant. Similar to Shiller et al. (2010b), van den Bunt et al. (2018b) found all children with dyslexia compensated in the opposite direction of the perturbation, with the only difference from typically reading peers being a weaker return to baseline during the ramp-down phase. In contrast, children with 16p11.2 deletion showed significantly weaker compensation than their TD peers (Demopoulos et al., 2018). Coughler et al. (2021) found a unique pattern where children with a specific deficit in language (DLD) demonstrated differential compensation in positive and negative shift conditions. Children with DLD showed larger compensation in the positive shift condition and compensated less in the negative shift condition compared to typically developing peers.

Clinical Findings Summary

Clinically, across formant and *f0* manipulation studies, a broad range of disorder areas were examined, from ASD, 16p11.2 deletion, SSD, dyslexia, fluency, to DLD. All of these disorders have been linked to impairments in or closely linked to auditory processing. Although there was a lack of methodological consistency, several studies found aberrant responses in some of the clinical populations compared with typically developing children. This included increased following responses and larger or smaller compensation responses compared to typically developing peers.

Results involving children who stutter further support developmental sensorimotor control changes into adulthood. In Daliri et al. (2018), children who stutter produced greater compensation than adults who stuttered, however, they did not differ compared to children who do not stutter. This suggests some shift in reliance in adults, which is further supported by the finding that adults who stutter did not show any adaptation, whereas children who stutter did successfully adapt in their F1 productions. Adults who stutter were no longer updating their stored motor programs through feedforward control unlike children who stutter. Kim et al. (2020) found very different findings, children who stutter showed no significant compensatory response, although similarly, adults showed a reduced compensation compared to TD adults. Of note here, compensatory responses significantly correlated with age, where greater compensation occurred with increased age. Potential

differences between these two studies may be due to differences in shifts applied, where Daliri et al. (2018) shifted the phonemic category of the vowels, and Kim et al. (2020) did not.

Neuroanatomically, the atypical mixed compensation responses seen in children with ASD in f_0 manipulations studies (some in the typical range and others overcompensating) may relate to findings that children with ASD have weaker white matter connections between left ventral premotor cortex, a key area in speech motor planning, and other cortical regions involved in speech production (Russo et al., 2008; Peeva et al., 2013). Although not collected in their studies (Russo et al., 2008; Scheerer et al., 2020b), weaker white matter connections could be associated with the overcompensation profile found in some children. Scheerer et al. (2020b) additionally found shorter response latencies in children with autism, indicating more research is needed to further understand what may be driving differences.

Similarly, the mixed findings found for children with SSD, with some showing typical responses (Shiller et al., 2010b) and others showing increased following responses (Terband et al., 2014), could be related to gray and white matter volume differences. Previous studies have found abnormal gray and white matter volume in areas relating to speech motor control, which is thought to be related to delays in synaptic pruning (Preston et al., 2014).

Of interest, the type of manipulation was shown to affect the direction of aberrant responses. Demopoulos et al. (2018) found children who were carriers of 16p11.2 deletion showed overcompensation compared to controls in response to pitch perturbation (unexpected shift), but undercompensation compared to controls in response to formant manipulated feedback (sustained shift). This further supports theories that these vocal and articulatory controls develop at differing rates and time points. However, it is important to take into consideration that different experimental designs (i.e., sudden vs. sustained shift) may be the driving factor resulting in these differing compensation responses. Further research is needed contrasting responses to sudden and sustained shifts using consistent manipulations.

Neurophysiological and Neuroimaging Findings

Fundamental Frequency Manipulation

Two studies explored EEG responses to f_0 altered auditory feedback (Liu et al., 2013; Scheerer et al., 2013b). Both found that P1 and N1 latency, and P1 amplitude decreased with age (Liu et al., 2013; Scheerer et al., 2013b). Scheerer et al. (2013b) also found that N1 amplitude increased with age, however, Liu et al. (2013) did not find a significant effect of age on N1 amplitude. P2 amplitude showed greater variability, but generally was shown to increase with age (Liu et al., 2013; Scheerer et al., 2013b). Notably, Liu et al. (2013) also found significant interactions between sex and age in the N1 and P2 potentials. Male participants generally produced larger N1 amplitudes, and within the group of older children, females demonstrated significantly shorter N1 latencies than males. Among the young children, males had significantly larger P2 amplitudes than females, and young females had significantly larger P2 amplitudes than older females (Liu et al., 2013). In terms of P2 response latencies, P2 latency was found to be age-dependent for males only, however, within the group of older children, females were found to have significantly shorter P2 response latencies compared to males (Liu et al., 2013). The variable age-sex interactions found by Liu et al. (2013) speaks to the complexity of factors influencing neural responses to auditory feedback.

Formant Frequency Manipulation

Two studies (van den Bunt et al., 2018b; Ohashi and Ostry, 2021) examined neural activation in relation to formant manipulation sensorimotor control studies. van den Bunt et al. (2018b) used fractional anisotropy to measure connectivity of the arcuate fasciculus/superior longitudinal fasciculus (AF/SLF). Fractional anisotropy of the AF did not directly relate to altered auditory feedback responses, but did correlate strongly with phonological awareness scores. When phonological awareness was controlled, higher fractional anisotropy was found to be associated with less adaptation during altered auditory feedback (van den Bunt et al., 2018b). Ohashi and Ostry (2021) found children and adults had distinct patterns of functional connectivity. In adults, compensation to altered auditory feedback was positively correlated with activation in the right inferior frontal gyrus (area 44) and associative sensory regions. In children, compensation was positively correlated with functional connectivity of the primary somatosensory cortex (S1)/primary motor cortex (M1) and posterior rostral cingulate zone (RCZ) and left anterior insular cortex. When contrasting younger and older children (over 9 years), older children showed an increasingly adult-like pattern of connectivity.

Neurophysiological and Neuroimaging Findings Summary

Neurophysiological and neuroimaging studies involved examining evoked potentials, diffusion-weighted imaging and resting-state functional connectivity. Evoked potentials have a well-established history of being an objective measure of maturation of the nervous system, which can increase our understanding of neurophysiological changes that underlie behavioral responses to sensory input (Eggermont, 1988). Due to small sample sizes and differing languages (English in Scheerer et al., 2013b, and Mandarin in Liu et al., 2013), age-dependent conclusions are guarded, although both utilized similar f_0 shifted paradigms. Developmental trends observed in P1-N1-P2 amplitudes and latencies mirrored general developmental trends found during passive listening tasks (Ponton et al., 2000; Wunderlich and Cone-Wesson, 2006; Fitzroy et al., 2015). These similarities support the existence of a developmental gradient in auditory integration.

The significant decreases in latency observed with age across both the P1 and the N1 components of the P1-N1-P2 complex (Liu et al., 2013; Scheerer et al., 2013b), alongside the decreases in behavioral response latencies, together provide significant support for the existence of age-related changes in the efficiency of auditory integration in the cortex. This in turn suggests that the

efficiency of information processing in cortical areas supporting sensory function influences the developmental trajectory of speech motor control. Findings of consistent decreases in P1 amplitude (Liu et al., 2013; Scheerer et al., 2013b) and increases in N1 amplitude (Scheerer et al., 2013b) across age during the processing of altered auditory feedback provides more evidence for the age-dependent shift from reliance on feedback to reliance on feedforward control identified through the analysis of response magnitudes.

Neuroanatomically, although findings examining fractional anisotropy of the left arcuate fasciculus were not found to directly relate to compensation responses (van den Bunt et al., 2018b), resting-state functional connectivity showed distinct patterns of connectivity which significantly related to compensation in speech sensorimotor adaptation tasks (Ohashi and Ostry, 2021). This finding supports the hypothesis that protracted neural plasticity during development relates to differences in performance in speech motor learning, demonstrating that speech motor adaptation abilities relate to cortical remodeling and reorganization occurring across development.

SUMMARY

Speech motor control, in particular, auditory feedback is key to the development of speech, however, much remains unknown about how this develops in children. The current scoping review explored pediatric studies that examined frequency altered auditory feedback, with findings divided into fundamental and formant frequency manipulation studies. The aim of this scoping review was to gain a broad overview of the current state of research in pediatric frequency altered auditory feedback, investigating how responses to shifted auditory feedback change throughout development, thus expanding our understanding of the developing speech motor control system, and highlighting potential future directions and gaps in the literature. Searches from six academic databases retrieved twenty-three articles that explored various implementations of frequency manipulated altered auditory feedback. Results found age-dependent developmental trajectories related to response latencies, magnitude of compensatory responses, variability and perceptual abilities, as well as relationships of compensation with literacy.

Age-Dependent Trajectory of Responses to Altered Auditory Feedback

The primary goal of this study was to gain a better understanding of how and when children use information from auditory feedback to regulate their speech. Results across both fundamental and formant frequency manipulated altered auditory feedback showed children above the age of four generally compensated for the altered auditory feedback in the opposite direction of the perturbation (MacDonald et al., 2012). This is consistent with previous research of pediatric responses to other forms of altered auditory feedback where children, like adults, adjusted their speech to perturbations of their vocal intensity, timing, and jaw/lip positioning (e.g., Chase et al., 1961; Siegel et al., 1976; Ménard et al., 2008). However, mixed findings across different measures were evident. For example, increased incidence of following responses, as well as larger and smaller compensatory responses in children compared to adults, suggests that pediatric populations may not be using auditory feedback for speech motor control in the same manner as adults. Results obtained using these different measures may provide key information about the developmental trajectory of auditory feedback control across childhood.

Future Directions

Although the reviewed studies provided relevant findings about the potential of age-dependent changes in auditory feedback control, further research is needed. This scoping review found several limitations and gaps within the field, highlighting the need for further high quality quantitative well-designed studies.

The most significant limitation across these studies is a lack of power due to small sample sizes. Several of the studies discussed included around 20 participants, with only three studies including more than 30 participants in each group (Scheerer et al., 2013b, 2020b; van den Bunt et al., 2018a). Some studies also failed to report effect sizes, making power computations not possible. Creating well-powered studies, with consistent reporting of effect sizes, would enable a more expansive systematic or meta-analysis in the future.

In terms of age ranges explored, very few studies of f_0 manipulation explored younger ages (primarily focusing on school age), whereas formant manipulation studies explored a broader range. Scheerer et al. (2013b) and Scheerer et al. (2020b) were the only studies to examine a broad age range, looking at children 4–17 (as well as adults 18–30 years) and 3–13 years, respectively, in pitch perturbation. A more expansive look at changes across development is particularly missing in formant manipulations studies. Expanding age ranges within studies, examining changes across childhood, between young and older children, would provide clearer information about developmental changes in responses. Additionally, utilizing longitudinal studies may clarify maturational changes, taking into account the increased variability found in younger children.

In light of the several other subsystems developing in parallel with speech motor control, more comprehensive data collection is necessary. While several of the studies discussed in this scoping review examined aspects of other systems (e.g., clinical measures, neurophysiological, and perceptual), no study provided a comprehensive examination, combining information about neural processing, and parallel skill development (e.g., speech, language, and literacy) in relation to behavioral performance (e.g., vocal response magnitude).

Overall, gaps in the literature highlight the need for more comprehensive, larger sample, broad age range studies, with multiple outcome measures (e.g., magnitude, response latency, language, phonological awareness, literacy, and auditory perception).

CONCLUSION

The current scoping review provides a detailed description of the current state of research on pediatric responses to conditions of f_0 and formant shifted altered auditory feedback, and highlights critical gaps in the literature. As discovered, only a small body of research exists to date that addresses pediatric responses to frequency altered auditory feedback. Within the 23 articles reviewed, significant variability was seen in methodological frameworks, manipulations applied, as well as languages of participants, and age ranges. Significant variability in the characteristics of behavioral responses across these studies also leads to difficulties in generalizing and identifying age-dependent trends.

While this review provides key information about age-related changes in auditory integration and the development of speech motor control, there is a pressing need for future research in this area in order to understand further the general cognitive development of speech motor control.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

KQdL and DB contributed to the initial conception and design of the study. KQdL created initial search terms, conducted the initial literature review, and wrote the first draft of the manuscript. CC updated search terms and literature review and updated the manuscript. All authors contributed to editing the manuscript, and approved the submitted version.

FUNDING

KQdL contribution was supported by the Ward family through a studentship at Holland Bloorview Kids Rehabilitation Hospital. Funding for the study was supported by Discovery Grants through the Natural Sciences and Engineering Research Council of Canada (RGPIN-2015-05803 to DB, RGPIN-2019-05143 to DP, and RGPIN-2018-05655 to JOC).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnhum.2022.858863/full#supplementary-material

REFERENCES

Alsius, A., Mitsuya, T., and Munhall, K. (2013). Does compensation in auditory feedback require attention? *Proc. Meet. Acoust.* 19:e060098. doi: 10.1121/1.4799040

Anstis, S. M., and Cavanagh, P. (1979). Adaptation to frequency-shifted auditory feedback. *Percept. Psychophys.* 26, 449–458. doi: 10.3758/bf03204284

Arksey, H., and O'Malley, L. (2005). Scoping studies: towards a methodological framework. *Int. J. Soc. Res. Methodol.* 8, 19–32. doi: 10.1080/1364557032000119616

Bauer, J. J., and Larson, C. R. (2003). Audio-vocal responses to repetitive pitch-shift stimulation during a sustained vocalization: improvements in methodology for the pitch-shifting technique. *J. Acoust. Soc. Am.* 114, 1048–1054. doi: 10.1121/1.1592161

Behroozmand, R., Karvelis, L., Liu, H., and Larson, C. R. (2009). Vocalization-induced enhancement of the auditory cortex responsiveness during voice F0 feedback perturbation. *Clin. Neurophysiol.* 120, 1303–1312. doi: 10.1016/j.clinph.2009.04.022

Behroozmand, R., Korzyukov, O., Sattler, L., and Larson, C. R. (2012). Opposing and following vocal responses to pitch-shifted auditory feedback: evidence for different mechanisms of voice pitch control. *J. Acoust. Soc. Am.* 132, 2468–2477. doi: 10.1121/1.4746984

Behroozmand, R., Liu, H., and Larson, C. R. (2011). Time-dependent neural processing of auditory feedback during voice pitch error detection. *J. Cogn. Neurosci.* 23, 1205–1217. doi: 10.1162/jocn.2010.21447

Belmonte, M. K., Saxena-Chandhok, T., Cherian, R., Muneer, R., George, L., and Karanth, P. (2013). Oral motor deficits in speech-impaired children with autism. *Front. Integr. Neurosci.* 7:47. doi: 10.3389/fnint.2013.00047

Bloodstein, O., and Bernstein-Ratner, N. (2008). *A Handbook on Stuttering*, 6th Edn. Clifton Park, NY: Delmar.

Bohland, J. W., and Guenther, F. H. (2006). An fMRI investigation of syllable sequence production. *Neuroimage* 32, 821–841. doi: 10.1016/j.neuroimage.2006.04.173

Burnett, T. A., Freedland, M. B., Larson, C. R., and Hain, T. C. (1998). Voice F0 responses to manipulations in pitch feedback. *J. Acoust. Soc. Am.* 103, 3153–3161. doi: 10.1121/1.423073

Burnett, T. A., and Larson, C. R. (2002). Early pitch-shift response is active in both steady and dynamic voice pitch control. *J. Acoust. Soc. Am.* 112, 1058–1063. doi: 10.1121/1.1487844

Burnett, T. A., Senner, J. E., and Larson, C. R. (1997). Voice F0 responses to pitch-shifted auditory feedback: a preliminary study. *J. Voice* 11, 202–211. doi: 10.1016/s0892-1997(97)80079-3

Cai, S., Beal, D. S., Ghosh, S. S., Tiede, M. K., Guenther, F. H., and Perkell, J. S. (2012). Weak responses to auditory feedback perturbation during articulation in persons who stutter: evidence for abnormal auditory-motor transformation. *PLoS One* 7:e41830. doi: 10.1371/journal.pone.0041830

Cai, S., Ghosh, S. S., Guenther, F. H., and Perkell, J. S. (2011). Focal manipulations of formant trajectories reveal a role of auditory feedback in the online control of both within-syllable and between-syllable speech timing. *J. Neurosci.* 31, 16483–16490. doi: 10.1523/jneurosci.3653-11.2011

Callan, D. E., Kent, R. D., Guenther, F. H., and Vorperian, H. K. (2000). An auditory-feedback-based neural network model of speech production that is robust to developmental changes in the size and shape of the articulatory system. *J. Speech Lang. Hear. Res.* 43, 721–736. doi: 10.1044/jslhr.4303.721

Caudrelier, T., Ménard, L., Perrier, P., Schwartz, J.-L., Gerber, S., Vidou, C., et al. (2019). Transfer of sensorimotor learning reveals phoneme representations in preliterate children. *Cognition* 192:103973. doi: 10.1016/j.cognition.2019.05.010

Chase, R. A., Sutton, S., First, D., and Zubin, J. (1961). A developmental study of changes in behavior under delayed auditory feedback. *J. Genet. Psychol.* 99, 101–112. doi: 10.1080/00221325.1961.10534396

Chen, S. H., Liu, H., Xu, Y., and Larson, C. R. (2007). Voice F0 responses to pitch-shifted voice feedback during English speech. *J. Acoust. Soc. Am.* 121, 1157–1163. doi: 10.1121/1.2404624

Civier, O., Tasko, S. M., and Guenther, F. H. (2010). Overreliance on auditory feedback may lead to sound/syllable repetitions: simulations of stuttering and fluency-inducing conditions with a neural model of speech production. *J. Fluency Disord.* 35, 246–279. doi: 10.1016/j.jfludis.2010.05.002

Coplan, J., and Gleason, J. R. (1988). Unclear speech: recognition and significance of unintelligible speech in preschool children. *Pediatrics* 82, 447–452.

Coughler, C., Hamel, E. M., Oram Cardy, J., Archibald, L. M., and Purcell, D. W. (2021). Compensation to altered auditory feedback in children with developmental language disorder and typical development. *J. Speech Lang. Hear. Res.* 64, 2363–2376. doi: 10.1044/2020_jslhr-20-00374

Covidence Systematic Review Software (2020). *Veritas Health Innovation*. Melbourne: Covidence Systematic Review Software.

Cowie, R., and Douglas-Cowie, E. (1992). *Postlingually Acquired Deafness: Speech Deterioration and the Wider Consequences*. Berlin: De Gruyter Mouton.

Cunningham, J., Nicol, T., Zecker, S., and Kraus, N. (2000). Speech-evoked neurophysiologic responses in children with learning problems: development and behavioral correlates of perception. *Ear Hear.* 21, 554–568. doi: 10.1097/00003446-200012000-00003

Daliri, A., Wieland, E. A., Cai, S., Guenther, F. H., and Chang, S.-E. (2018). Auditory-motor adaptation is reduced in adults who stutter but not in children who stutter. *Dev. Sci.* 21:e12521. doi: 10.1111/desc.12521

de Boysson-Bardies, B. (2001). *How Language Comes to Children: From Birth to Two Years*. Cambridge, MA: Mit Press.

Demopoulos, C., Kothare, H., Mizuiri, D., Henderson-Sabes, J., Fregeau, B., Tjernagel, J., et al. (2018). Abnormal speech motor control in individuals with 16p11.2 deletions. *Sci. Rep.* 8:1274. doi: 10.1038/s41598-018-19751-x

Donath, T. M., Natke, U., and Kalveram, K. T. (2002). Effects of frequency-shifted auditory feedback on voice F 0 contours in syllables. *J. Acoust. Soc. Am.* 111, 357–366. doi: 10.1121/1.1424870

Eggermont, J. J. (1988). On the rate of maturation of sensory evoked potentials. *Electroencephalogr. Clin. Neurophysiol.* 70, 293–305. doi: 10.1016/0013-4694(88)90048-x

Elman, J. L. (1981). Effects of frequency-shifted feedback on the pitch of vocal productions. *J. Acoust. Soc. Am.* 70, 45–50. doi: 10.1121/1.386580

Fitzroy, A. B., Krizman, J., Tierney, A., Agouridou, M., and Kraus, N. (2015). Longitudinal maturation of auditory cortical function during adolescence. *Front. Hum. Neurosci.* 9:530. doi: 10.3389/fnhum.2015.00530

Franken, M. K., Acheson, D. J., McQueen, J. M., Hagoort, P., and Eisner, F. (2018). Opposing and following responses in sensorimotor speech control: why responses go both ways. *Psychon. Bull. Rev.* 25, 1458–1467. doi: 10.31234/osf.io/tskxq

Franken, M. K., Acheson, D. J., McQueen, J. M., Hagoort, P., and Eisner, F. (2019). Consistency influences altered auditory feedback processing. *Q. J. Exp. Psychol.* 72, 2371–2379. doi: 10.1177/1747021819838939

Ghosh, S. S., Tourville, J. A., and Guenther, F. H. (2008). A neuroimaging study of premotor lateralization and cerebellar involvement in the production of phonemes and syllables. *J. Speech Lang. Hear. Res.* 51, 1183–1202. doi: 10.1044/1092-4388(2008/07-0119

Golfinopoulos, E., Tourville, J. A., Bohland, J. W., Ghosh, S. S., Nieto-Castanon, A., and Guenther, F. H. (2011). fMRI investigation of unexpected somatosensory feedback perturbation during speech. *Neuroimage* 55, 1324–1338. doi: 10.1016/j.neuroimage.2010.12.065

Guenther, F. H. (1994). A neural network model of speech acquisition and motor equivalent speech production. *Biol. Cybernet.* 72, 43–53. doi: 10.1007/bf00206237

Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychol. Rev.* 102, 594–621. doi: 10.1037/0033-295x.102.3.594

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *J. Commun. Disord.* 39, 350–365. doi: 10.1016/j.jcomdis.2006.06.013

Guenther, F. H. (2016). *Neural Control of Speech*. Cambridge, MA: The MIT Press.

Guenther, F. H., Hampson, M., and Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychol. Rev.* 105, 611–633. doi: 10.1037/0033-295x.105.4.611-633

Guenther, F. H., and Vladusich, T. (2012). A neural theory of speech acquisition and production. *J. Neurolinguist.* 25, 408–422. doi: 10.1016/j.jneuroling.2009.08.006

Hain, T. C., Burnett, T. A., Kiran, S., Larson, C. R., Singh, S., and Kenney, M. K. (2000). Instructing subjects to make a voluntary response reveals the presence of two components to the audio-vocal reflex. *Exp. Brain Res.* 130, 133–141. doi: 10.1007/s002219900237

Heinks-Maldonado, T. H., Nagarajan, S. S., and Houde, J. F. (2006). Magnetoencephalographic evidence for a precise forward model in speech production. *NeuroReport* 17, 1375–1379. doi: 10.1097/01.wnr.0000233102.43526.e9

Heller Murray, E. S., and Stepp, C. E. (2020). Relationships between vocal pitch perception and production: a developmental perspective. *Sci. Rep.* 10:3912. doi: 10.1038/s41598-020-60756-2

Hillyard, S. A., and Picton, T. W. (1978). On and off components in the auditory evoked potential. *Percept. Psychophys.* 24, 391–398. doi: 10.3758/bf03199736

Houde, J. F., and Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science* 279, 1213–1216. doi: 10.1126/science.279.5354.1213

Houde, J. F., and Jordan, M. I. (2002). Sensorimotor adaptation of speech I. *J. Speech Lang. Hear. Res.* 45, 295–310. doi: 10.1044/1092-4388(2002/023)

Houde, J. F., Kort, N. S., Niziolek, C. A., Chang, E. F., and Nagarajan, S. S. (2013). Neural evidence for state feedback control of speaking. *Proc. Meet. Acoust.* 19:e060178. doi: 10.1121/1.4799495

Houde, J. F., and Nagarajan, S. S. (2011). Speech production as state feedback control. *Front. Hum. Neurosci.* 5:82. doi: 10.3389/fnhum.2011.00082

Hu, H., Liu, Y., Guo, Z., Li, W., Liu, P., Chen, S., et al. (2015). Attention modulates cortical processing of pitch feedback errors in voice control. *Sci. Rep.* 5, 1–8. doi: 10.1038/srep07812

Jones, J. A., and Munhall, K. G. (2000). Perceptual calibration of F0 production: evidence from feedback perturbation. *J. Acoust. Soc. Am.* 108:1246. doi: 10.1121/1.1288414

Jones, J. A., and Munhall, K. G. (2002). The role of auditory feedback during phonation: studies of mandarin tone production. *J. Phonet.* 30, 303–320. doi: 10.1006/jpho.2001.0160

Jones, J. A., and Munhall, K. G. (2005). Remapping auditory-motor representations in voice production. *Curr. Biol.* 15, 1768–1772. doi: 10.1016/j.cub.2005.08.063

Katseff, S., Houde, J., and Johnson, K. (2012). Partial compensation for altered auditory feedback: a tradeoff with somatosensory feedback? *Lang. Speech* 55, 295–308. doi: 10.1177/0023830911417802

Kent, R. D. (1999). "Motor control: neurophysiology and functional development," in *Clinical Management of Motor Speech Disorders in Children*, eds A. Caruso and E. Strand (New York, NY: Thieme Medical Publishers).

Kent, R. D. (2004). "Models of speech motor control: implications from recent developments in neurophysiological and neurobehavioral science," in *Speech Motor Control in Normal and Disordered Speech*, eds B. Maassen, R. Kent, H. F. M. Peters, P. H. H. M. van Lieshout, and W. Hulstijn (Oxford: Oxford University Press), 1–28. doi: 10.1093/oso/9780198795421.003.0001

Kent, R. D., and Vorperian, H. K. (1995). Development of the craniofacial-oral-laryngeal anatomy: a review. *J. Med. Speech Lang. Pathol.* 3, 145–190.

Keough, D., Hawco, C., and Jones, J. A. (2013). Auditory-motor adaptation to frequency-altered auditory feedback occurs when participants ignore feedback. *BMC Neurosci.* 14:25. doi: 10.1186/1471-2202-14-25

Kim, K. S., Daliri, A., Flanagan, J. R., and Max, L. (2020). Dissociated development of speech and limb sensorimotor learning in stuttering: speech auditory-motor learning is impaired in both children and adults who stutter. *Neuroscience* 451, 1–21. doi: 10.1101/2020.09.23.310797

Kotecha, R., Pardos, M., Wang, Y., Wu, T., Horn, P., Brown, D., et al. (2009). Modeling the developmental patterns of auditory evoked magnetic fields in children. *PLoS One* 4:e4811. doi: 10.1371/journal.pone.0004811

Kraus, N., McGee, T., Carrell, T., Sharma, A., Micco, A., and Nicol, T. (1993). Speech-evoked cortical potentials in children. *J. Am. Acad. Audiol.* 4, 238–248. doi: 10.1016/0168-5597(93)90063-u

Larson, C. R. (1998). Cross-modality influences in speech motor control. *J. Commun. Disord.* 31, 489–503. doi: 10.1016/s0021-9924(98)00021-5

Larson, C. R., Burnett, T. A., Bauer, J. J., Kiran, S., and Hain, T. C. (2001). Comparison of voice F 0 responses to pitch-shift onset and offset conditions. *J. Acoust. Soc. Am.* 110, 2845–2848. doi: 10.1121/1.1417527

Larson, C. R., Burnett, T. A., Kiran, S., and Hain, T. C. (2000). Effects of pitch-shift velocity on voice F 0 responses. *J. Acoust. Soc. Am.* 107, 559–564. doi: 10.1121/1.428323

Larson, C. R., Sun, J., and Hain, T. C. (2007). Effects of simultaneous perturbations of voice pitch and loudness feedback on voice F 0 and amplitude control. *J. Acoust. Soc. Am.* 121, 2862–2872. doi: 10.1121/1.2715657

Lester-Smith, R. A., Daliri, A., Enos, N., Abur, D., Lupiani, A. A., Letcher, S., et al. (2020). The Relation of Articulatory and Vocal Auditory-Motor Control in Typical Speakers. *J. Speech Lang. Hear. Res.* 63, 3628–3642. doi: 10.1044/2020_JSLHR-20-00192

Levac, D., Colquhoun, H., and O'Brien, K. K. (2010). Scoping studies: advancing the methodology. *Implement. Sci.* 5:69. doi: 10.1186/1748-5908-5-69

Liberman, A. M. (1989). Reading is hard just because listening is easy. *Brain Read.* 95, 197–205. doi: 10.1007/978-1-349-10732-2_14

Liu, H., and Larson, C. R. (2007). Effects of perturbation magnitude and voice F 0 level on the pitch-shift reflex. *J. Acoust. Soc. Am.* 122, 3671–3677. doi: 10.1121/1.2800254

Liu, H., Meshman, M., Behroozmand, R., and Larson, C. R. (2011). Differential effects of perturbation direction and magnitude on the neural processing of voice pitch feedback. *Clin. Neurophysiol.* 122, 951–957. doi: 10.1016/j.clinph.2010.08.010

Liu, H., Russo, N. M., and Larson, C. R. (2010). Age-related differences in vocal responses to pitch feedback perturbations: a preliminary study. *J. Acoust. Soc. Am.* 127, 1042–1046. doi: 10.1121/1.3273880

Liu, P., Chen, Z., Jones, J. A., Wang, E. Q., Chen, S., Huang, D., et al. (2013). Developmental sex-specific change in auditory-vocal integration: ERP evidence in children. *Clin. Neurophysiol.* 124, 503–513. doi: 10.1016/j.clinph.2012.08.024

Liu, P., Chen, Z., Larson, C. R., Huang, D., and Liu, H. (2010). Auditory feedback control of voice fundamental frequency in school children. *J. Acoust. Soc. Am.* 128:1306. doi: 10.1121/1.3467773

MacDonald, E. N., Goldberg, R., and Munhall, K. G. (2010). Compensations in response to real-time formant perturbations of different magnitudes. *J. Acoust. Soc. Am.* 127, 1059–1068. doi: 10.1121/1.3278606

MacDonald, E. N., Johnson, E. K., Forsythe, J., Plante, P., and Munhall, K. G. (2012). Children's development of self-regulation in speech production. *Curr. Biol.* 22, 113–117. doi: 10.1016/j.cub.2011.11.052

Mattingly, I. G. (1972). "Reading, the linguistic process, and linguistic awareness," in *Language by Ear and by Eye*, eds J. F. Kavenagh and I. G. Mattingly (Cambridge, MA: MIT Press), 23–34.

Ménard, L., Perrier, P., Aubin, J., Savariaux, C., and Thibeault, M. (2008). Compensation strategies for a lip-tube perturbation of French [u]: an acoustic and perceptual study of 4-year-old children. *J. Acoust. Soc. Am.* 124, 1192–1206. doi: 10.1121/1.2945704

Mitsuya, T., MacDonald, E. N., Munhall, K. G., and Purcell, D. W. (2015). Formant compensation for auditory feedback with English vowels. *J. Acoust. Soc. Am.* 138, 413–424. doi: 10.1121/1.4923154

Möbius, B., and Dogil, G. (2002). *Phonemic and Postural Effects on the Production of Prosody*. Stuttgart: University of Stuttgart.

Munhall, K. G., MacDonald, E. N., Byrne, S. K., and Johnsrude, I. (2009). Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate. *J. Acoust. Soc. Am.* 125, 384–390. doi: 10.1121/1.3035829

Namasivayam, A. K., Pukonen, M., Goshulak, D., Yu, V. Y., Kadis, D. S., Kroll, R., et al. (2013). Relationship between speech motor control and speech intelligibility in children with speech sound disorders. *J. Commun. Disord.* 46, 264–280. doi: 10.1016/j.jcomdis.2013.02.003

Natke, U., Donath, T. M., and Kalveram, K. T. (2003). Control of voice fundamental frequency in speaking versus singing. *J. Acoust. Soc. Am.* 113, 1587–1593. doi: 10.1121/1.1543928

Natke, U., and Kalveram, K. T. (2001). Effects of frequency-shifted auditory feedback on fundamental frequency of long stressed and unstressed syllables. *J. Speech Lang. Hear. Res.* 44, 577–584. doi: 10.1044/1092-4388(2001/045)

Ohashi, H., and Ostry, D. J. (2021). Neural development of speech sensorimotor learning. *J. Neurosci.* 41, 4023–4035. doi: 10.1523/jneurosci.2884-20.2021

Oller, D. K., and Eilers, R. E. (1988). The role of audition in infant babbling. *Child Dev.* 59, 441–449. doi: 10.2307/1130323

Patel, S., Nishimura, C., Lodhavia, A., Korzyukov, O., Parkinson, A., Robin, D. A., et al. (2014). Understanding the mechanisms underlying voluntary responses to pitch-shifted auditory feedback. *J. Acoust. Soc. Am.* 135, 3036–3044. doi: 10.1121/1.4870490

Peeva, M. G., Tourville, J. A., Agam, Y., Holland, B., Manoach, D. S., and Guenther, F. H. (2013). White matter impairment in the speech network of individuals with autism spectrum disorder. *NeuroImage Clin.* 3, 234–241. doi: 10.1016/j.nicl.2013.08.011

Perkell, J., Matthies, M., Lane, H., Guenther, F., Wilhelms-Tricarico, R., Wozniak, J., et al. (1997). Speech motor control: acoustic goals, saturation effects, auditory feedback and internal models. *Speech Commun.* 22, 227–250. doi: 10.1016/s0167-6393(97)00026-5

Polich, J., Ladish, C., and Burns, T. (1990). Normal variation of P300 in children: age, memory span, and head size. *Int. J. Psychophysiol.* 9, 237–248. doi: 10.1016/0167-8760(90)90056-J

Ponton, C. W., Eggermont, J. J., Kwong, B., and Don, M. (2000). Maturation of human central auditory system activity: evidence from multi-channel evoked potentials. *Clin. Neurophysiol.* 111, 220–236. doi: 10.1016/s1388-2457(99)00236-9

Preston, J. L., Molfese, P. J., Mencl, W. E., Frost, S. J., Hoeft, F., Fulbright, R. K., et al. (2014). Structural brain differences in school-age children with residual speech sound errors. *Brain Lang.* 128, 25–33. doi: 10.1016/j.bandl.2013.11.001

Purcell, D. W., and Munhall, K. G. (2006a). Compensation following real-time manipulation of formants in isolated vowels. *J. Acoust. Soc. Am.* 119, 2288–2297. doi: 10.1121/1.2173514

Purcell, D. W., and Munhall, K. G. (2006b). Adaptive control of vowel formant frequency: evidence from real-time formant manipulation. *J. Acoust. Soc. Am.* 120, 966–977. doi: 10.1121/1.2217714

Raharjo, I., Kothare, H., Nagarajan, S. S., and Houde, J. F. (2021). Speech compensation responses and sensorimotor adaptation to formant feedback perturbations. *J. Acoust. Soc. Am.* 149, 1147–1161. doi: 10.1121/10.0003440

Ramus, F., and Szenkovits, G. (2008). What phonological deficit? *Q. J. Exp. Psychol.* 61, 129–141. doi: 10.1080/17470210701508822

Rueckl, J. G., Paz-Alonso, P. M., Molfese, P. J., Kuo, W. J., Bick, A., Frost, S. J., et al. (2015). Universal brain signature of proficient reading: evidence from four contrasting languages. *Proc. Natl. Acad. Sci. U.S.A.* 112, 15510–15515. doi: 10.1073/pnas.1509321112

Russo, N., Larson, C., and Kraus, N. (2008). Audio-vocal system regulation in children with autism spectrum disorders. *Exp. Brain Res.* 188, 111–124. doi: 10.1007/s00221-008-1348-2

Scheerer, N. E., Behich, J., Liu, H., and Jones, J. A. (2013a). ERP correlates of the magnitude of pitch errors detected in the human voice. *Neuroscience* 240, 176–185. doi: 10.1016/j.neuroscience.2013.02.054

Scheerer, N. E., Liu, H., and Jones, J. A. (2013b). The developmental trajectory of vocal and event-related potential responses to frequency-altered auditory feedback. *Eur. J. Neurosci.* 38, 3189–3200. doi: 10.1111/ejn.12301

Scheerer, N. E., Jacobson, D. S., and Jones, J. A. (2016). Sensorimotor learning in children and adults: exposure to frequency-altered auditory feedback during speech production. *Neuroscience* 314, 106–115. doi: 10.1016/j.neuroscience.2015.11.037

Scheerer, N. E., Jacobson, D. S., and Jones, J. A. (2020a). Sensorimotor control of vocal production in early childhood. *J. Exp. Psychol. Gen.* 149, 1071–1077. doi: 10.1037/xge0000706

Scheerer, N. E., Jones, J. A., and Iarocci, G. (2020b). Exploring the relationship between prosodic control and social competence in children with and without autism spectrum disorder. *Autism Res.* 13, 1880–1892. doi: 10.1002/aur.2405

Scheerer, N. E., and Jones, J. A. (2012). The relationship between vocal accuracy and variability to the level of compensation to altered auditory feedback. *Neurosci. Lett.* 529, 128–132. doi: 10.1016/j.neulet.2012.09.012

Sharma, A., Kraus, N., McGee, T. J., and Nicol, T. G. (1997). Developmental changes in P1 and N1 central auditory responses elicited by consonant-vowel syllables. *Electroencephalogr. Clin. Neurophysiol.* 104, 540–545. doi: 10.1016/s0168-5597(97)00050-6

Shiller, D. M., Gracco, V. L., and Rvachew, S. (2010a). Auditory-motor learning during speech production in 9-11-year-old children. *PLoS One* 5:e12975. doi: 10.1371/journal.pone.0012975

Shiller, D. M., Rvachew, S., and Brosseau-Lapré, F. (2010b). Importance of auditory perceptual target to the achievement of speech production accuracy. *Can. J. Speech Lang. Pathol. Audiol.* 34, 181–192.

Shiller, D. M., and Rochon, M.-L. (2014). Auditory-perceptual learning improves speech motor adaptation in children. *J. Exp. Psychol. Hum. Percept. Perform.* 40, 1308–1315. doi: 10.1037/a0036660

Siegel, G. M., Pick, H. L., Olsen, M. G., and Sawin, L. (1976). Auditory feedback on the regulation of vocal intensity of preschool children. *Dev. Psychol.* 12, 255–261. doi: 10.1037/0012-1649.12.3.255

Sivasankar, M., Bauer, J. J., Babu, T., and Larson, C. R. (2005). Voice responses to changes in pitch of voice or tone auditory feedback. *J. Acoust. Soc. Am.* 117, 850–857. doi: 10.1121/1.1849933

Smith, A., and Goffman, L. (2004). "Interaction of motor and language factors in the development of speech production," in *Speech Motor Control in Normal and Disordered Speech*, eds B. Maassen, R. Kent, H. F. M. Peters, P. H. H. M. van Lieshout, and W. Hulstijn (Oxford: Oxford University Press), 225–252.

Stemple, J. C., Glaze, L. E., and Gerdeman, B. K. (2000). *Clinical Voice Pathology: Theory and Management*, 4th Edn. Boston, MA: Cengage Learning.

Strand, E. A., McCauley, R. J., Weigand, S. D., Stoeckel, R. E., and Baas, B. S. (2013). A motor speech assessment for children with severe speech disorders: reliability and validity evidence. *J. Speech Lang. Hear. Res.* 56, 505–520. doi: 10.1044/1092-4388(2012/12-0094)

Terband, H., and Maassen, B. (2010). Speech motor development in childhood apraxia of speech: generating testable hypotheses by neurocomputational modeling. *Folia Phoniatr. Logopaedica* 62, 134–142. doi: 10.1159/000287212

Terband, H., van Brenk, F., and van Doornik-van der Zee, A. (2014). Auditory feedback perturbation in children with developmental speech sound disorders. *J. Commun. Disord.* 51, 64–77. doi: 10.1016/j.jcomdis.2014.06.009

Tonnquist-Uhlen, I., Borg, E., and Spens, K. E. (1995). Topography of auditory evoked long-latency potentials in normal children, with particular reference to the N1 component. *Electroencephalogr. Clin. Neurophysiol.* 95, 34–41. doi: 10.1016/0013-4694(95)00044-y

Tourville, J. A., and Guenther, F. H. (2011). The DIVA model: a neural theory of speech acquisition and production. *Lang. Cogn. Process.* 26, 952–981. doi: 10.1080/01690960903498424

Tourville, J. A., Reilly, K. J., and Guenther, F. H. (2008). Neural mechanisms underlying auditory feedback control of speech. *NeuroImage* 39, 1429–1443. doi: 10.1016/j.neuroimage.2007.09.054

van Brenk, F., and Terband, H. (2020). Compensatory and adaptive responses to real-time formant shifts in adults and children. *J. Acoust. Soc. Am.* 147, 2261–2270. doi: 10.1121/10.0001018

van den Bunt, M. R., Groen, M. A., Frost, S., Lau, A., Preston, J. L., Gracco, V. L., et al. (2018a). Sensorimotor control of speech and children's reading ability. *Sci. Stud. Read.* 22, 503–516. doi: 10.1080/10888438.2018.1491583

van den Bunt, M. R., Groen, M. A., van der Kleij, S. W., Noordenbos, M. W., Segers, E., Pugh, K. R., et al. (2018b). Deficient response to altered auditory feedback in dyslexia. *Dev. Neuropsychol.* 43, 622–641. doi: 10.1080/87565641.2018.1495723

van den Bunt, M. R., Groen, M. A., Ito, T., Francisco, A. A., Gracco, V. L., Pugh, K. R., et al. (2017). Increased response to altered auditory feedback in dyslexia: a weaker sensorimotor magnet implied in the phonological deficit. *J. Speech Lang. Hear. Res.* 60, 654–667. doi: 10.1044/2016_jslhr-l-16-0201

Villacorta, V. M., Perkell, J. S., and Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *J. Acoust. Soc. Am.* 122, 2306–2319. doi: 10.1121/1.2773966

Wunderlich, J. L., and Cone-Wesson, B. K. (2006). Maturation of CAEP in infants and children: a review. *Hear. Res.* 212, 212–223. doi: 10.1016/j.heares.2005.11.008

Zarate, J. M., and Zatorre, R. J. (2008). Experience-dependent neural substrates involved in vocal pitch regulation during singing. *Neuroimage* 40, 1871–1887. doi: 10.1016/j.neuroimage.2008.01.026

Zhang, Z. (2016). Mechanics of human voice production and control. *J. Acoust. Soc. Am.* 140, 2614–2635. doi: 10.1121/1.4964509

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Coughler, Quinn de Launay, Purcell, Oram Cardy and Beal. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Human Neuroscience

ORIGINAL RESEARCH
published: 07 July 2022
doi: 10.3389/fnhum.2022.890065

Inter-Trial Formant Variability in Speech Production Is Actively Controlled but Does Not Affect Subsequent Adaptation to a Predictable Formant Perturbation

Hantao Wang[1] *and Ludo Max*[1,2]*

[1] Department of Speech and Hearing Sciences, University of Washington, Seattle, WA, United States, [2] Haskins Laboratories, New Haven, CT, United States

OPEN ACCESS

Edited by:
Jeffery A. Jones,
Wilfrid Laurier University, Canada

Reviewed by:
Takemi Mochida,
Nippon Telegraph and Telephone, Japan
David Jenson,
Washington State University, United States
Ding-lan Tang,
University of Wisconsin-Madison, United States

***Correspondence:**
Ludo Max
LudoMax@uw.edu

Specialty section:
This article was submitted to Speech and Language, a section of the journal Frontiers in Human Neuroscience

Received: 05 March 2022
Accepted: 14 June 2022
Published: 07 July 2022

Citation:
Wang H and Max L (2022) Inter-Trial Formant Variability in Speech Production Is Actively Controlled but Does Not Affect Subsequent Adaptation to a Predictable Formant Perturbation. Front. Hum. Neurosci. 16:890065. doi: 10.3389/fnhum.2022.890065

Despite ample evidence that speech production is associated with extensive trial-to-trial variability, it remains unclear whether this variability represents merely unwanted system noise or an actively regulated mechanism that is fundamental for maintaining and adapting accurate speech movements. Recent work on upper limb movements suggest that inter-trial variability may be not only actively regulated based on sensory feedback, but also provide a type of workspace exploration that facilitates sensorimotor learning. We therefore investigated whether experimentally reducing or magnifying inter-trial formant variability in the real-time auditory feedback during speech production (a) leads to adjustments in formant production variability that compensate for the manipulation, (b) changes the temporal structure of formant adjustments across productions, and (c) enhances learning in a subsequent adaptation task in which a predictable formant-shift perturbation is applied to the feedback signal. Results show that subjects gradually increased formant variability in their productions when hearing auditory feedback with reduced variability, but subsequent formant-shift adaptation was not affected by either reducing or magnifying the perceived variability. Thus, findings provide evidence for speakers' active control of inter-trial formant variability based on auditory feedback from previous trials, but–at least for the current short-term experimental manipulation of feedback variability–not for a role of this variability regulation mechanism in subsequent auditory-motor learning.

Keywords: speech motor control, variability, adaptation, auditory feedback, acoustics, articulation

INTRODUCTION

Over the years, the variability involved in human speech production has generated substantial empirical and theoretical interest. Both the physiological processes and acoustic output of speech production are inherently variable: even for a single speaker, no two repetitions of the same syllable are exactly the same in terms of muscle activation, kinematics, or acoustics (MacNeilage, 1970; Perkell and Klatt, 1986; Lindblom, 1990; Patri et al., 2015). Recently, it has started to become clear that such intra-individual variability at the behavioral level may reflect not only system noise but

also functionally relevant adjustments in movement planning. Identifying the contribution of both these components will be critical for a better understanding of the sensorimotor control principles involved in spoken language.

To date, most experimental studies on the role of variability in speech production have taken an *observational* approach. That is, researchers typically have observed specific aspects of production variability in selected experimental conditions (without directly manipulating variability itself), and assessed the relationship with other measures of production or perception. For example, in the area of phonation, when subjects were asked to match a target tone by vocalizing with the same pitch and duration, those with greater production variability during the baseline phase exhibited stronger compensatory responses when unpredictable pitch perturbations were introduced in the auditory feedback signal (Scheerer and Jones, 2012). As an example from speech articulation, production variability for vowels has been shown to be linked to the speaker's categorical perceptual boundary between vowels (Chao et al., 2019). Various studies also examined production variability in relation to aspects of perception, but quantified variability *across* different speaking conditions or consonant contexts (e.g., how different is /ɛ/ in "bed" vs. in "tech"), and thus did not address pure trial-to-trial variability in one particular phonetic context (e.g., Perkell et al., 2008; Franken et al., 2017).

Other groups have examined the potential relationship between observed trial-to-trial variability in speech acoustics and the extent of auditory-motor *learning* in a formant-shift adaptation paradigm. Purcell and Munhall (2006) reported a significant correlation between the lag 1 autocorrelation of trial-to-trial differences in a speaker's first formant (F1) during a baseline phase with unaltered auditory feedback and the extent of subsequent adaptation in response to a F1 perturbation. The relevance of this report is unclear, however, as calculating the lag 1 autocorrelation based on *differences* between neighboring trials can be a form of overdifferencing (Cryer and Chan, 2008). For example, it can be mathematically demonstrated that, after differencing, even a white noise time series has a lag 1 autocorrelation of –0.5. Thus, finding a negative lag 1 autocorrelation based on differenced data does not necessarily mean that, in the original time series of formant data, trials were actually adjusted based on the preceding trial. In subsequent work, the same group quantified variability of vowel production as the standard deviations of a speaker's F1 and F2 distributions during the baseline phase (MacDonald et al., 2011). Using pooled data from seven experiments with a total of 116 participants, they found no significant correlation between these different metrics of variability and the extent of adaptation to an F1 perturbation. In a more recent study, the same group did report a significant correlation between baseline F1 standard deviation and F1 adaptation, but they also cautioned–on the basis of a permutation test applied to the prior data–that this was most likely a chance result (Nault and Munhall, 2020).

Thus, the question whether individual speakers' baseline formant variability relates to their extent of auditory-motor learning in a formant-shift adaptation task remains unanswered to date. Interestingly, a study on upper limb sensorimotor control has suggested that reach movement trial-to-trial variability during a baseline phase does, in fact, facilitate early learning when adapting to a perturbing force field, possibly because greater variability offers more exploration of the task space (Wu et al., 2014). Even for upper limb movements, however, the generalizability and interpretation of this single study remain unclear (He et al., 2016; Singh et al., 2016; Murillo et al., 2017; Sternad, 2018; van der Vliet et al., 2018).

A more powerful approach toward addressing the issue of a potential relationship between sensorimotor variability and sensorimotor learning may consist of investigating variability with *experimental,* rather than observational, research methods. Direct experimental manipulation of inter-trial motor and/or sensory variability would allow one to ask multiple more specific questions. First, is inter-trial variability itself under active control by the central nervous system? In other words, can we find evidence of adjustments that compensate for increases or decreases in perceived variability of a specific performance measure? Second, does either the change in perceived variability of a performance measure or any active motor compensation for that perceived change affect sensorimotor adaptation in a new environment where that same aspect of performance is predictably perturbed?

To start investigating speech variability with such experimental methods, it is possible to adapt an approach taken in upper limb studies that magnified or attenuated visual feedback errors by a certain ratio (van Beers, 2009; Wong et al., 2009; Patton et al., 2013; van der Kooij et al., 2015). By aiming to manipulate the magnitude of feedback error in each trial, those studies also magnified and/or attenuated the *dispersion* of feedback error across trials. Hence, similar manipulations can be used to answer the above formulated question whether the inter-trial variability for a particular parameter of motor performance is actively controlled by the central nervous system. Specifically, motor behavior can be analyzed for any evidence of adjustments that compensate for the magnified or attenuated feedback variability. It should be noted that, in this context, a study's ability to both magnify and attenuate variability is critical from a methodological perspective. If an experimental paradigm only magnifies perceived variability by increasing the size of perceived movement errors, it is not possible to unambiguously attribute any resulting decrease in motor variability to the across-trials statistics *per se* vs. a preference for avoiding larger errors. If, on the other hand, a manipulation that attenuates perceived variability by minimizing perceived error leads to a compensatory *increase* in motor variability, then an interpretation based on the feedback statistics across trials is much more compelling as there are no theoretical reasons to expect a preference for avoiding smaller motor errors.

A reaching movement study by van Beers (2009) implemented such separate feedback conditions: movement endpoint errors were unaltered, reduced in magnitude by 50%, or increased in magnitude by 50%. Although the study did not specifically focus on compensation in terms of motor variability, van Beers (2009) found that the temporal structure of motor adjustments across trials differed among the visual feedback conditions: the sample lag 1 autocorrelation for movement endpoints was close to zero

when errors in the feedback (and thus inter-trial variability) were not manipulated, negative when feedback errors were magnified, and positive when feedback errors were attenuated. The findings were interpreted in terms of which model of motor learning best explains subjects' trial-to-trial adjustments, taking into account separate sources of central motor planning noise and peripheral motor execution noise. For natural movements with unperturbed feedback, van Beers (2009) concluded that trial-to-trial corrections are proportional to the magnitude of the previous error in such a way that movement variability is *minimized*, and it was suggested that this strategy is likely to underlie other forms of motor learning.

Lastly, a few upper limb studies have examined the effect of error magnification or attenuation on sensorimotor learning of a separate perturbation such as a visuomotor rotation. Results from those studies indicate that error magnification leads to more complete and faster adaptation whereas error attenuation has the opposite effect (Patton et al., 2013; van der Kooij et al., 2015). Despite this observed difference in adaptation, it has been argued that the adaptive learning mechanism itself, as quantified by a simple state-space model with the two parameters retention rate and error sensitivity, would remained unchanged between the different sensory feedback conditions (van der Kooij et al., 2015). However, other models of sensorimotor learning suggest that important parameters such as error sensitivity may be influenced by the prior history of feedback errors, a mechanism not captured by the simple state-space model (Herzfeld et al., 2014). Clearly, the effect of experimental manipulations of error magnitude and inter-trial variability on sensorimotor learning remains poorly understood even for upper limb movements.

Unfortunately, for speech articulation, work with experimental manipulations of feedback variability is only just starting to appear (see Tang et al., 2021), and the effect of manipulating the inter-trial variability of a specific parameter (e.g., frequency of one or more formants) on sensorimotor adaptation to a separate, predictable perturbation of the same parameter (e.g., a consistent formant shift) remains entirely unexplored. We therefore investigated whether an experimental magnification or attenuation of *perceived* inter-trial formant variability during speech production (a) leads to compensatory adjustments in *produced* formant variability, (b) induces changes in the temporal structure of formant adjustments across productions, and (c) affects subsequent auditory-motor learning when the speaker is exposed to a predictable formant-shift perturbation. Here, as the first step in this line of work, we implemented a relatively short-term formant variability manipulation (75 trials) and we looked for an effect on formant-shift adaptation in a subsequent task.

MATERIALS AND METHODS

General Procedure

Twenty-eight right-handed adult native speakers of American English (20 women, 8 men, age $M = 22.93$ years, $SD = 3.93$ years, range = 18–31) with no self-reported history of speech, hearing or neurological disorders participated after providing written informed consent (all procedures were approved by the Institutional Review Board at the University of Washington). Based on a pure tone hearing screening, all participants had monaural thresholds at or below 20 dB HL at all octave frequencies from 250 Hz to 4 kHz in both ears.

The experiment was conducted in a sound-attenuated booth. First, participants completed a practice session with unaltered auditory feedback to familiarize themselves with the instrumentation set-up by producing 7 blocks of three target words. Each block consisted of the monosyllabic words "talk," "tech," and "tuck" in randomized order. These words were presented individually on a monitor in front of the participant, each word remaining visible for 3 s. To help participants maintain a consistent speaking style, visual feedback about speech intensity and duration was presented on the monitor after each production. The target intensity was between 72 and 80 dB SPL, and the target vowel duration was 100–400 ms.

The actual experiment then included a *Pre-test* and two versions of a *Variability* task that were each immediately followed by an *Adaptation* task (**Figure 1A**). The *Pre-test* served to determine each participant's median frequencies for the first and second formant (F1, F2) for the three target words (details below). During productions of the same words in the *Variability* tasks, inter-trial formant variability in the auditory feedback was either manipulated (*magnified* for one group of 4 men and 10 women, *attenuated* for the other group of 4 men and 10 women) or left unaltered (a control condition completed by both groups). Each *Variability* task was followed by an *Adaptation* task during which participants again produced the same target words but this time while hearing auditory feedback with a consistent upward perturbation of F1 and F2 (details below). The order of completing the manipulated and control versions of the *Variability* task (each immediately followed by an identical Adaptation task) was counterbalanced across participants.

In all of the above tasks, each participant's speech output was captured with a microphone (SM 58, Shure) positioned 15 cm from the mouth and connected to an audio interface (Babyface Pro, RME, Haimhausen, Germany) and computer located outside the soundbooth (**Figure 1B**). The computer used MATLAB (The MathWorks, Natick, MA, United States) to present the visual stimuli, manipulate real-time auditory feedback when necessary, and record the participant's speech. Auditory feedback manipulations were implemented with the publicly available MATLAB software "Audapter"[1] (Cai et al., 2008; Tourville et al., 2013). The output of the audio interface was amplified (HeadAmp6 Pro, ART ProAudio, Niagara Falls, NY, United States), and played back to the participant *via* insert earphones (ER-3A, Etymotic Research Inc., Grove Village, IL, United States). Before each participant's experiment, the

[1]Audapter settings for the present study were as follows: sampling rate 48,000 Hz, downsampling factor 3, nDelay factor 3, linear prediction model order 17 for male participants and 15 for female participants. The total feedback loop latency of the specific hardware and software setup is 11.37 milliseconds (Kim et al., 2020b). Given that Audapter detects vowel onsets and offsets based on a short-time root mean square (RMS) intensity threshold, we determined the optimal RMS threshold for vowel detection for each individual participant based on visual inspection of the RMS intensity contours of the last five trials from the practice/familiarization session.

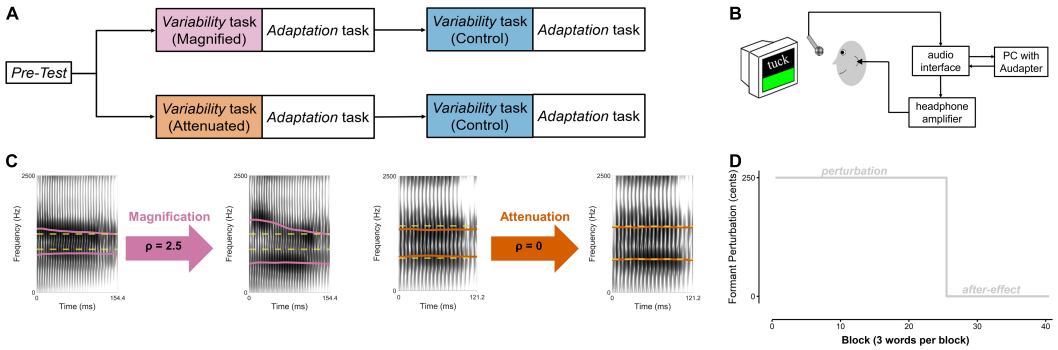


FIGURE 1 | (A) Speech tasks completed by two groups of participants. Within each group, order of the experimental condition (Magnified or Attenuated feedback variability) and the Control condition was counterbalanced across participants. **(B)** Instrumentation setup. **(C)** Example spectrograms of Difference-shifted trials in the Magnified and Attenuated conditions of the *Variability* task. Dashed yellow line: pre-test median formant frequencies (F1 and F2, in Hz). Solid magenta and orange lines: produced (left of arrow) and heard (right of arrow) formants in the Magnified and Attenuated conditions. **(D)** Time course of the formant-shift feedback perturbation in the *Adaptation* task.

feedback system was calibrated such that speech input with an intensity of 75 dB SPL at the microphone resulted in 72 dB SPL output in the earphones (Cornelisse et al., 1991). For this calibration procedure, the intensity of the auditory feedback in the earphones was measured using a 2 cc coupler (Type 4946, Bruel & Kjaer Inc., Norcross, GA, United States) connected to a sound level meter (Type 2250A Hand Held Analyzer with Type 4947 $^1/_2''$ Pressure Field Microphone, Bruel & Kjaer Inc., Norcross, GA, United States).

Pre-test

In the *Pre-test*, participants produced 30 blocks of the three target words with unaltered auditory feedback. During the production of each word, F1 and F2 were tracked by Audapter in real time. After the task was completed, a custom-written MATLAB script extracted the average F1 and F2 values (in Hz) across the middle portion of each production (defined as the window 40–60% into the vowel), calculated the across-trials median F1 and F2 for each of the participant's vowels /ɔ/ ("talk"), /ɛ/ ("tech") and /ʌ/ ("tuck"), and identified the actual production closest to the pair of F1 and F2 median values for each vowel (closeness was defined based on Euclidean distance in F1-F2 space). The mid-vowel F1 and F2 values from the participant's three productions identified in this manner–productions hereafter referred to as the pre-test medians for each vowel–were used to determine the magnitude of the feedback variability manipulation in the *Variability* task. There was a short break (∼2 min) between the *Pre-test* and the first *Variability* task.

Variability Task

Participants performed the *Variability* task once with auditory feedback in which F1 and F2 variability was experimentally manipulated (either *magnified* or *attenuated*, depending on the participant's group assignment) and once with unaltered auditory feedback as a control condition. In each *Variability* task, they produced 25 blocks of the three target words (for this first study with a variability perturbation, the number of trials was chosen based on published data regarding the number of trials that is sufficient for participants to reach maximum compensation in studies with other perturbations; see, for example, Kim et al., 2020a). In the *magnified* and *attenuated* conditions, formant variability in the auditory feedback was manipulated by modifying the difference between the formants in a given trial and the pre-test median for that vowel.

Specifically, a new mode of formant shifting, Difference-shift, was implemented by modifying Audapter's source code. In the new Difference-shift mode, the user supplies a target frequency for each formant (F^T) and a modification ratio (ρ). Within each frame, Audapter shifts the formant frequencies according to the equation $F^{fb} = F^T + \rho \times (F^c - F^T)$, where F^{fb} is the formant frequency in the feedback and F^c is the formant frequency of the current production (both in Hz). Thus, Difference-shift modifies the difference between the current formant value and the target frequency by the modification ratio. For example, if the user enters 550 Hz as the target frequency for F1 and ρ = 2.5, then for an actual F1 value of 600 Hz, the Difference-shift mode shifts the output F1 to 675 Hz (550 + 2.5 50). When $\rho = 1$, the Difference-shift mode magnifies the difference between the produced formant value and the target frequency, whereas the difference is attenuated when $\rho < 1$.

In both the Magnified and the Attenuated conditions, the pre-test median of F1 and F2 for each vowel was supplied as the target formant frequency F^T. To magnify the difference between the current production and the target formant frequency, ρ was set to 2.5 in the Magnified condition. To minimize the difference between the current production and the target, ρ was set to 0 in the Attenuated condition. Examples of individual productions and the corresponding manipulated feedback for each condition are included in **Figure 1C**. Note that if $\rho = 0$, the Difference-shift would theoretically always shift the formant frequency to the target frequency, regardless of the current production. However, due to the intrinsic limitations of real-time formant tracking and the digital filtering techniques used to alter the signal, the actual ratio between produced frequency and Difference-shifted output frequency is not always identical to the supplied modification ratio. Given this situation

that, in reality, $\rho = 0$ reduces (but does not completely eliminate) feedback variability, it was chosen as the preferred ratio for the Attenuated condition. The overall effectiveness of the feedback perturbation for magnifying and attenuating feedback formant variability is described below in the Section "Results."

Adaptation Task

Each *Adaptation* task followed immediately after one of the *Variability* tasks, and was identical after the manipulated and control versions of the *Variability* task. In both cases, it consisted of a perturbation phase (25 blocks) and an after-effect phase (15 blocks) (**Figure 1D**). No variability manipulation was applied, but, at the start of the perturbation phase, a sudden 250 cents[2] upshift of F1 and F2 was introduced by Audapter. This formant shift was turned off, and participants received unaltered auditory feedback, during the after-effect phase. There was a short break (∼2 min) between the end of the first *Adaptation* task and the beginning of the second *Variability* task.

Data Extraction and Analysis

The speech signal from all tasks (*Pre-test* task, *Variability* tasks, and *Adaptation* tasks) was digitized by Audapter. Using a custom-written MATLAB script, we examined the production data from all tasks offline to exclude productions containing production errors (e.g., mispronunciations or yawning; 0.45% of productions were rejected for this reason), manually marked the onset and offset of the vowel in each production based on visual inspection of its waveform and spectrogram, and extracted the first two formant frequencies (F1 and F2) as tracked by the linear predictive coding algorithm implemented in Praat (Boersma, 2001). To disentangle feedforward adaptive learning vs. online feedback-driven corrections within trials, F1 and F2 formant values for each trial were extracted both across an initial portion of the vowel (5–30% into its total duration) and a middle portion of the vowel (40–60% into total duration). Additionally, to verify accuracy of the auditory feedback manipulation in the experimental conditions of the *Variability* task (Magnified and Attenuated variability), we extracted F1 and F2 also across the same middle portion of the vowel in the recorded feedback signal.

Statistical analyses for the *Variability* task and the *Adaptation* task made use of paired two-sample *t*-tests or, in a few cases, one-sample *t*-tests, with the significance level set at 0.05. When multiple statistical comparisons were carried out as one family of tests, *p*-values were adjusted with the Holm–Bonferroni method (Holm, 1979). Cohen's *d* was used for effect size calculations (Cohen, 1988). All statistical tests were conducted in the R software (R Core Team, 2019).

Analysis of the Variability Task

Formant frequencies measured for the initial and middle portions of vowels from the *Variability* task were normalized by conversion from Hz to cents. The medians (F1 and F2) of each vowel from each subject's pre-test productions, also measured offline across the initial and the middle portions separately, were chosen as the reference frequency for the conversion. Similarly, the formants measured from the middle portion of the vowel in the auditory feedback signal were also converted with reference to each subject's pre-test median frequencies for the middle portion.

A primary focus of analysis for the *Variability* task was the participants' production variability. To quantify this production variability with a measure directly related to the nature of the perturbation itself (i.e., distance to the pre-test median formants), we formulated a distance index (DI), $DI = \sqrt{F1^2 + F2^2}$, where F1 and F2 are a trial's formant frequencies already expressed in cents relative to the pre-test median. For each production, two DI's, $DI_{initial}$ and DI_{mid}, were calculated with the formant values that had been extracted from the non-overlapping initial and middle portions of that trial's vowel. For the auditory feedback signal, there was only one DI measurement per trial, DI_{fb}, as formant frequencies had been extracted only for the middle portion of the vowel.

First, to verify the effectiveness of our formant feedback variability magnification and attenuation by the Difference-shift implementation in Audapter, the ratio between the average DI_{fb} and average DI_{mid} of each participant's experimental *Variability* task was compared to the ideal ratio based on the perturbation algorithm (assuming perfect formant tracking and signal processing). Second, to examine the effect of feedback variability manipulation on production variability (Wong et al., 2009), we compared both $DI_{initial}$ and DI_{mid} between the Control condition and the experimental (Magnified or Attenuated) conditions. To explore the possibility of gradual changes in production variability during the course of the *Variability* task, these variables were considered not only for the whole task (25 blocks of 3 trials each) but also block-by-block and stage-by-stage (with a stage operationally defined as a series of 5 consecutive blocks). Third, to examine possible online feedback-based corrections in response to the variability manipulations, we also calculated the within-trial difference between $DI_{initial}$ and DI_{mid} [note that this approach shows similarities with the "centering" measure used in previous studies of online feedback corrections (Niziolek et al., 2013; Niziolek and Kiran, 2018), but differs from it in that our DI measures determine each trial's distance to the median production from the *Pre-test* in cents rather than distance to the median production of the analyzed dataset itself in mels]. For each experimental condition (Magnified, Attenuated) and each control condition (completed by the Magnified and Attenuated groups separately), we used one-sample *t*-tests to determine whether the within trial changes in DI were statistically significantly different from zero (i.e., whether or not "centering" toward the pre-test median occurred). For each group separately, we then used paired *t*-tests to determine whether any within-trial changes differed between the experimental and control condition.

Although analogous to the nature of the variability perturbation itself, one potential problem with the DI-based analysis is that it is theoretically possible for a participant to increase or decrease the average distance between their trial

[2]The conversion formula between cents and Hz is: $F_{cents} = 1200 \times \log_2(\frac{F_{Hz}}{R_{Hz}})$, where R_{Hz} is a reference frequency. 100 cents = 1 semitone. For the perturbation in the *Adaptation* part, $F_{cents} = 250$, $R_{Hz} = F^c$, and $F_{Hz} = F^{fb}$. A 250 cents upshift approximately equals a 15.5% increase in Hz.

formant frequencies and the pre-test medians without increasing the actual dispersion of these trials in two-dimensional (F1, F2) acoustic space. For example, although extremely unlikely for real speech, it is theoretically possible that the formants for all trials could be moved further away from the pre-test median (thereby increasing DI) but always to the same location in acoustic space. For this reason, we followed up on statistically significant DI effects by also determining for each participant the size of the area in acoustic space covered by the relevant productions (i.e., trials produced in the Control condition or in a given stage of the experimental conditions). The size of this area was determined by means of 95% confidence ellipses, calculated based on formant frequencies from the initial portion of the vowels.

A secondary focus of the *Variability* task was to investigate possible effects of the variability manipulations on the temporal structure of formant adjustments across trials. Consistent with the approach used in previous non-speech studies (van Beers, 2009; van der Vliet et al., 2018), we compared between Control and experimental conditions the sample lag 1 autocorrelation function, ACF(1), calculated for the sequence of averaged formant frequencies (i.e., mean of F1 and F2) obtained at the initial portion of the vowel in each trial. Formally, $ACF\,(1) = \frac{1}{N}\frac{\sum_{n=1}^{N-1}(F[n+1]-\overline{F})(F[n]-\overline{F})}{\sum_{n=1}^{N}(F[n]-\overline{F})^2}$, where $N = 75$, $F[n] = (F1_{initial}[n] + F2_{initial}[n])/2$, and $\overline{F} = \frac{1}{N}\sum_{n=1}^{N} F[n]$.

Analysis of the Adaptation Task

Given that adaptation refers to adjustments in movement planning based on prior experience (as opposed to online feedback-driven corrections), only the formant frequencies measured at the initial portion of the vowel were used for analysis of the *Adaptation* task. These formant frequencies were normalized to cents with reference to the median formants of each vowel in blocks 16–25 of the *Variability* task immediately prior to the onset of the *Adaptation* task. The frequencies of F1 and F2, in cents, were averaged for each trial as in several of our prior studies (e.g., Kim et al., 2020a; Shiller et al., 2020).

We compared three metrics between the perturbation phases from the Control and experimental conditions: early adaptation extent, early adaptation rate, and final adaptation extent. Early adaptation extent was calculated by determining the average formant frequency across the first 15 trials of the perturbation phase. Early adaptation rate was defined as the slope of a linear regression function based on the formant frequencies of the same 15 trials. Final adaptation extent was calculated by determining

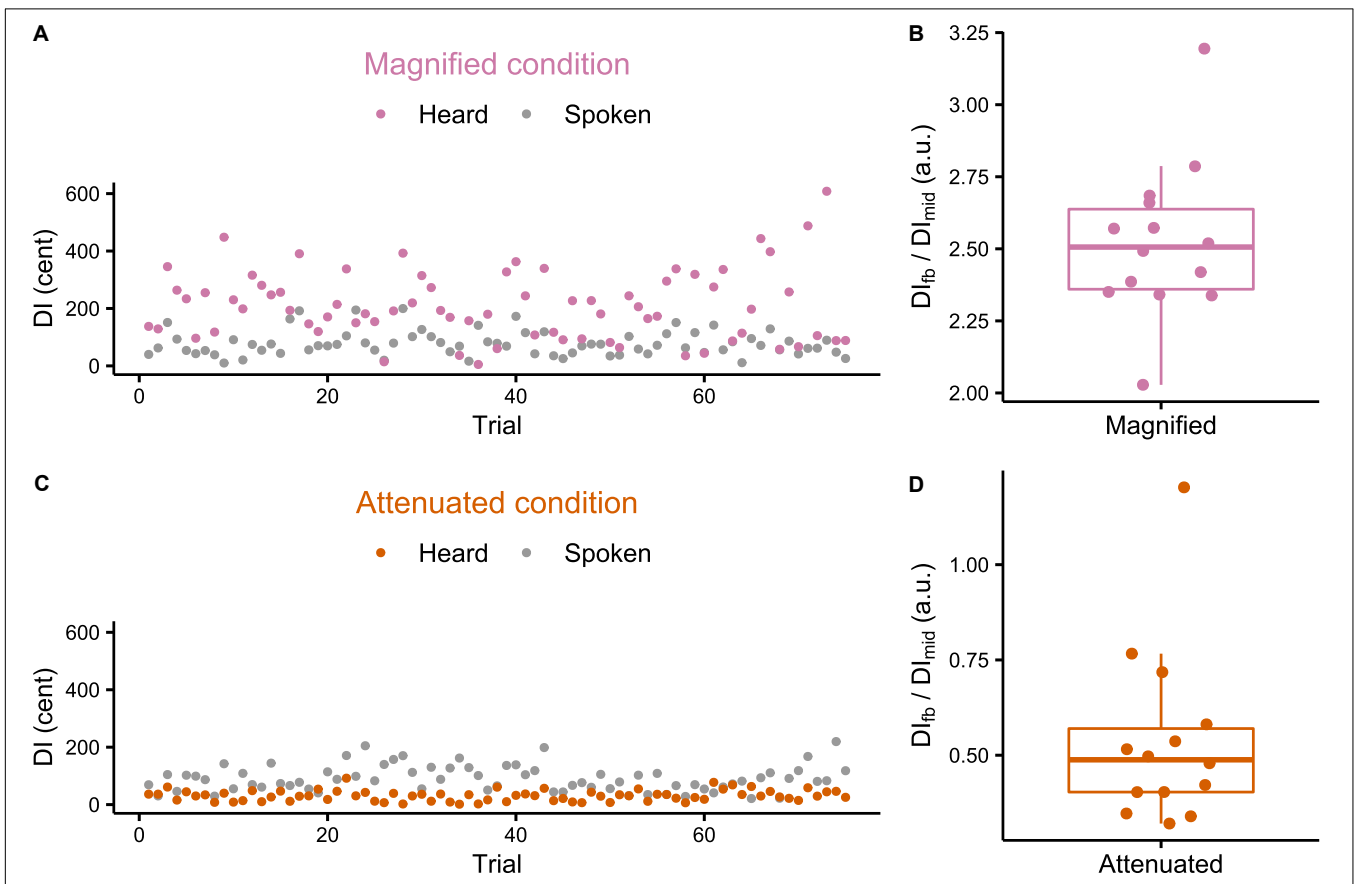


FIGURE 2 | (A,C) Example individual participant data for production and feedback Distance Index (DI) of each trial in the *Variability* task under Magnified and Attenuated conditions. **(B,D)** Boxplots with symbols depicting each participant's ratio between average feedback DI and average production DI (both measured mid-vowel) across all trials of the *Variability* task with Magnified or Attenuated feedback variability.

the average formant frequency across the last 15 trials of the perturbation phase of the task.

RESULTS

Variability Task

Effectiveness of the Feedback Variability Manipulations

Individual participant data for DI calculated for both the produced and heard trials from the *Variability* task are presented in **Figure 2**. **Figures 2A,C** each show that the feedback manipulation was effective for the two selected participants from the Magnified and Attenuated conditions, respectively. **Figures 2B,D** show for *all* individual participants the ratio between the average DI of the formants in the manipulated feedback, DI_{fb}, and that of the produced formants, DI_{mid}, for the Magnified and Attenuated conditions, respectively. In the Magnified condition, the group mean of this ratio was 2.52 ($SD = 0.27$), a value very close to the intended modification ratio $\rho = 2.5$ (which is also the theoretical value of DI_{fb}/DI_{mid} if the Difference-shift had worked perfectly in every frame of every trial). In the Attenuated condition (ratio $\rho = 0$), however, there was one outlier participant ($DI_{fb}/DI_{prod} = 1.202$) for whom the Difference-shift mode failed to achieve the goal of attenuating formant variability in the auditory feedback. With the outlier removed, the group mean of the DI_{fb}/DI_{mid} ratio was 0.49 ($SD = 0.14$) and all remaining ratios were less than 1, confirming that the goal of attenuating feedback variability was achieved. The data from the participant with the unsuccessful feedback perturbation in the Attenuated condition were excluded from all further analyses.

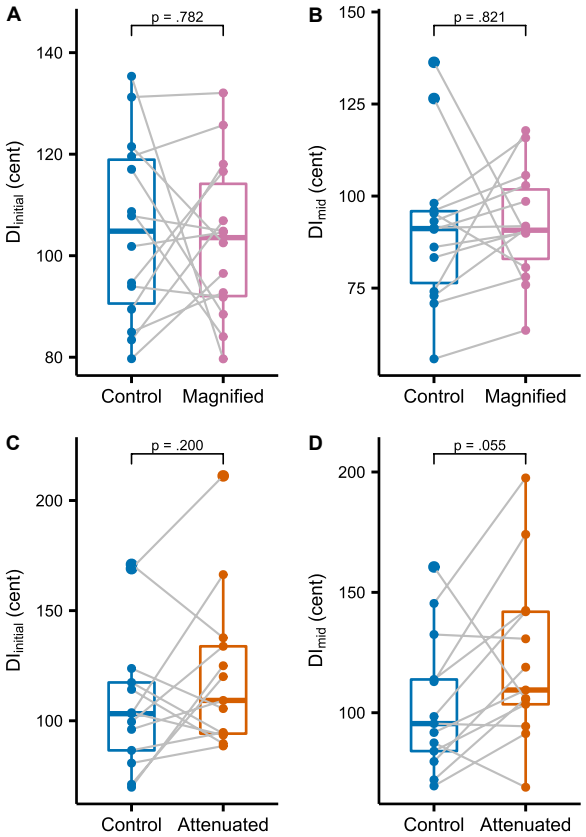


FIGURE 3 | (A,C) Boxplots with symbols depicting each participant's average $DI_{initial}$ for the entire *Variability* task in Control and Magnified or Control and Attenuated conditions. **(B,D)** Boxplots with each participant's average DI_{mid} for the entire *Variability* task in the same conditions.

Production Variability

The first set of production variability analyses compared the Control condition with both experimental conditions at the whole-task level for the target vowel's initial portion ($DI_{initial}$, **Figure 3A** for the Magnified condition, **Figure 3C** for the Attenuated condition) and middle portion (DI_{mid}, **Figures 3B,D**). As compared with the Control condition, no statistically significant change in $DI_{initial}$ was found for either the Magnified or the Attenuated condition [$t(13) = -0.282$, $p = 0.782$, $d = 0.075$, and $t(12) = 1.358$, $p = 0.200$, $d = -0.376$, respectively]. Similarly, there was also no significant change in DI_{mid} for either the Magnified or Attenuated condition [$t(13) = 0.231$, $p = 0.821$, $d = -0.062$, and $t(12) = 2.122$, $p = 0.055$, $d = 0.588$, respectively].

The second set of production variability analyses examined whether a response to the auditory feedback manipulations might develop over time with continuing exposure. Therefore, these analyses considered the time course of the $DI_{initial}$ and DI_{mid} variables per block of 3 trials and per stage of 5 blocks. **Figure 4** shows group data for the change in $DI_{initial}$ from block to block (**Figures 4A,B**) and stage to stage (**Figures 4C,D**) under the Control and experimental conditions. For the group that completed Control and Magnified conditions, the data show no change in formant production $DI_{initial}$ within either of those conditions. Statistical comparisons of $DI_{initial}$ between the first stage and each of the following stages confirmed the absence of an adjustment in this distance metric with Magnified feedback variability (**Table 1** and **Figure 4C**). In contrast, for the group that completed Control and Attenuated conditions, $DI_{initial}$ showed a statistically significant increase from Stage 1 to Stage 2 and from Stage 1 to Stage 3 in the condition with Attenuated feedback variability whereas no statistically significant change was observed in the same group's Control condition (**Table 1** and **Figure 4D**). Visualizations of the Attenuated condition individual participant data for $DI_{initial}$ in Stage 1 and Stage 3, and of the extent and direction for individual changes in this variable over the same time period, are included in **Figures 4E–H** (analysis techniques based on Wilcox and Erceg-Hurn, 2012; Bieniek et al., 2016; Rousselet et al., 2017). The data show a robust trend across individuals as 11 of 13 participants increased their formant production $DI_{initial}$ in the first half of the task with Attenuated feedback variability.

Similar results were obtained when considering DI_{mid} from block to block (**Figures 5A,B**) and stage to stage (**Figures 5C,D**): DI_{mid} showed no change in either group's Control condition, also no change in the Magnified condition, but a statistically significant increase from Stage 1 to Stages 2, 3, and 5 in the

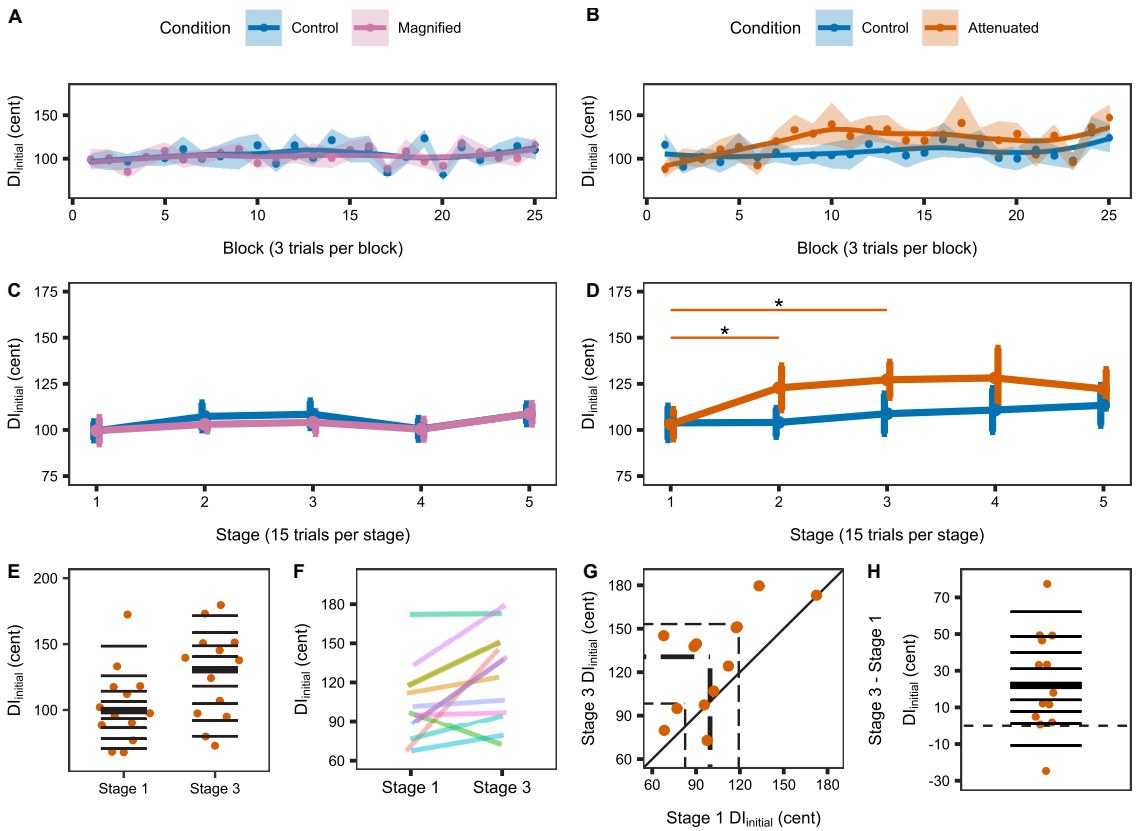


FIGURE 4 | (A,B) Change in $DI_{initial}$ across the Variability task by block (i.e., 3 trials) for the Magnified and Attenuated feedback variability conditions. Dots represent the group mean DI per block. Shaded regions indicate standard error of the mean (SEM). Solid lines are loess smoothed fits (span = 0.6). **(C,D)** Change in $DI_{initial}$ across the Variability task by stage (i.e., 15 trials) for the Magnified and Attenuated feedback variability conditions. Error bars indicate SEM. Asterisks indicate adjusted $p < 0.05$ (see **Table 1**). **(E–H)** Individual participant data for the significant change from Stage 1 to Stage 3 in the Attenuated condition: **(E)** Stripchart of $DI_{initial}$ in Stage 1 and Stage 3. Horizontal lines indicate deciles; bold line is the median. **(F)** Stripchart with each participant's Stage 1 and Stage 3 data linked. **(G)** Scatterplot of Stage 1 by Stage 3 data. The diagonal line denotes no difference between stages. Participants in the upper left half increased $DI_{initial}$ in Stage 3. Dashed lines mark quartiles. **(H)** Stripchart of the difference in $DI_{initial}$ between Stage 3 and Stage 1. Horizontal lines indicate deciles; the bold line is the median; the dashed line is at zero (no difference between stages).

Attenuated condition (**Table 1** and **Figure 5D**). The individual participant data for Stage 1 and Stage 3 in this condition with Attenuated feedback variability show a highly consistent increase in formant production DI_{mid} during the first half of the task (**Figures 5E–H**).

We examined the change from $DI_{initial}$ to DI_{mid} as an indicator of potential within-trial corrections in the conditions with Magnified or Attenuated formant variability in the auditory feedback. For participants assigned to the Attenuated group, within-trial changes were not statistically significantly different from zero for either the experimental condition [$t(12) = 0.164$, $p = 0.872$, $d = 0.046$] or the control condition [$t(12) = -1.285$, $p = 0.446$, $d = -0.356$]. For participants in the Magnified group, within-trial changes were statistically significant, but this was the case for both the experimental condition [$t(13) = -4.117$, $p = 0.002$, $d = -1.100$] and the control condition [$t(13) = -3.650$, $p = 0.003$, $d = -0.975$]. For neither group were within-trial changes in the experimental condition statistically different from those in the control condition with unaltered feedback variability [Attenuated group: $t(12) = -0.997$, $p = 0.339$, $d = -0.288$; Magnified group: $t(13) = -0.962$, $p = 0.354$, $d = 0.264$].

Given that the Attenuated condition showed a statistically significant increase in $DI_{initial}$ (as well as DI_{mid}) from Stage 1 to Stage 3, **Figure 6** shows the individual participants' inter-trial dispersion of formant frequencies in 2D (F1, F2) acoustic space for Stages 1 and 3 of the Attenuated variability condition together with equivalent data from the *Pre-test*. All data were extracted from the initial portion of the vowels. Although the comparison of 95% confidence ellipse areas for Stage 1 versus Stage 3 did not reach statistical significance [$t(12) = -1.894$, $p = 0.083$], this comparison was associated with a medium effect size ($d = -0.525$), and 9 of 13 individual participants increased the ellipse area in Stage 3 as compared with Stage 1. Of the four participants who decreased ellipse size, only two showed a change that fell within the range of changes (but with opposite sign) observed for the subjects with increasing ellipses; the other two subjects showed only minimal changes.

TABLE 1 | Adjusted *p*-values (paired *t*-tests, Holm–Bonferroni method) for comparisons of $DI_{initial}$ (top section) and DI_{mid} (bottom section) between Stage 1 (first 5 blocks of 3 trials) and each subsequent stage (also 15 trials) in the Control, Magnified, and Attenuated feedback variability conditions (the two participant groups completing Magnified or Attenuated variability conditions each completed their own Control conditions, labeled Control M and Control A).

	Stage 1 vs. 2	Stage 1 vs. 3	Stage 1 vs. 4	Stage 1 vs. 5
		$DI_{initial}$		
Control M	$t(13) = -1.519$, $p = 0.612$, $d = -0.406$	$t(13) = -1.278$, $p = 0.612$, $d = -0.342$	$t(13) = -0.234$, $p = 0.819$, $d = -0.062$	$t(13) = -1.514$, $p = 0.612$, $d = -0.405$
Magnified	$t(13) = -0.506$, $p = 1.000$, $d = -0.135$	$t(13) = -0.680$, $p = 1.000$, $d = -0.182$	$t(13) = -0.111$, $p = 1.000$, $d = -0.030$	$t(13) = -1.717$, $p = 0.440$, $d = -0.459$
Control A	$t(12) = -0.041$, $p = 1.000$, $d = -0.011$	$t(12) = -0.561$, $p = 1.000$, $d = -0.156$	$t(12) = -1.153$, $p = 1.000$, $d = -0.320$	$t(12) = -0.944$, $p = 1.000$, $d = -0.262$
Attenuated	$t(12) = -2.800$, $p = 0.048$*, $d = -0.777$	$t(12) = -3.189$, $p = 0.031$*, $d = -0.884$	$t(12) = -2.330$, $p = 0.076$, $d = -0.646$	$t(12) = -2.051$, $p = 0.076$, $d = -0.569$
		DI_{mid}		
Control M	$t(13) = -0.426$, $p = 1.000$, $d = -0.114$	$t(13) = -0.077$, $p = 1.000$, $d = -0.021$	$t(13) = -1.347$, $p = 0.804$, $d = -0.360$	$t(13) = -0.918$, $p = 1.000$, $d = -0.245$
Magnified	$t(13) = 0.473$, $p = 1.000$, $d = 0.126$	$t(13) = 0.676$, $p = 1.000$, $d = 0.181$	$t(13) = 0.501$, $p = 1.000$, $d = 0.134$	$t(13) = -2.526$, $p = 0.101$, $d = -0.675$
Control A	$t(12) = 0.659$, $p = 1.000$, $d = 0.183$	$t(12) = -0.126$, $p = 1.000$, $d = -0.035$	$t(12) = -1.133$, $p = 1.000$, $d = -0.314$	$t(12) = -0.427$, $p = 1.000$, $d = -0.118$
Attenuated	$t(12) = -3.039$, $p = 0.021$*, $d = -0.843$	$t(12) = -4.996$, $p = 0.001$*, $d = -1.386$	$t(12) = -1.929$, $p = 0.078$, $d = -0.535$	$t(12) = -3.406$, $p = 0.016$*, $d = -0.945$

Statistically significant differences (*) were found only for the Attenuated feedback variability manipulation, in particular for the comparisons Stage 1 vs. Stage 2 and Stage 1 vs. Stage 3.

Autocorrelation Structure

To assess the temporal structure of formant adjustments across the entire series of productions in the manipulated auditory feedback conditions, we determined the sample lag 1 autocorrelation [ACF(1)] of the time series consisting of averaged F1 and F2 values from the initial vowel portion of each trial in the *Variability* task (**Figure 7**). It should be noted that the large sample 95% confidence interval of ACF(1) for a white noise process with sample size $N = 75$ (i.e., the number of trials in each analyzed time series) is (-0.22, 0.22) (Brockwell and Davis, 2016). Most of the individual ACF(1) data from all conditions in the current study fell within this bound, indicating that, from a statistical perspective, it is likely that most production sequences were generated by white noise processes. There were no statistically significant differences in ACF(1) between either of the two experimental conditions and the Control condition [Magnified: $t(13) = 0.670$, $p = 0.515$, $d = 0.179$; Attenuated: $t(12) = -0.324$, $p = 0.752$, $d = 0.090$].

Adaptation Task

Figure 8 shows group mean formant frequencies produced throughout the *Adaptation* tasks that followed immediately after different conditions of the *Variability* task (data are in cents relative to the end of the preceding *Variability* task, measured at the initial portion of the vowel, and averaged across F1 and F2 and across the 3 trials per block). Recall that separate groups of participants completed the Magnified and Attenuated experimental conditions of the *Variability* task, and that, therefore, each group completed their own Control condition of the *Variability* task with no feedback perturbation. The Control versus experimental condition within-group comparisons in **Figure 8** suggest that adaptation was not affected by the prior formant feedback variability manipulations. Statistical testing confirmed the absence of any significant differences between Control and Magnified or between Control and Attenuated for early adaptation extent (average formant frequency of the first 15 adaptation trials; **Figure 9A**), learning rate during early adaptation (slope of a linear regression line over the formant frequencies of the first 15 adaptation trials; **Figure 8B**), or final adaptation extent (average formant frequency of the last 15 perturbation trials; **Figure 8C**). The *p* values for all statistical comparisons are included with the data visualizations in **Figure 9**.

DISCUSSION

Previous observational studies have led to the suggestion that inter-trial motor variability may be related to both enhanced online feedback-based compensation (a study on fundamental frequency in speech, Scheerer and Jones, 2012) and enhanced adaptive learning (a study on upper limb reach movements, Wu et al., 2014). However, neither of these results have been consistently supported by other empirical data (Scheerer and Jones, 2012; He et al., 2016; Singh et al., 2016), alternative explanations for the findings have been offered (He et al., 2016; Singh et al., 2016; Murillo et al., 2017; Sternad, 2018; van der Vliet et al., 2018), and further investigation is clearly warranted (Dhawale et al., 2017). Moreover, results from an experimental study that directly manipulated feedback variability for reaching movements by magnifying or attenuating the size of target errors suggested that the temporal structure of adjustments across trials, indexed by the sample lag 1 autocorrelation [ACF(1)] for movement endpoints, changed with manipulated feedback (van Beers, 2009). In the same study, the adjustments across trials were consistent with predictions made by state-space models often used to characterize learning mechanisms in sensorimotor adaptation experiments (van Beers, 2009). Thus, inter-trial motor variability itself may represent a form

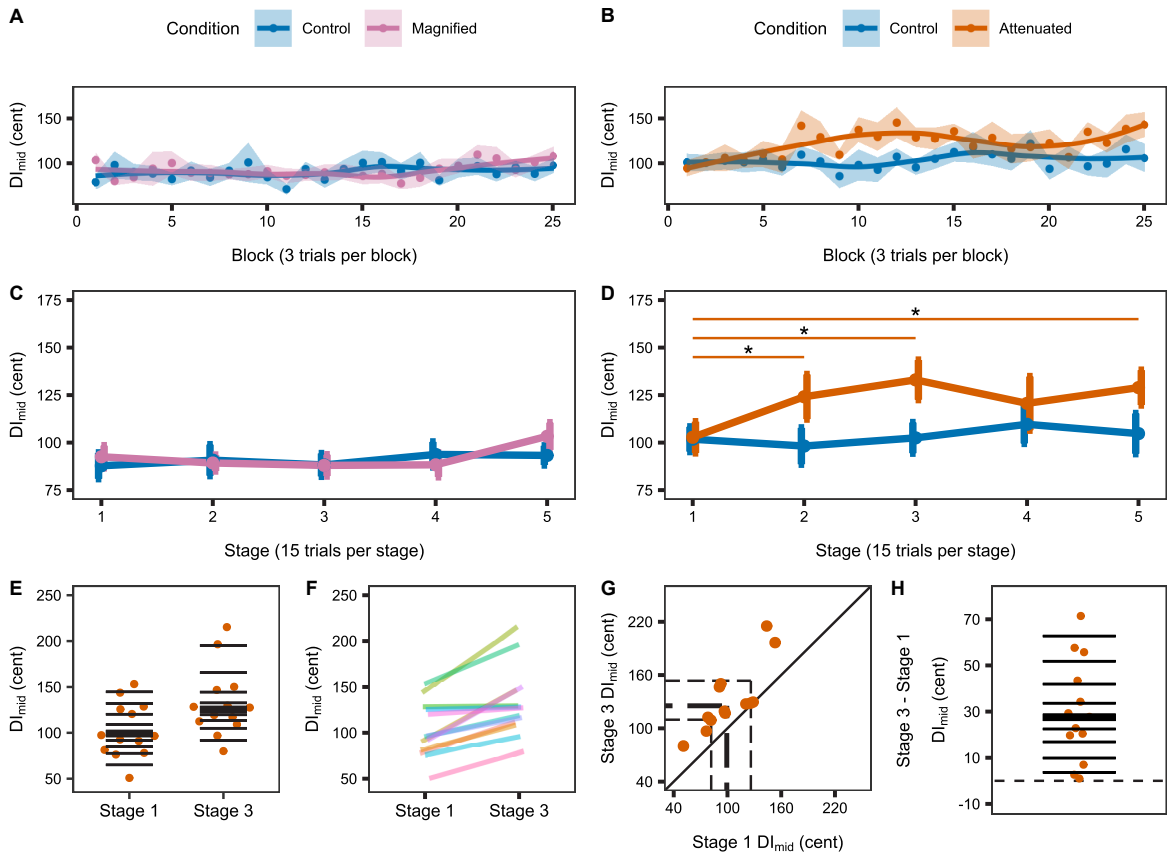


FIGURE 5 | (A,B) Change in DI_{mid} across the Variability task by block (i.e., 3 trials) for the Magnified and Attenuated feedback variability conditions. Dots represent the group mean DI per block. Shaded regions indicate standard error of the mean (SEM). Solid lines are loess smoothed fits (span = 0.6). **(C,D)** Change in DI_{mid} across the Variability task by stage (i.e., 15 trials) for the Magnified and Attenuated feedback variability conditions. Error bars indicate SEM. Asterisks indicate adjusted $p < 0.05$ (see **Table 1**). **(E–H)** Individual participant data for the significant change from Stage 1 to Stage 3 in the Attenuated condition: **(E)** Stripchart of DI_{mid} in Stage 1 and Stage 3. Horizontal lines indicate deciles; bold line is the median. **(F)** Stripchart with each participant's Stage 1 and Stage 3 data linked. **(G)** Scatterplot of Stage 1 by Stage 3 data. The diagonal line denotes no difference between stages. Participants in the upper left half increased DI_{mid} in Stage 3. Dashed lines mark quartiles. **(H)** Stripchart of the difference in DI_{mid} between Stage 3 and Stage 1. Horizontal lines indicate deciles; the bold line is the median; the dashed line is at zero (no difference between stages).

of trial-by-trial learning. On the other hand, the authors of a reaching movement experiment combining error feedback magnification or attenuation with a constant perturbation that elicits visuomotor adaptation concluded that variability manipulation did not alter the underlying adaptive learning mechanisms (van der Kooij et al., 2015), despite observed changes in adaptation behavior (Patton et al., 2013; van der Kooij et al., 2015).

We sought to clarify, for sensorimotor control of speech articulation, whether *experimental manipulations* of inter-trial feedback variability (here variability of formant frequencies in the real-time auditory feedback) (a) lead to speaker adjustments in inter-trial production variability, suggestive of an active regulation mechanism; (b) lead to changes in the temporal structure of adjustments across trials [ACF(1)], suggestive of trial-by-trial learning; and (c) affect learning in a subsequent auditory-motor adaptation paradigm with a constant formant-shift perturbation. To manipulate inter-trial formant variability in the feedback, we implemented a novel real-time formant manipulation algorithm that can either magnify or attenuate the difference between the formants in a current production and target formants operationally defined as the median formant values from a *Pre-test*.

Active Regulation of Variability

After the *Pre-test* with unaltered auditory feedback, participants completed two conditions of a *Variability* task (each followed by an *Adaptation* task): one was a Control condition with unaltered formant feedback, and the other condition had either Magnified or Attenuated formant variability in the auditory feedback, depending on the participant's group assignment. Signal processing algorithms generating the feedback signal in these experimental conditions increased or decreased the distance between the formants produced in a given trial and the participants' median formants for the same word in the *Pre-test*. We therefore quantified participants' productions with a DI that expressed produced formant frequencies also in terms of their distance to the pre-test median.

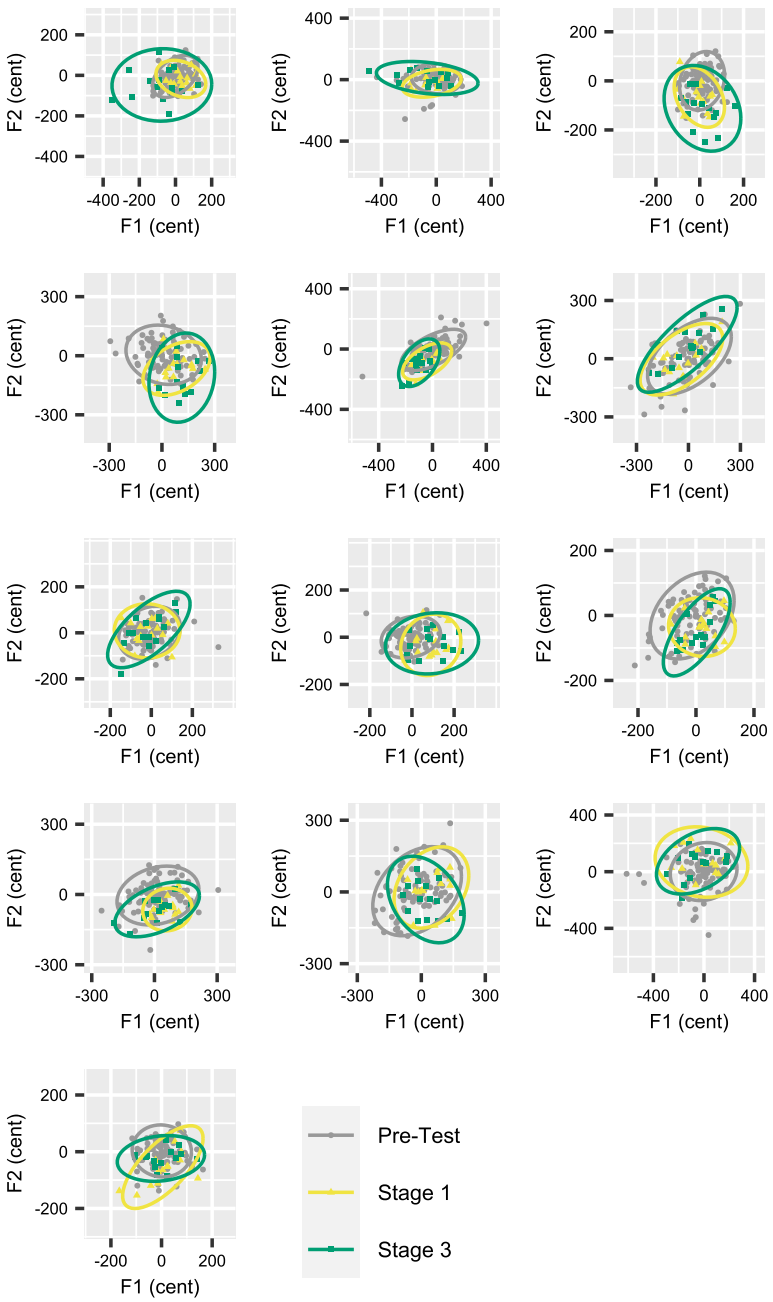


FIGURE 6 | Individual participant data (one participant per panel) for inter-trial formant dispersion in acoustic vowel space (F1 by F2). Data based on 95% confidence ellipses, calculated for formant frequencies extracted from the initial portion of the vowels. Each participant's data from Stages 1 and 3 (15 trials per stage) in the Attenuated feedback variability condition are shown together with their data from the *Pre-test* (90 trials). Nine of 13 participants increased ellipse area in Stage 3 as compared with Stage 1. Participants are ordered (by row) from greatest to smallest ellipse area increase.

Compared with each group's own Control condition, the condition with Magnified feedback variability did not result in an adjustment in distance, but the condition with Attenuated feedback variability led to a gradual *increase* in distance between produced trials and the pre-test median (thus opposing the feedback manipulation). This increasing distance between produced formants and pre-test median formants was detected in both the initial portion of the vowel (5–30% into the total vowel duration; results in **Figure 4**) and the middle portion of the vowel (40–60% into the total vowel duration; results in **Figure 5**) portions of the vowel, and, thus, reflects gradual changes in movement planning rather than online within-vowel corrections. In fact, neither of the experimental conditions affected the extent of within-vowel corrections as compared with the same

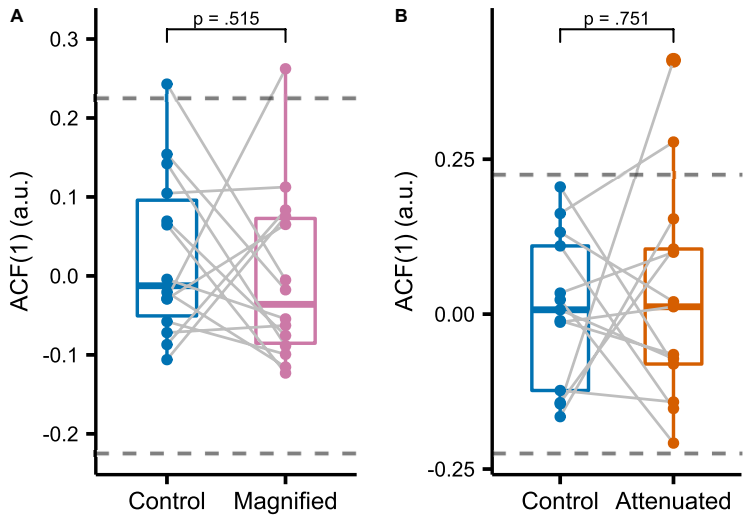


FIGURE 7 | Sample lag 1 autocorrelation functions [ACF(1)] for formant data measured in the initial portion of the vowel and averaged across F1 and F2 for Control versus Magnified **(A)** and Control versus Attenuated **(B)** conditions of the *Variability* task. Dashed lines indicate the large sample 95% confidence interval of ACF(1) for a white noise process with sample size 75 (the number of trials per condition). Each dot represents an individual participant.

participants' Control condition. As it is theoretically possible for DI to increase even in the absence of an increase in variability (e.g., if a participant moved their formants further from the pre-test median but always to the same location in acoustic F1F2 space), we followed up by determining the size of the area in acoustic space covered by each participants' productions. This analysis confirmed that during the early stages of exposure to Attenuated variability feedback, most—but not all—participants did actually increase the overall spread of their productions in the two-dimensional acoustic space (i.e., increased formant production variability; results in **Figure 6**).

It is not straightforward to compare this finding of active variability regulation with those from prior limb motor control studies that magnified and/or attenuated the dispersion of feedback across trials as a by-product of manipulating the magnitude of target error in each trial. The study by Wong et al. (2009) only *increased* the size of perceived target errors (and thus feedback dispersion), and, consequently, one cannot necessarily attribute the resulting decrease in motor variability to the magnified feedback variability as opposed to a control strategy that seeks to avoid large errors on each trial individually. The study by van Beers (2009) did implement both magnified and attenuated target errors, but focused on the temporal structure of movement endpoint adjustments across trials (see below Section "Temporal Structure"). Nevertheless, for reaching movements with unperturbed visual feedback, van Beers (2009) reported that trial-to-trial adjustments are made in such a way that movement variability is minimized.

Our data from speech articulation are not consistent with the idea that the central nervous system generally aims to minimize variability. In fact, these data suggest a strikingly different situation: when the feedback perturbation magnified inter-trial formant variability, this extended variability was tolerated and not opposed, but when the perturbation attenuated inter-trial formant variability, articulation was gradually adjusted such that the acoustic output counteracted the perturbation. Thus, overall, the present data are consistent with the interpretation that a sufficiently large level of feedback variability is desirable, and that this level of variability is actively regulated through adjustments in motor planning.

In light of this overall support for the hypothesis that variability is actively regulated, it is reasonable to wonder why the increase in production variability in the Attenuated condition was not statistically significant in some of the later stages of the task. As shown in **Figure 4**, the increase in $DI_{initial}$ relative to stage 1 was significant in stages 2 and 3 but not in stages 4 and 5 (in both cases $p = 0.076$ with medium effect sizes). Closer inspection reveals that, at Stage 4, the mean $DI_{initial}$ value had further increased, but the standard error of the mean was also larger at this stage. At Stage 5, the mean $DI_{initial}$ value did decrease, but it never returned to its original value from stage 1. As shown in **Figure 5**, the increase in DI_{mid} relative to stage 1 was still statistically significant in the last stage of the task, only not in the preceding Stage 4 ($p = 0.078$, medium effect size). Thus, there was a trend for the increased production variability to be not sustained at its maximum level in the later stages of the task, but any attempts at interpreting the specific results for Stage 4 would be purely speculative.

It should also be acknowledged that an alternative explanation might be offered for the absence of formant variability regulation in the Magnified condition of our *Variability* task. Specifically, one could argue that the highly practiced speech movements may have been performed with minimized variability from the very beginning of the task, and that, therefore, a floor effect prevented further reduction of this variability in the Magnified condition. This would be a reasonable argument as the lower bound of variability seems to be physiologically constrained by the stochastic nature of events in the peripheral motor system

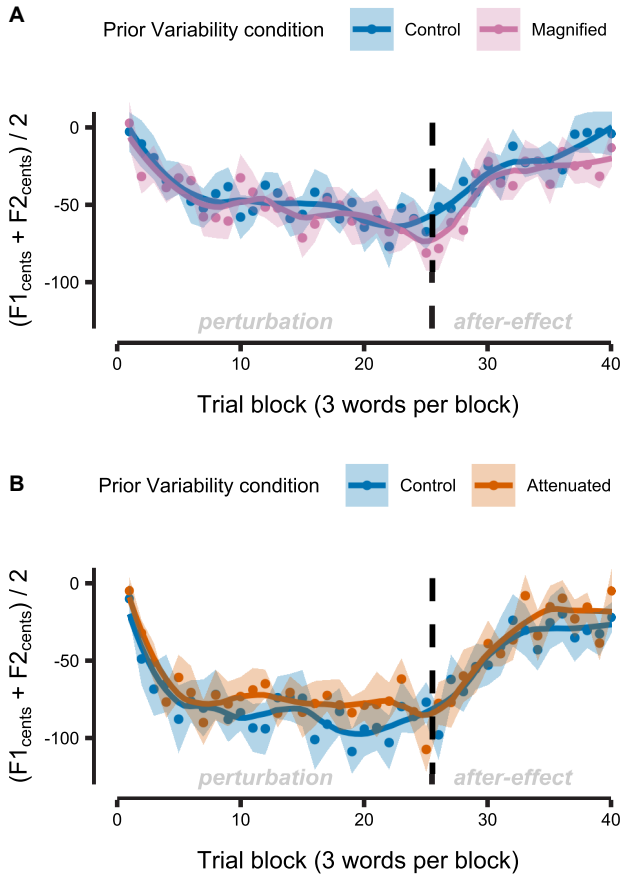


FIGURE 8 | Group-level formant-shift adaptation data after completion of the *Variability* task's Control and Magnified conditions **(A)** or after the Control and Attenuated conditions **(B)**. Dots represent group mean formant frequencies per block (3 trials) and averaged across F1 and F2. Shaded regions indicate standard error of the mean. Solid lines are loess smoothed fits (span = 0.3).

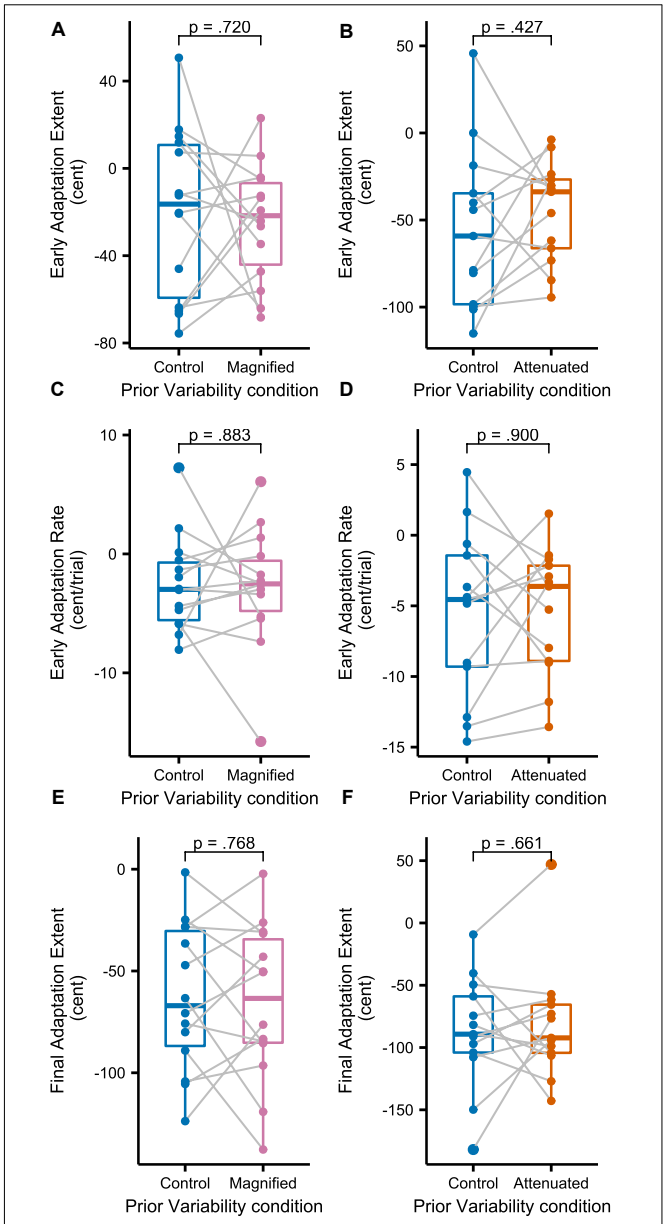


FIGURE 9 | Boxplots with symbols depicting each participant's early adaptation extent **(A,B)**, early adaptation rate **(C,D)**, and final adaptation extent **(E,F)** for formant-shift adaptation completed after the *Variability* task's Control and Magnified or Control and Attenuated conditions. Full statistics for these data: **(A)** $t(13) = 0.366$, $p = 0.720$, $d = 0.098$; **(B)** $t(12) = -0.822$, $p = 0.427$, $d = 0.228$; **(C)** $t(13) = 0.150$, $p = 0.883$, $d = 0.040$; **(D)** $t(13) = -0.128$, $p = 0.900$, $d = 0.035$; **(E)** $t(13) = 0.301$, $p = 0.768$, $d = 0.081$; **(F)** $t(12) = -0.450$, $p = 0.661$, $d = 0.125$.

such as synaptic transmission (Calvin and Stevens, 1968) and muscle contraction (Clamann, 1969; Hamilton et al., 2004), which together are referred to as execution or performance noise in theoretical models of motor control (Van Beers et al., 2004; Cheng and Sabes, 2006; van Beers, 2009; Dhawale et al., 2017; van der Vliet et al., 2018). Only a separate component of motor variability, namely, planning or state noise (Cheng and Sabes, 2006; van Beers, 2009), may be subject to regulation by the central nervous system (Wu et al., 2014; Dhawale et al., 2017, 2019). Total system noise always comprises both execution and planning noise, and, thus, cannot be regulated to a level lower than that of the execution noise itself. In fact, work on limb motor control has estimated the planning noise to be substantially smaller than the execution noise, the former accounting for only about 20∼30% of total motor variability (Cheng and Sabes, 2007; van Beers, 2009; van der Vliet et al., 2018).

However, the argument that the central nervous system does control speech movements in such a way that total system noise is minimized is not compatible with our results from the Attenuated condition. There would be no reason to implement adjustments in the direction of *more* variability in this condition if the controller seeks to minimize total system noise (given that the Attenuated feedback signal indicates a variability level that is minimized even below the presumed lower bound in typical speech). Consequently, our results from the two conditions taken together support the aforementioned interpretation that, at least for speech articulation, a certain non-minimal level of

feedback variability is desirable and actively maintained, possibly in function of providing sensorimotor exploration (Wu et al., 2014; Dhawale et al., 2017, 2019). Moreover, this conclusion implies that the speech motor control system not only calculates and keeps track of distribution features for key aspects of the auditory feedback signal (e.g., dispersion measures such as variance of the formant frequencies), but also compares these features with the expected distributions and then updates future movement planning accordingly (Parrell and Houde, 2019). If our findings are replicated in future studies, computational and conceptual models of speech motor control will need to start incorporating such more complex feedback mechanisms, analogous to suggestions that have been made in the non-speech motor control literature (e.g., Herzfeld et al., 2014; Dhawale et al., 2019).

Temporal Structure

If articulatory adjustments in the Attenuated condition of the *Variability* task relied on error-based learning mechanisms similar to those driving auditory-motor adaptation with predictable formant perturbations (Houde and Jordan, 1998; Daliri and Dittman, 2019), then the temporal structure of adjustments across trials—such as indexed by the lag 1 autocorrelation [ACF(1)] of the overall sequence of productions—would be expected to vary depending on the feedback manipulation (van Beers, 2009). It should be noted at this time that the authors of one previous publication on variability in formant production suggested that their ACF(1) results indicated trial-to-trial adjustments even for speech produced without any auditory perturbation (Sitek et al., 2013). However, the lag 1 autocorrelation of –0.47 in that study was calculated based on *differences* between pairs of successive trials, thus introducing the problem of overdifferencing that we have discussed above in the Introduction (recall that after differencing even a white noise time series has a lag 1 autocorrelation of –0.5). With regard to the specific perturbation-related questions investigated in the present study, our results (illustrated in **Figure 7**) showed no statistically significant difference in ACF(1) for the sequences of trials produced in the conditions with Attenuated or Magnified formant feedback variability versus the Control condition with unaltered auditory feedback.

The lack of significant difference in ACF(1) between the Control and experimental conditions (Attenuated and Magnified) is not consistent with work by van Beers (2009). In the latter study, comparisons with a Control condition showed that ACF(1) decreased in a Magnified condition and increased in an Attenuated condition, in keeping with the prediction of a state-space model of adaptive learning based on sensory feedback (Cheng and Sabes, 2006). In fact, in our own study, most of the ACF(1) values for the sequences of productions fell within the 95% confidence interval of a white noise process, suggesting no feedback-based learning. One possible interpretation is of course that the speech control system simply does not modify productions based on auditory feedback from the immediately preceding trial. Although this control system clearly shows adaptation to predictable auditory perturbations (Houde and Jordan, 1998; Villacorta et al., 2007; Shiller et al., 2020), it is possible that such learning mechanisms are inactive in the absence of consistently maintained predictable perturbations (cf. Gonzalez Castro et al., 2014; Herzfeld et al., 2014). However, the statistically significant increase in formant production variability in the Attenuated condition does indicate a previously undocumented form of adaptive learning process during this *Variability* task.

We therefore speculate that the employed ACF(1) analysis may fail to capture the specific form of feedback-based learning in the *Variability* task. Several observations support this hypothesis. First, the state-space model of motor control predicts that when the parameter of error sensitivity (also known as adaptation rate) is very low, the ACF(1) of the trial sequence for each of the feedback manipulations implemented in the current experiment would be small, and the trial sequence would resemble a white noise process (van Beers, 2009; van der Vliet et al., 2018). It has been estimated recently that, in comparison with limb motor control studies which generally reported error sensitivity in the range of 30–50% (Baddeley et al., 2003; Cheng and Sabes, 2007; van Beers, 2009; van der Kooij et al., 2015), the error sensitivity for speech auditory-motor adaptation is, on average, as small as 4.8% (Daliri and Dittman, 2019). Second, it is known from previous studies that adaptive learning in speech production can differ between different vowels and words, and a given participant may even adapt for one vowel but follow the perturbation for another vowel (Houde and Jordan, 1998; Max and Maffett, 2015). In the current study's *Variability* task, three different target words ("talk," "tech," "tuck") were produced in pseudo-random order. Feedback-based learning under such circumstances may be very complex (e.g., How much does feedback from a trial of "tech" affect the production of "talk"? What is the influence of some trials being preceded by the same word and other trials by a different word?), especially if the history of feedback prior to the last trial is also taken into account (Herzfeld et al., 2014). Such complexity is not captured by the simple ACF(1) index. Third, the statistically significant change in formant production variability during the Attenuated condition of the *Variability* task indicates that the production sequence may be non-stationary, which renders ACF(1) difficult to interpret. Unfortunately, despite these various disadvantages of ACF(1), it is unclear which alternative measurements may be used to reveal the temporal structure of feedback-based adaptive learning in conditions with altered formant feedback variability.

Effect of Variability on Adaptation

Immediately after having been exposed to Attenuated or Magnified formant feedback variability, participants completed a conventional speech auditory-motor adaptation task with a predictable upward shift of all formants. This *Adaptation* task allowed us to assess the potential effect of prior formant feedback variability on formant production learning. If sensorimotor learning is affected by the extent of perceived inter-trial variability (Herzfeld et al., 2014; Wu et al., 2014), then participants' formant adaptation profiles can be expected to differ after experiencing Attenuated versus Magnified formant feedback variability. On the other hand, if inter-trial variability has no effect on the mechanisms underlying adaptive

learning (van der Kooij et al., 2015), then participants' formant adaptation profiles can be expected to remain unchanged between conditions.

Results shown in **Figures 8**, **9** indicate that three different measures of formant adaptation—early adaptation extent, early adaptation rate, and final adaptation extent—were all statistically indistinguishable between the Control condition and the two experimental conditions (Magnified or Attenuated feedback variability). In other words, the prior manipulation of formant feedback variability, or the participants' motor adjustments to this manipulation, had no effect at all on the subsequent formant shift adaptation task. This result aligns with the conclusion of van der Kooij et al. (2015) who conducted a series of visuomotor rotation reach experiments with magnified or attenuated visual feedback errors. Although those authors observed behavioral differences across the feedback manipulation conditions, state-space model estimates of the underlying learning *mechanism* remained unchanged. In the current study, even the behavioral measures showed no differences at all in each group's comparison of formant adaptation after the control versus experimental condition of the *Variability* task. Hence, our results for speech articulation suggest no direct relationship between formant variability perceived in a preceding task and the adaptive learning of formant output adjustments when subsequently exposed to a persistent formant perturbation.

The absence of an effect of formant feedback variability on formant production adaptation may relate to the aforementioned low error-sensitivity parameter in speech auditory-motor adaptation (Daliri and Dittman, 2019). Of course, it is also possible that this outcome is entirely specific to certain methodological aspects of our study. For example, we only implemented a relatively *short-term* feedback variability manipulation (75 trials), and examined formant-shift adaptation in a *subsequent* task. Future studies should also address the effect of longer-term variability manipulations and variability manipulations implemented during the auditory-motor adaptation task itself. Moreover, it might prove fruitful to develop methodological approaches that are able to dissociate the effects of manipulations that alter sensory variability (as implemented here) versus direct manipulations of motor variability (which alter both motor and sensory variability).

CONCLUSION

In sum, by experimentally manipulating inter-trial formant variability in the auditory feedback signal for speech, the present study yielded three novel findings. First, formant production variability in speech production appears to be actively regulated to a desirable level rather than merely minimized. Second, under the conditions investigated here, the temporal structure of inter-trial formant changes was not affected by experimental manipulations of formant feedback variability. Third, for these specific test conditions, subsequent auditory-motor adaptation in a standard formant shift perturbation task was also not affected by the formant feedback manipulations. We hope that future empirical studies will be able to investigate the generalizability of these findings, and that future theoretical work will provide conceptual and computational accounts of the active regulation of inter-trial variability in the sensorimotor control of speech production.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Washington IRB. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

HW collected and analyzed the data. Both authors designed the experiments and data analysis procedures, interpreted the data, wrote the manuscript, contributed to the article, and approved the submitted version.

FUNDING

This research was supported by grants R01DC014510 and R01DC017444 from the National Institute on Deafness and Other Communication Disorders.

ACKNOWLEDGMENTS

We thank Kwang S. Kim, Ph.D., for contributions during the process of designing and implementing the auditory feedback perturbations.

REFERENCES

Baddeley, R. J., Ingram, H. A., and Miall, R. C. (2003). System identification applied to a visuomotor task: near-optimal human performance in a noisy changing task. *J. Neurosci.* 23, 3066–3075. doi: 10.1523/jneurosci.23-07-03066.2003

Bieniek, M. M., Bennett, P. J., Sekuler, A. B., and Rousselet, G. A. (2016). A robust and representative lower bound on object processing speed in humans. *Eur. J. Neurosci.* 44, 1804–1814. doi: 10.1111/ejn.13100

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot Int.* 5, 341–345.

Brockwell, P. J., and Davis, R. A. (2016). *Introduction to Time Series and Forecasting*. Berlin: Springer.

Cai, S., Boucek, M., Ghosh, S., Guenther, F., and Perkell, J. (2008). "A system for online dynamic perturbation of formant trajectories and results from perturbations of the mandarin triphthong /iau/,"in *Proceedings of the 8th ISSP*, University of Strasbourg, Strasbourg.

Calvin, W. H., and Stevens, C. F. (1968). Synaptic noise and other sources of randomness in motoneuron interspike intervals. *J. Neurophysiol.* 31, 574–587. doi: 10.1152/jn.1968.31.4.574

Chao, S.-C., Ochoa, D., and Daliri, A. (2019). Production variability and categorical perception of vowels are strongly linked. *Front. Hum. Neurosci.* 13:96. doi: 10.3389/fnhum.2019.00096

Cheng, S., and Sabes, P. N. (2006). Modeling sensorimotor learning with linear dynamical systems. *Neural Comput.* 18, 760–793. doi: 10.1162/089976606775774651

Cheng, S., and Sabes, P. N. (2007). Calibration of visually guided reaching is driven by error-corrective learning and internal dynamics. *J. Neurophysiol.* 97, 3057–3069. doi: 10.1152/jn.00897.2006

Clamann, H. P. (1969). Statistical analysis of motor unit firing patterns in a human skeletal muscle. *Biophys. J.* 9, 1233–1251. doi: 10.1016/S0006-3495(69)86448-9

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Routledge Academic.

Cornelisse, L. E., Gagne, J. P., and Seewald, R. C. (1991). Ear level recordings of the long-term average spectrum of speech. *Ear Hear.* 12, 47–54. doi: 10.1097/00003446-199102000-00006

Cryer, J. D., and Chan, K.-S. (2008). *Time Series Analysis: With Applications to R*. Berlin: Springer. doi: 10.1007/978-0-387-75959-3

Daliri, A., and Dittman, J. (2019). Successful auditory motor adaptation requires task-relevant auditory errors. *J. Neurophysiol.* 122, 552–562. doi: 10.1152/jn.00662.2018

Dhawale, A. K., Miyamoto, Y. R., Smith, M. A., and Ölveczky, B. P. (2019). Adaptive regulation of motor variability. *Curr. Biol.* 29, 3551.e7–3562.e7. doi: 10.1016/j.cub.2019.08.052

Dhawale, A. K., Smith, M. A., and Ölveczky, B. P. (2017). The role of variability in motor learning. *Ann. Rev. Neurosci.* 40, 479–498. doi: 10.1146/annurev-neuro-072116-031548

Franken, M. K., Acheson, D. J., McQueen, J. M., Eisner, F., and Hagoort, P. (2017). Individual variability as a window on production-perception interactions in speech motor control. *J. Acoust. Soc. Am.* 142, 2007–2018. doi: 10.1121/1.5006899

Gonzalez Castro, L. N., Hadjiosif, A. M., Hemphill, M. A., and Smith, M. A. (2014). Environmental consistency determines the rate of motor adaptation. *Curr. Biol.* 24, 1050–1061. doi: 10.1016/j.cub.2014.03.049

Hamilton, A. F. D. C., Jones, K. E., and Wolpert, D. M. (2004). The scaling of motor noise with muscle strength and motor unit number in humans. *Exp. Brain Res.* 157, 417–430. doi: 10.1007/s00221-004-1856-7

He, K., Liang, Y., Abdollahi, F., Fisher Bittmann, M., Kording, K., and Wei, K. (2016). The statistical determinants of the speed of motor learning. *PLoS Comput. Biol.* 12:e1005023. doi: 10.1371/journal.pcbi.1005023

Herzfeld, D. J., Vaswani, P. A., Marko, M. K., and Shadmehr, R. (2014). A memory of errors in sensorimotor learning. *Science* 345, 1349–1353. doi: 10.1126/science.1253138

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6, 65–70.

Houde, J. F., and Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science* 279, 1213–1216. doi: 10.1126/science.279.5354.1213

Kim, K. S., Daliri, A., Flanagan, J. R., and Max, L. (2020a). Dissociated development of speech and limb sensorimotor learning in stuttering: speech auditory-motor learning is impaired in both children and adults who stutter. *Neuroscience* 451, 1–21. doi: 10.1016/j.neuroscience.2020.10.014

Kim, K. S., Wang, H., and Max, L. (2020b). It's about time: minimizing hardware and software latencies in speech research with real-time auditory feedback. *J. Speech Lang. Hear. Res.* 63, 2522–2534. doi: 10.1044/2020_JSLHR-19-00419

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. *Speech Prod. Speech Model.* 55, 403–439. doi: 10.1007/978-94-009-2037-8_16

MacDonald, E. N., Purcell, D. W., and Munhall, K. G. (2011). Probing the independence of formant control using altered auditory feedback. *J. Acoust. Soc. Am.* 129:955. doi: 10.1121/1.3531932

MacNeilage, P. F. (1970). Motor control of serial ordering of speech. *Psychol. Rev.* 77, 182–196. doi: 10.1037/h0029070

Max, L., and Maffett, D. G. (2015). Feedback delays eliminate auditory-motor learning in speech production. *Neurosci. Lett.* 591, 25–29.

Murillo, D. B., Sánchez, C. C., Moreside, J., Vera-García, F. J., and Moreno, F. J. (2017). Can the structure of motor variability predict learning rate? *J. Exp. Psychol.* 43, 596–607. doi: 10.1037/xhp0000303

Nault, D. R., and Munhall, K. G. (2020). Individual variability in auditory feedback processing: responses to real-time formant perturbations and their relation to perceptual acuity. *J. Acoust. Soc. Am.* 148:3709. doi: 10.1121/10.0002923

Niziolek, C. A., and Kiran, S. (2018). Assessing speech correction abilities with acoustic analyses: evidence of preserved online correction in persons with aphasia. *Int. J. Speech Lang. Pathol.* 20, 659–668. doi: 10.1080/17549507.2018.1498920

Niziolek, C. A., Nagarajan, S. S., and Houde, J. F. (2013). What does motor efference copy represent? evidence from speech production. *J. Neurosci.* 33, 16110–16116. doi: 10.1523/JNEUROSCI.2137-13.2013

Parrell, B., and Houde, J. (2019). Modeling the role of sensory feedback in speech motor control and learning. *J. Speech Lang. Hear. Res.* 62, 2963–2985. doi: 10.1044/2019_JSLHR-S-CSMC7-18-0127

Patri, J. F., Diard, J., and Perrier, P. (2015). Optimal speech motor control and token-to-token variability: a Bayesian modeling approach. *Biol. Cybernet.* 109, 611–626. doi: 10.1007/s00422-015-0664-4

Patton, J. L., Wei, Y. J., Bajaj, P., and Scheidt, R. A. (2013). Visuomotor learning enhanced by augmenting instantaneous trajectory error feedback during reaching. *PLoS One* 8:e46466. doi: 10.1371/journal.pone.0046466

Perkell, J. S., and Klatt, D. H. (1986). *Invariance and Variability in Speech Processes*. Mahwah, NJ: Lawrence Erlbaum Associates.

Perkell, J. S., Lane, H., Ghosh, S., Matthies, M. L., Tiede, M., Guenther, F., et al. (2008). "Mechanisms of vowel production: auditory goals and speaker acuity," in *Proceeding of the Paper Presented at the 8th International Seminar on Speech Production*, Groningen.

Purcell, D. W., and Munhall, K. G. (2006). Adaptive control of vowel formant frequency: evidence from real-time formant manipulation. *J. Acoust. Soc. Am.* 120, 966–977. doi: 10.1121/1.2217714

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Rousselet, G. A., Pernet, C. R., and Wilcox, R. R. (2017). Beyond differences in means: robust graphical methods to compare two groups in neuroscience. *Eur. J. Neurosci.* 46, 1738–1748. doi: 10.1111/ejn.13610

Scheerer, N. E., and Jones, J. A. (2012). The relationship between vocal accuracy and variability to the level of compensation to altered auditory feedback. *Neurosci. Lett.* 529, 128–132. doi: 10.1016/j.neulet.2012.09.012

Shiller, D. M., Mitsuya, T., and Max, L. (2020). Exposure to auditory feedback delay while speaking induces perceptual habituation but does not mitigate the disruptive effect of delay on speech auditory-motor learning. *Neuroscience* 446, 213–224. doi: 10.1016/j.neuroscience.2020.07.041

Singh, P., Jana, S., Ghosal, A., Murthy, A., and Goldberg, M. E. (2016). Exploration of joint redundancy but not task space variability facilitates supervised motor learning. *Proc. Natl. Acad. Sci. U.S.A.* 113, 14414–14419. doi: 10.1073/pnas.1613383113

Sitek, K. R., Mathalon, D. H., Roach, B. J., Houde, J. F., Niziolek, C. A., and Ford, J. M. (2013). Auditory cortex processes variation in our own speech. *PLoS One* 8:e82925. doi: 10.1371/journal.pone.0082925

Sternad, D. (2018). It's not (only) the mean that matters: variability, noise and exploration in skill learning. *Curr. Opin. Behav. Sci.* 20, 183–195. doi: 10.1016/j.cobeha.2018.01.004

Tang, D., Parrell, B., and Niziolek, C. A. (2021). Variability is actively regulated in speech. *bioRxiv* [Preprint]. doi: 10.1101/2021.10.08.462639

Tourville, J. A., Cai, S., and Guenther, F. (2013). Exploring auditory-motor interactions in normal and disordered speech. *Proc. Meet. Acoust.* 19:060180. doi: 10.1121/1.4800684

van Beers, R. J. (2009). Motor learning is optimally tuned to the properties of motor noise. *Neuron* 63, 406–417. doi: 10.1016/j.neuron.2009.06.025

Van Beers, R. J., Haggard, P., and Wolpert, D. M. (2004). The role of execution noise in movement variability. *J. Neurophysiol.* 91, 1050–1063. doi: 10.1152/jn.00652.2003

van der Kooij, K., Brenner, E., van Beers, R. J., and Smeets, J. B. J. (2015). Visuomotor adaptation: how forgetting keeps us conservative. *PLoS One* 10:e0117901. doi: 10.1371/journal.pone.0117901

van der Vliet, R., Frens, M. A., de Vreede, L., Jonker, Z. D., Ribbers, G. M., Selles, R. W., et al. (2018). Individual differences in motor noise and adaptation rate are optimally related. *Eneuro* 5: ENEURO.0170-18.2018. doi: 10.1523/ENEURO.0170-18.2018

Villacorta, V. M., Perkell, J. S., and Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *J. Acoust. Soc. Am.* 122, 2306–2319. doi: 10.1121/1.2773966

Wilcox, R. R., and Erceg-Hurn, D. M. (2012). Comparing two dependent groups via quantiles. *J. Appl. Stat.* 39, 2655–2664. doi: 10.1080/02664763.2012.724665

Wong, J., Wilson, E. T., Malfait, N., and Gribble, P. L. (2009). The influence of visual perturbations on the neural control of limb stiffness. *J. Neurophysiol.* 101, 246–257. doi: 10.1152/jn.90371.2008

Wu, H. G., Miyamoto, Y. R., Castro, L. N. G., Ölveczky, B. P., and Smith, M. A. (2014). Temporal structure of motor variability is dynamically regulated and predicts motor learning ability. *Nat. Neurosci.* 17, 312–321. doi: 10.1038/nn.3616

Author Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Deafness and Other Communication Disorders or the National Institutes of Health.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Wang and Max. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Human Neuroscience

TYPE Original Research
PUBLISHED 15 July 2022
DOI 10.3389/fnhum.2022.879981

Check for updates

OPEN ACCESS

EDITED BY
Jeffery A. Jones,
Wilfrid Laurier University, Canada

REVIEWED BY
Rory A. DePaolis,
James Madison University, United States
Takemi Mochida,
Nippon Telegraph and Telephone, Japan

*CORRESPONDENCE
Mark K. Tiede
tiede@haskins.yale.edu

SPECIALTY SECTION
This article was submitted to Speech and Language, a section of the journal Frontiers in Human Neuroscience

RECEIVED 20 February 2022
ACCEPTED 28 June 2022
PUBLISHED 15 July 2022

CITATION
Goldenberg D, Tiede MK, Bennett RT and Whalen DH (2022) Congruent aero-tactile stimuli bias perception of voicing continua.
Front. Hum. Neurosci. 16:879981.
doi: 10.3389/fnhum.2022.879981

COPYRIGHT
© 2022 Goldenberg, Tiede, Bennett and Whalen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Congruent aero-tactile stimuli bias perception of voicing continua

Dolly Goldenberg[1], Mark K. Tiede[1]*, Ryan T. Bennett[2] and D. H. Whalen[1,3,4]

[1]Haskins Laboratories, New Haven, CT, United States, [2]Department of Linguistics, University of California, Santa Cruz, Santa Cruz, CA, United States, [3]The Graduate Center, City University of New York (CUNY), New York, NY, United States, [4]Department of Linguistics, Yale University, New Haven, CT, United States

Multimodal integration is the formation of a coherent percept from different sensory inputs such as vision, audition, and somatosensation. Most research on multimodal integration in speech perception has focused on audio-visual integration. In recent years, audio-tactile integration has also been investigated, and it has been established that puffs of air applied to the skin and timed with listening tasks shift the perception of voicing by naive listeners. The current study has replicated and extended these findings by testing the effect of air puffs on gradations of voice onset time along a continuum rather than the voiced and voiceless endpoints of the original work. Three continua were tested: bilabial ("pa/ba"), velar ("ka/ga"), and a vowel continuum ("head/hid") used as a control. The presence of air puffs was found to significantly increase the likelihood of choosing voiceless responses for the two VOT continua but had no effect on choices for the vowel continuum. Analysis of response times revealed that the presence of air puffs lengthened responses for intermediate (ambiguous) stimuli and shortened them for endpoint (non-ambiguous) stimuli. The slowest response times were observed for the intermediate steps for all three continua, but for the bilabial continuum this effect interacted with the presence of air puffs: responses were slower in the presence of air puffs, and faster in their absence. This suggests that during integration auditory and aero-tactile inputs are weighted differently by the perceptual system, with the latter exerting greater influence in those cases where the auditory cues for voicing are ambiguous.

KEYWORDS

sensory integration, action-perception, multimodal speech perception, perceptual units, tactile perception

Introduction

In multisensory (or multimodal) integration, information from different sensory modalities, such as sight, audition, or somatosensation, are integrated by the human perceptual and nervous system into a coherent percept (see Rosenblum, 2008a; Stein and Stanford, 2008; Stein et al., 2009; Spence and Bayne, 2015; for review and discussion).

This integration occurs even though the input from different sensory modalities is processed at different speeds (Eagleman, 2008): for instance, auditory input reaches the cortex in less than half the time of visual input (Molholm et al., 2002). Animal studies with single neurons indicate that there are differences in the way that multimodal and unimodal signals are processed (Stein and Stanford, 2008), consistent with human use of separate, multimodal regions for some tasks (e.g., Banati et al., 2000; Calvert, 2001). Direct comparisons of neural processing speeds for haptic input are more difficult, since possible contact points on the skin are distributed over the entire body, not just the area of the eyes and ears. To complicate matters further, the speed of processing is affected by factors such as stimulus intensity (e.g., Colonius and Diederich, 2004), previous experience (Miyazaki et al., 2006), or the way stimuli are presented (Harrar and Harris, 2008), all of which can affect the salience of correspondence between different sensory inputs during the process of integration.

A relevant question is how sensations associated with different afferent timings become integrated and perceived as a single coherent event. One possibility could be a dynamic recalibration of expectations. Eagleman and Holcombe (2002) and Haggard et al. (2002) demonstrated that participants perceive two events from different modalities (haptic and visual, in this case) as being closer temporally than they are in fact because they perceive them as part of the same event: a flash of light that appeared after the participants have pressed a button was perceived as immediately subsequent to the button press even though it was objectively later than that. Stetson et al. (2006) suggested that participant expectations of the relative timing of motor acts and sensory consequences can shift, even to the extent that they can switch places: the later event can be perceived as earlier. This shows that sensory inputs, processed at different speeds but associated with the same event, can be part of one coherent percept.

Multisensory integration in speech perception is the combined use of different sensory modalities in the construction of a speech percept. Most current research on multimodal integration focuses on vision and audition: vision has been demonstrated to enhance the perception of speech when integrated with auditory stimuli in both suboptimal acoustic conditions such as background noise or strong foreign accent (Sumby and Pollack, 1954; Reisberg et al., 1987; MacLeod and Summerfield, 1990) and cases of increased cognitive load such as complicated structure or content (Reisberg et al., 1987; Arnold and Hill, 2001). Visual cues have also been demonstrated to facilitate language acquisition both in children (Mills, 1987) and adults acquiring a second language (Hardison, 2007), and to improve the speech perception of individuals with hearing impairments, especially individuals with cochlear implants (e.g., Geers and Brenner, 1994; Grant and Seitz, 2000; Lachs et al., 2001; Kaiser et al., 2003). Conversely, it has been shown that incongruent visual and auditory cues can modify perception of the acoustic signal in adults (McGurk and MacDonald, 1976; Massaro et al., 1993) and infants (Burnham and Dodd, 1996; Rosenblum et al., 1997). This body of evidence suggests that visual and auditory cues are integrated, along with other cues, in the process of speech perception (Rosenblum et al., 2017).

In recent years, evidence has accumulated demonstrating that tactile information may also be integrated with other modalities in general (e.g., Banati et al., 2000; Lee et al., 2019), and in the perception of speech in particular. In early studies, the effects of tactile information on perception was shown for participants that either had explicit knowledge of the task (Fowler and Dekle, 1991; Gick et al., 2008), or were trained to make a connection between the tactile and the auditory cues (Sparks et al., 1978; Reed et al., 1989; Bernstein et al., 1991). However, later studies have established that tactile information influences auditory perception of uninformed and untrained listeners as well (Gick and Derrick, 2009; Ito et al., 2009; Derrick and Gick, 2013).

Ito et al. (2009) used a robotic device to pull facial skin, creating patterns of facial skin deformation in listeners, that normally accompany the production of the vowels /ɛ/ and /æ/. They showed that by timing these deformations to auditory stimuli, the perceptual judgments of a synthetic vowel continuum ranging from /ɛ/to/æ/ were shifted in the expected direction of the bias. For example, when the skin was pulled upward (a deformation consistent with /ɛ/) the word "head" was preferred, whereas when the skin was pulled downward (consistent with /æ/) the word "had" was preferred. However, deformations applied rearward (orthogonal to directions consistent with vowel production) had no effect on the perceptual judgments. Ito et al. concluded that somatosensory cues can modulate speech perception, but only when these are congruent with those expected in production.

Gick and Derrick (2009) studied the effect of applying air puffs to the back of the hand and the center of the neck at the suprasternal notch on auditory perception of a voicing contrast. In their experiment, native speakers of North-American English were asked to determine whether they heard a syllable with an initial voiceless stop or a syllable with an initial voiced stop. The stimuli, the syllables /ba/, /pa/, /da/, and /ta/ produced by a male native speaker of North-American English, were partially masked by white noise in order to increase ambiguity. During some trials, while the participants heard the stimuli, puffs of air were applied to the back of the participant's hand, on their suprasternal notch, or as a control beside and tangent to headphones they wore. During the control trials the puff had no direct contact with hair or skin, and was released only into the air near the headphones. The participants were blindfolded; thus, they had no visual information about the application of the air puffs. The duration of the air puffs reflected the duration of the turbulent part of a naturally produced English aspirated

consonant. The presence of airflow facilitated the identification of voiceless stops and reduced the identification of voiced stops. Since no such effect was found for the participants in the control group where no direct tactile information was provided, Gick and Derrick concluded that tactile information can modulate speech perception similar to the way vision does.

In a later study, the effect of tactile stimulation of the ankle on auditory perception was tested (Derrick and Gick, 2013). The motivation for using the ankle was two-fold. First, it is a distal location relative to the source of aspiration in typical speaking situations. Thus, while speakers may have experience with feeling air puffs on the back of their hand while they were speaking, or, at least to some extent, with feeling air puffs on the neck while others were speaking, it is unlikely they have similar experience with feeling air puffs on their ankles. Moreover, even if such experience does exist, it is not frequent or robust, thus it is not likely that participants associate the feeling or a puff of air on their ankle with the production of certain speech sounds. Second, the ankle is distant from the ear, and its representation in the somatosensory cortex is distant from the ear's representation in the somatosensory cortex (Penfield and Rasmussen, 1950). Since comparison of the ankle results to the hand and neck results from Gick and Derrick (2009) did not reveal significant differences, Derrick and Gick concluded that integration is a full-body process and that the association between the felt puff of air and the produced aspirated sound does not depend on direct experience.

Evidence for multimodal speech perception addresses the debate over the nature of the objects of speech perception. From a general auditory point of view (e.g., Klatt, 1979; Stevens, 1981, 1989; Massaro, 1987; Diehl et al., 2004; Hickok and Poeppel, 2007; Yi et al., 2019) the objects of speech perception are sounds. From an ecological or direct perception point of view, represented in the field of speech by Direct Realism (e.g., Fowler, 1981, 1984, 1996), these objects are physical events in the actual world–vocal tract gestures. From the point of view of Motor Theory (Liberman et al., 1967; Liberman and Mattingly, 1985; Liberman and Whalen, 2000) and Articulatory Phonology (Browman and Goldstein, 1986, 1989, 1992; Galantucci et al., 2009) the objects of speech perception are abstract representations of vocal tract gestures rather than physical events as such. The general auditory approaches assume that perception of speech sounds is the same as perception of non-speech sounds. According to this view, the same mechanisms of audition and perceptual learning are used for perception of all types of sounds. Thus, from this perspective, the objects of speech perception may be acoustic or auditory objects, or acoustic landmarks which convey information about the gestures that produced them (Stevens, 2002; Yun et al., 2020). These approaches posit an intermediate representation constructed from sensory input. That is, listeners identify acoustic patterns or features by matching them to stored acoustic representations. In contrast to the non-auditory approaches which assume listeners recover gestures in some form, according to the auditory view listeners perceive "the acoustic consequences of gestures" (Diehl et al., 2004, p. 168) (though see Stevens, 2002). It is assumed that all the relevant information for perception of speech is included in the acoustic signal and is recoverable by general mechanisms of perceptual learning.

But an argument in favor of the non-auditory approaches arises from evidence for multisensory integration, which suggests that the objects of speech perception are distinct from units of non-speech auditory perception [see Goldstein and Fowler (2003) and Rosenblum (2008b) for examples and discussion]. The argument is that if visual or other sensory cues participate in the process of speech perception, the objects of speech perception cannot be auditory, or at least not exclusively auditory, and evidence supporting integration from multiple modalities serves to strengthen this position. However, this argument relies on the interpretation of these experimental findings as supporting multimodal integration in speech perception. For the air puff studies of Gick and Derrick (2009) and Massaro (2009), Derrick and Gick (2013) has argued that it is possible that the participants interpreted the airflow, when it was provided, as aspiration and relied on this interpretation in making their decision. That is, the criticism is that the participants may have based their responses *only* on tactile information without any integration with the auditory cues. The possibility that Gick and Derrick's findings were simply the result of a general response to tactile stimuli was tested in Gick and Derrick (2009). A tap condition, in which contact with the same test locations was made using a metal solenoid plunger, established that while aero-tactile stimuli were able to shift speech perception, taps on the skin of the participants did not (see supplementary material, Gick and Derrick, 2009). Derrick and Gick (2013) argue that the results of this test are not just a control for a general attention effect caused by the addition of another type of stimuli, but also suggest that the integration of the tactile signal with the auditory signal is dependent upon it being perceived as "event-relevant, as opposed to merely synchronous" (Derrick and Gick, 2013, p. 406).

However, this test does not rule out Massaro's suggestion that there was no integration, since it is still possible that speech perception during the experiment was unimodal, that is, based solely on aero-tactile information when it was provided, and on auditory information when aero-tactile information was not provided. The stimuli in Gick and Derrick (2009) and Derrick and Gick (2013) were masked by background noise. This made the acoustic stimuli less informative than they could have been under perfect acoustic conditions. Therefore, it might have been the case that the tactile stimulus was the most prominent signal, and as a result a unimodal response was

made to it. The current study aims at investigating this question further. Specifically, we use voice onset time (VOT) continua systematically ranging over eight steps from voiceless to voiced sounds rather than endpoint stimuli only (as in the work by Gick and Derrick). This design enables us to show that biasing effects of air puffs are least at the endpoints and greatest for the ambiguous stimuli near the perceptual boundary, supporting interpretation of the tactile cues as forming an integrated rather than unimodal response.

Our prediction is that if integration is not part of the process, then all the sounds along the continuum should be equally affected by aero-tactile cues, and so trials accompanied by air puffs will be perceived unimodally as being voiceless. However, if instead the results show an interaction between the effect of air puff and the effect of step along the continuum this would suggest that aero-tactile information is taken into account along with the auditory information provided, in cases when auditory information is not sufficient for disambiguation, or when the tactile information is not congruent with the auditory information. Such a result would show that participants are using a context-weighted blend of sensory cues in perceiving and categorizing speech sounds, thus providing an example of multi-sensory integration in the perception of speech.

As an additional test for saliency of tactile cues, a continuum consisting of vowel sounds ranging from /ɛ/ to /ɪ/ in a/hVd/ context was included as a control. While higher vowels are produced with a more constricted oral passage (Jaeger, 1978), both endpoints have approximately equal airflow and are thus not expected to be sensitive to aero-tactile cues. A contrast between an effect of air-puffs on perception of the VOT continua and a lack of it for the vowel continuum would further support an interpretation that cues are integrated only when relevant, that is, that the aero-tactile information is taken into account only in cases where aspiration (or amount of air produced by the speaker) is relevant for the distinction being made.

Materials and methods

Participants

In a survey, 42 monolingual native speakers of American English participated in the experiment (24 females; age range 18–56, mean age 28.7, SD = 11.5). Only right-hand dominant participants were recruited. The participants were all residents of Southern Connecticut at the time of the experiment. Their level of education ranged from high school graduates to graduate students. The participants were recruited with flyers and by word of mouth. All were naive to the purpose of the study and had no self-reported speech or hearing defects. All participants provided informed consent overseen by the Yale Human Research Protection Program and were compensated for their time.

Stimuli

Acoustic stimuli

Voice onset time is the interval between the release of a stop consonant and the onset of voicing following or preceding the release (Lisker and Abramson, 1964; Abramson and Whalen, 2017). In American English stops are habitually produced with a positive average VOT. The duration of the positive VOT is longer for voiceless stops than for voiced stops and varies with place of articulation: the more distant the place of articulation from the lips, the longer the VOT. Average VOT durations for American English stops are summarized in **Table 1**. Note that VOT varies with context: it is shorter for stops when following an obstruent than when following a nasal, a glide, or a vowel. For stops in onset positions it is shortest for those in clusters that begin with /s/ (Randolph, 1989).

Our endpoint stimuli were taken from a recording of a male monolingual native speaker of American English. He produced six tokens of each of the syllables /pa/,/ba/,/ka/, and /ga/. These were used to obtain his average values for VOT for these utterances. Two eight-step VOT continua were then created, one for the bilabial and one for the velar place of articulation. The continua were created by removing the initial burst from one of the voiceless exemplars (/pa/or/ka/) and then systematically shortening the aspiration in log-scaled steps, with the final step matching the mean aspiration duration of the voiced token. Aspiration durations for each step of the VOT continua appear in **Table 2**. A non-linear (logarithmic) step size was chosen because psycho-acoustic perception tends to follow Weber's law (subjective sensation is proportional to the logarithm of the stimulus intensity); e.g., Fastl and Zwicker (2006). See Rosen and Howell (1981) for results on VOT, and Stevens (2000, p. 228) for a similar effect on the perception of duration of burst.

An additional continuum consisting of vowel sounds ranging from /ɛ/ to /ɪ/ in an /hVd/context was included for use as a control. It was synthesized from endpoint recordings of a male monolingual native speaker of North-American English producing "head" and "hid," by linearly interpolating F1 and F2 values within the vowel over the eight continuum steps, using an

TABLE 1 Average VOT durations for American English stops (Byrd, 1993).

Place of articulation	VOT length (ms)	
	Voiceless	Voiced
Bilabial	44	18
Alveolar	49	24
Velar	52	27

TABLE 2 VOT continua steps showing length of retained aspiration at each step (ms).

Step no.	VOT length (ms)	
	Bilabial continuum	Velar continuum
1	98	81
2	58	56
3	37	42
4	24	35
5	18	31
6	14	28
7	12	27
8	11	26

iterative Burg algorithm to shift the location of filter poles and zeros in resynthesis (Purcell and Munhall, 2006).[1]

A pre-test of each continuum conducted online as a Mechanical Turk task was used to assess the quality of the stimuli. The test was run with an independent group of participants that did not take part in the main study (N = 41). They were asked to choose whether they heard a "pa" or a "ba" (in the bilabial condition), or "ka" or "ga" (in the velar condition), and rate the goodness of the token on a five step Likert scale. The sounds from the two continua (/pa/-/ba/ and /ka/-/ga/) were presented in the same test. A similar pre-test was conducted for the vowel continuum in which additional 20 participants were asked to choose whether they heard "head" or "hid," and to rate the goodness of the token. The order of presentation was randomized in both pre-tests. The results of the pretests are plotted in **Figure 1**.

1 The 24 sound files used as acoustic stimuli are available as Supplementary Material from https://tinyurl.com/2p8tjfnh.

The bilabial category boundary is approximately centered between its endpoints, that is, its bias (4.2) is close to its midpoint (4.5). The bias was calculated as the 50% crossover point of the psychometric function for the continuum, computed across all listeners. Acuity (a measure of boundary slope) was computed as the difference between the 25 and 75% probabilities for the discrimination function. The velar category boundary is not as centralized and is skewed toward voicelessness (bias = 3.6); that is, longer VOTs were necessary for /ka/ responses. The velar acuity (2.0) is shallower than that of the bilabial (1.1), possibly due to this skew. Finally, the category boundary for the vowel control continuum is also approximately centered (bias = 4.7, acuity = 1.5). The goodness ratings for all three continua are higher at the margins than at the intermediate steps of the continuum, which reflects the fact that the ambiguous sounds were harder to categorize, as expected.

Tactile (air puff) stimuli

To deliver air puff stimuli the following equipment was employed. A three gallon air compressor (Campbell Hausfeld) was connected to a solenoid valve (Parker) used to gate airflow by 1/4-inch polyethylene tubing. The solenoid was toggled by a programmable relay controller device (KMtronic). A pressure transducer (PSC, model 312) and a flow meter (Porter-Parker MPC series) were connected to the tubing in order to monitor pressure and flow data. Solenoid control of airflow, presentation of audio stimuli and data recording were performed using a custom Matlab (Mathworks) procedure that was written for this experiment. The tubing was inserted into a soundproof room through a cable port and stabilized using a table microphone stand (see **Figure 2** for a diagram of the system).

In a given trial the signal to open the air valve solenoid was given by the Matlab procedure, which also controlled acoustic stimulus presentation through the computer's sound card such that the acoustic onset of each of the stimulus was coincident with the onset of the air puff from the tube. Detectable air

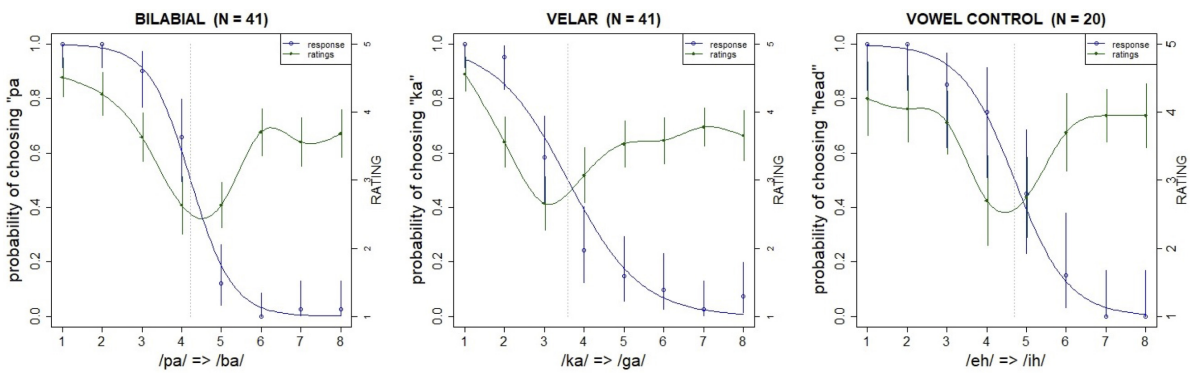


FIGURE 1
Viability test results for the continua: left scale (blue line) shows probability of choosing voiceless ("pa" or "ka") or "head" relative to step (dotted vertical line marks 50% crossover point); right scale (green line) shows Likert scale ratings by step. Error bars show 95% confidence intervals.

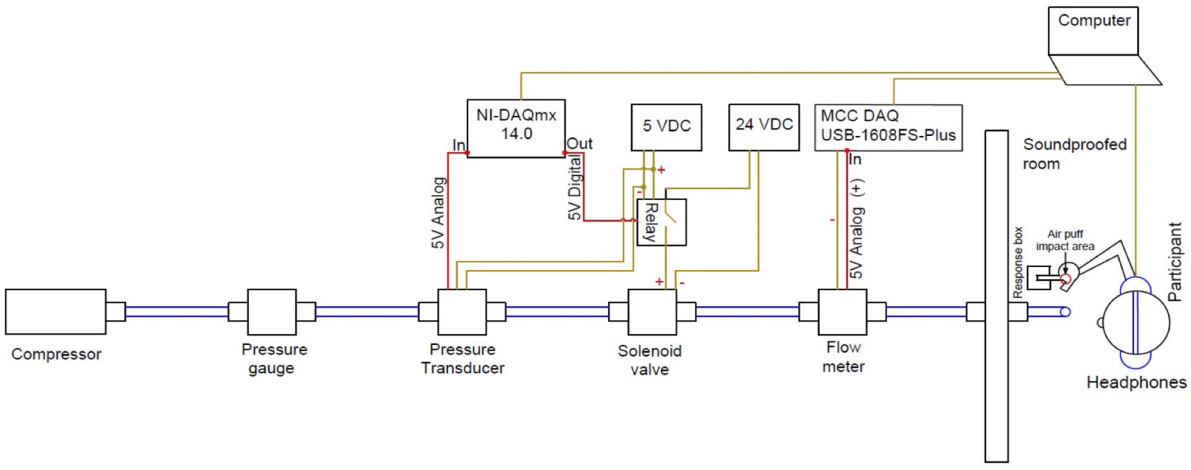


FIGURE 2
The aero-tactile stimulus presentation system.

turbulence exiting the tube was 87 ms in duration for the bilabial condition and 92 ms in duration for the velar condition. These timings reflect the mean aspiration time (that is, VOT) of the six voiceless tokens that the model speaker produced, thus simulating the temporal properties of the stimuli. The speaker's mean VOTs fall within the VOT range of initial aspirated stops in American English (54–100 ms, Lisker and Abramson, 1967; Cooper, 1991; Byrd, 1993). The airflow at the exit point of the tube was 5 Standard Liter Per Minute (SLPM). Note that this rate is lower than the average airflow of typical speech (8 SLPM, Isshiki and von Leden, 1964), and significantly lower than the average airflow of voiceless stop consonants in CV syllables (about 56 SLPM, Isshiki and Ringel, 1964). A lower rate was used to better align with the reduction in speed that occurs once aspiration exits the mouth, and additionally to reduce the possibility that the puff would be audible. The exit point of the tube was placed 5 cm away from the participant's skin, creating an area of initial impact with a diameter of 2–3 cm [similar to Derrick et al. (2009)]. The air puffs were applied on the dorsal surface of the right hand between the thumb and forefinger (see **Figure 3A**). A microphone placed near the exit of the tube was used to record airflow turbulence during each trial, to verify that air puff stimuli (when scheduled) were delivered with the expected timing.

Procedure

Each experimental session included two parts, an initial test to verify that the air puffs were felt but not heard, seen or otherwise perceived, and the main part, which tested participant responses to the auditory stimuli in the presence and absence of air puffs. Stimuli were presented to the participants through ear-enclosing headphones (Sennheiser HD 202 II).

Puff detection test

In the first part of the experiment the participants heard a short tone (500 Hz, 1,000 ms long) in each trial, which was either followed by a 50 ms long air puff, or not followed by a puff. They were presented with two blocks of 50 trials each, in which 25 of the trials were accompanied by air puffs and 25 were not, presented in randomized order. In the first block the participant's right hand was located next to the exit of the tube such that they could feel the puff on the back of their hand (see **Figure 3A**). They were asked to press the "yes" key on a response box with their left hand if they felt or otherwise detected a puff, or the "no" key if they did not. In the second block, the task was the same, but their right hand was positioned on their lap, completely removed from the exit point of the tube (**Figure 3B**). The goal of this part of the experiment was to verify that the participants felt the puff on their hand but did not hear or see or otherwise detect it. In order to reduce the chances of hearing the puff of air, a small desk fan was used to provide a low level of background noise throughout the experiment. The fan was pointed to the wall and away from the participant. On average this portion of the experiment lasted about 5 min.

Perturbed continua testing

In the second part of the experiment, the participant's right hand was located such that they could feel the puff of air on the back of their hand (**Figure 3A**). In this part, blocks were presented during which sounds drawn from one of the three continua were tested: from /pa/ to /ba/,/ka/ to /ga/,or /hɛd/to/hɪd/. Only one continuum type was used within a given block. Each block included six repetitions of each step of the continuum in randomized order; three instances were accompanied by air puffs and three were not, also randomly ordered. Within a session, each participant heard ten blocks: either five velar blocks and five bilabial blocks, five bilabial

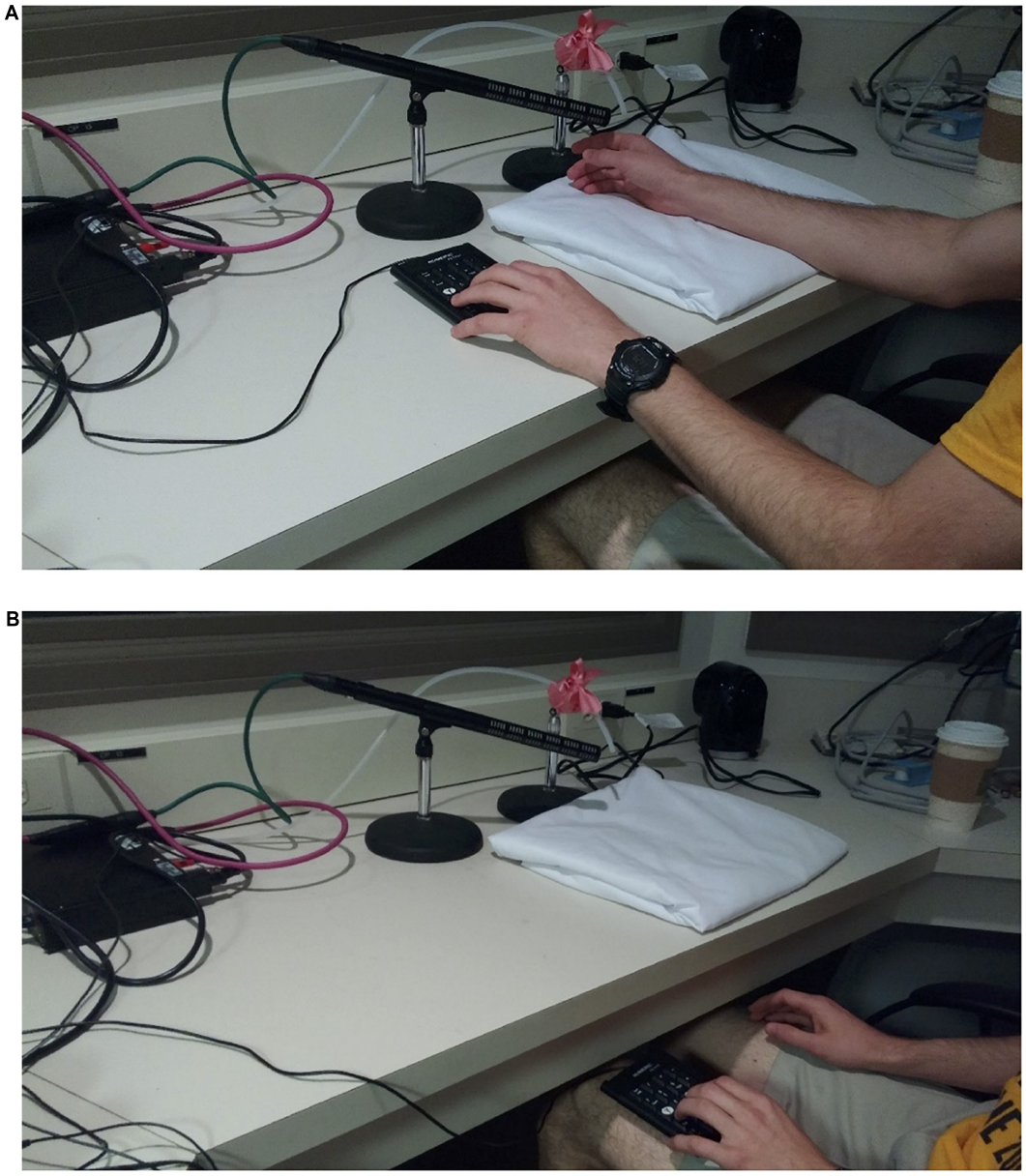


FIGURE 3
(A) Puff delivery setup: participant right hand placed near outflow of airtube, left hand on response button box. Microphone records air puff delivery for verification of timing. **(B)** Puff detection test setup: participant right hand positioned away from outflow of airtube. This test determines whether participant can detect airflow from cues other than tactile hand sensation.

blocks and five vowel blocks, or five velar blocks and five vowel blocks, with choices counterbalanced through the participant pool. This resulted in a total of 240 separate judgments [5 blocks × 3 repetitions × 2 puff conditions (−/+) × 8 continuum steps]. These numbers were chosen on the basis of piloting to keep the session to an approximate 45 min length, and for the same reason participants judged only two of the three possible continua during their session. Overall, 33 were tested for the bilabial continuum, 32 were tested for the velar continuum, and 19 were tested for the vowel control continuum.

Participants were asked to identify the stimulus they heard and to press the corresponding button on a response box on a computer screen: either "P" or "B" to indicate whether they heard /pa/or/ba/ during the bilabial blocks, "K" or "G" to indicate whether they heard /ka/or/ga/ during the velar blocks, and "head" or "hid" to indicate the word they heard during the vowel blocks. They were asked to respond as soon as they had made a decision, but were not constrained in time available for response. The reason for avoiding overt time pressure was our expectation that perceptual decisions involving

multimodal stimuli are more difficult, particularly when these are incongruent, as demonstrated by increased reaction times for McGurk studies with mismatched stimuli [see Alsius et al. (2017) for review]. Because we did not have *a priori* knowledge of how puffs could potentially delay formation of an integrated percept and did not wish to truncate that process we opted instead for participant-driven responses.

The presentation order of the continuum auditory stimuli and the accompanying tactile information (puff present vs. absent) were pseudo-randomized throughout each block. The blocks alternated such that there were no consecutive blocks of the same kind. For half the participants, the right button on the response box indicated a syllable with a voiceless consonant (e.g., "pa"). For the other half, the right button indicated a syllable with a voiced consonant. A similar counterbalancing was performed for the vowel blocks. In each trial the Matlab control procedure presented the audio stimulus, gated the air puff (or not), and recorded the participant choices from the response box as well as their response time. New trials began 1 s after each button-press response.

Results

Puff detection test

In the first block of the detection test, when their hand was close to the exit point of the tube, participants correctly discriminated puff/no puff conditions at an average rate of 98.1% (s.d. 2.6), with the worst performer at 90%. An exact binomial test confirms that these recognition percentages were well above chance ($p < 0.01$). In the second block, with their hand positioned away from the tube and everything else the same, participants were at chance: 50.4% (s.d. 2.6); best performer 57% (binomial test n.s.). These results confirm that the participants felt the puff of air on their hand, but could not hear, see, or otherwise detect it.

Perturbed continua testing

In 387 of all trials (1.9%) an air puff was requested but not delivered, or not requested but delivered, due to communication lapses with the solenoid controller. These problematic trials were identified using RMS peaks associated with (or missing from) the puff, measured from an acoustic recording made during the experiment (see microphone in **Figure 3A**) and were excluded from analysis. Although there was no time pressure to respond, an additional 85 trials were excluded because the button-press response time exceeded 8 s (∼5 s.d.), which was considered sufficiently long that the answer was potentially suspect. The data were then modeled with logistic regression in R (R Core Team, 2018) to estimate the effects of puffs on the perceptual boundary. **Figure 4** shows the estimated psychometric functions, pooled across speakers, in the presence and absence of air puffs. The vertical axis represents the probability of choosing a voiceless token or /ɛ/ (that is, "pa" in the case of the bilabial continuum, "ka" in the case of the velar continuum, or "head" in the case of the vowel continuum). The horizontal axis shows the 8 steps along the continuum. The baseline condition, without puff, is shown in blue lines with circles, and the condition with air puffs is shown in red lines with crosses. Vertical solid lines show the bias (50% crossover point), and vertical dotted lines mark the 25 and 75% probability points along each curve; the distance between these points gives the acuity (a measure of the slope of the boundary). The shift of the bias to the right in the presence of air puffs in the two VOT continua reflects the fact that there were more voiceless responses in this condition; this contrasts with the control vowel continuum which shows no shift in bias under puffs.

Quantifying the effect of puffs on perceived categories

A generalized linear mixed-effects model (GLMM) computed with the lme4 package (Bates et al., 2015) was used to assess the significance of the puffs contrast for each of the continua separately as they differ in step size, skewness, and type (the VOT continua were created by manipulating VOT duration, whereas the vowel continuum was created by manipulating formant structure). In this model[2] the dependent variable (the probability of choosing a voiceless or "head" response) was predicted by the fixed effects of PUFF (−/+) and continuum STEP, with random intercepts by participant ID [random slopes by participant were not supported by model comparison, $\chi^2(2) = 0.5094$, $p = 0.775$]. The results, summarized in **Table 3**, show a significant shift under +PUFF for the two VOT continua in the direction of increased judgment of voicelessness (bilabial $z = 3.16^{**}$, velar $z = 2.53^{*}$), and no effect of PUFF on the vowel continuum ($z = -0.31$). Marginal R^2 for these models (a measure of effect size), representing the proportion of variance explained by fixed factors alone, was computed using the method of Nakagawa and Schielzeth (2013), as implemented by Lefcheck and Sebastian Casallas (2014). The effect of STEP was significant for all continuum types. The addition of interaction terms for PUFF and STEP did not improve the fit of the model, in all three cases.

Comparison of effect sizes for the three continua

In order to compare the relative magnitudes of the puff effect we computed a second GLMM on the data combined from all three continua. In this model[3] the probability of choosing a

2 glmer (RESP∼PUFF + STEP + (1| ID), family = binomial).

3 glmer[RESP∼PUFF * CONT + CSTEP + (1 + CONT| ID), family = binomial].

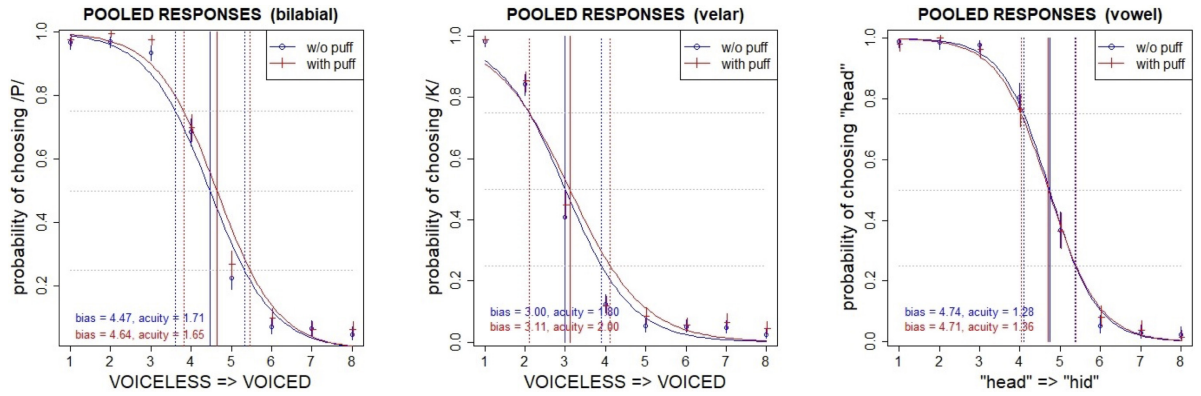


FIGURE 4
Perceived category boundaries, pooled across speakers, with (red) and without (blue) an air puff. Vertical lines show the bias (50%) crossover, which is systematically shifted in the direction of voiced responses for +puff trials in the bilabial (left) and velar (center) continua, but not in the control vowel continuum (right). 95% confidence intervals are indicated for each pooled response.

voiceless or "head" response was predicted by the fixed effects of PUFF and CONTinuum type and their interaction, and a continuous CSTEP covariate, with random slopes for CONT by participant ID [random slopes for PUFF were not supported by model comparison, $\chi^2(3) = 0.4445$, $p = 0.931$]. The results are shown in Table 4.

For this model the baseline (intercept) encodes the response for −PUFF, Vowel continuum, and CSTEP = 1, and the corresponding odds show the overwhelming preference for voiceless or "head" responses under this condition (1872.2–1). The significant main effect for CONTvel ($z = -6.85^{**}$) reflects the leftward skew of the velar continuum (illustrated in Figure 4); i.e., in the direction of increased voiced responses over baseline. The continuous CSTEP covariate (continuum step) has the expected negative correlation with stimulus VOT and vowel quality (voiceless > voiced, "head" > "hid"). Because of the inclusion of the non-responsive vowel control, the overall effect of +PUFF is not significant, but its interactions with the two VOT continua show significant positive shifts in the direction of increased voiceless responses over baseline (velar $z = 1.76$, bilabial $z = 2.27^{*}$). This is confirmed through *post hoc* (Tukey HSD) comparisons of +PUFF > −PUFF, which show velar $z = 2.48^{*}$ and bilabial $z = 3.35^{**}$. The odds ratios for these interactions show the ratio by which the odds ratios for the main effects (CONTvel/CONTvow, CONTbil/CONTvow) changes for +PUFF; i.e., their relative increase over baseline. Interpreted as an effect size this indicates that +PUFF had a greater effect on the bilabial continuum (odds ratios = 1.36) than the velar continuum (odds ratios = 1.27); however, the 95% confidence intervals overlap for these values, and the significance within this model for the velar interaction is marginal.

TABLE 3 Output of the GLMM response model for each continuum.

Continuum	−Air PUFF (baseline) vs. +Air PUFF			
	Coefficients	*z*-value	*P*-value	Marginal R^2
Bilabial	0.244	3.160	0.0016**	0.733
Velar	0.216	2.533	0.0113*	0.699
Vowel	−0.037	−0.313	0.7540 n.s.	0.817

For the two VOT continua the effect of +PUFF was to increase the likelihood of a voiceless response; the vowel control continuum was unaffected. R^2 shows the proportion of variance explained by the fixed factors alone. (**$p < 0.01$; *$p < 0.05$.)

Analysis of individual results

To assess the degree to which individual participants were sensitive to the air puff effect we computed separate logistic regression models for each, with response predicted by the fixed effect of PUFF and STEP as a continuous covariate.[4] About two thirds of the participants who heard the bilabial continuum showed a shift toward increased probability of voiceless responses (23/33; binomial test $p < 0.02$), as did about three quarters of the participants who heard the velar continuum (24/32; binomial test $p < 0.01$). About half of the participants who heard the vowel continuum showed small and non-significant shifts toward 'head' responses (9/19; n.s.). See Table 5 for summary statistics.

Analysis of response times

Response times were measured as the duration in milliseconds from the onset of the audio stimulus (which was coincident with the start of the air puff, if present), to the button-press event. For analysis they were log-scaled in order to normalize a right-skewed distribution. Figure 5 illustrates the mean response times pooled across participants, by PUFF, CONTinuum type, and STEP along the continuum.

4 glm (RESP∼PUFF + CSTEP, family = binomial).

TABLE 4 Output of GLMM combining continua to show relative effect sizes (using odds ratios).

	Coefficients	*z*-value	*P*-value	Odds ratios	95% confidence intervals
(Intercept)	7.53485	30.22	0.000	1872.162	(1148.36, 3052.16)
+PUFF	−0.03315	−0.31	0.758	n.s.	
CONTvel	−2.72139	−6.85	0.000	0.066	(0.030, 0.143)
CONTbil	−0.45373	−1.57	0.117	n.s.	
STEP	−1.59468	−66.77	0.000	0.203	(0.194, 0.213)
+PUFF:CONTvel	0.23953	1.76	0.078	1.271	(0.973, 1.659)
+PUFF:CONTbil	0.23953	2.21	0.023	1.360	(1.043, 1.772)

Marginal R^2 for this model is 0.756. The interaction terms show the ratio by which the odds ratio of each VOT continuum relative to the Vowel baseline changes for +PUFF, with a larger magnitude observed for the bilabial continuum than the velar.

An overall effect of CONTinuum type was observed, with bilabial responses slower than velar responses in general, and both significantly slower than vowel control responses.

A linear mixed-effects model[5] computed using lme4 with significance assessed using the lmerTest package (Kuznetsova et al., 2017) in R was used to predict the $\log_{10}$ response time from the fixed effects of PUFF, CONTinuum, and (discrete) continuum STEPs. Model comparison supported the complete interaction between fixed factors and the inclusion of random slopes and intercepts for each by participant. The analysis modeled discrete rather than continuous steps along the continuum to investigate how response time interacted with stimulus, with the expectation that responses to stimuli in the ambiguous range of each continuum would be slower. Significant results are shown in Table 6.

The pattern of main effects confirms that response times are slower for the ambiguous intermediate steps (4, 5, 6), and that responses for the two VOT continua are slower overall than for the vowel control baseline, with the bilabial responses slower than the velar. The interaction of STEP with the velar continuum reflects its left-skewed crossover, such that step 3 (closest to the boundary and thus most ambiguous) is significantly slower, while subsequent steps are faster relative to baseline. The negative coefficient for the interaction of +PUFF and the bilabial continuum suggests an overall facilitation effect (responses are faster than baseline), which Figure 5 suggests is active on the voiceless end of the continuum (steps 1, 2). This effect was likely due to the congruent nature of the added information, namely, consistent with what would be felt on the hand if placed near the mouth during production of a voiceless stop. The interaction of steps 5, 6, and 7 with the bilabial continuum shows that these responses were significantly faster than baseline *without* puffs, and significantly slower than baseline *with* puffs, indicating that over this portion of the continuum puffs represented an incongruent and thus inhibitory distraction, perhaps because of the mismatch of consistent puff duration with reduced aspiration for these tokens. Similar reaction time effects of secondary information for unambiguous portions of a continuum have been previously reported (e.g., Whalen, 1984). The differential effects of air puffs on response times argue against a unimodal effect and instead suggest that tactile cues are weighted according to both relevance and congruence.

5 lmer [LRT∼PUFF * STEP * CONT + (1 + PUFF + STEP + CONT| | ID)].

TABLE 5 Summary of the individual models computed for the participants.

Continuum	Mean coefficient	s.d. of coefficient	Range of coefficient
Bilabial	0.26766	0.479	−0.87388: 1.66863
Velar	0.21979	0.546	−0.83977: 0.98542
Vowel	−0.00845	−0.548	−0.99308: 1.02929

Discussion

The current study found that presence of air puffs significantly increased the likelihood of choosing voiceless responses for the two VOT continua, and consequently the category boundaries for both VOT continua were shifted toward the voiced end of each continuum in the presence of air puffs. The effect was found to be larger for the bilabial continuum than for the velar continuum, though not significantly so. The observed difference may be due to the unbalanced (left-skewed) velar continuum. Air puffs had no effect on choices for the control vowel continuum.

In this work VOT continua were used rather than endpoints alone to address the critique raised by Massaro (2009). Gick and Derrick (2009) and Derrick and Gick (2013) used CV exemplars presented in background noise. Because this degraded acoustic signal might not be sufficient for categorization listeners could be simply disregarding it, and instead be relying solely on the presence or absence of the aero-tactile cue. However, the current study shows that an air-puff alone in each trial was not sufficient for deciding the category, and that listeners instead weighted the tactile cue both by relevance (no effect on the vowel continuum)

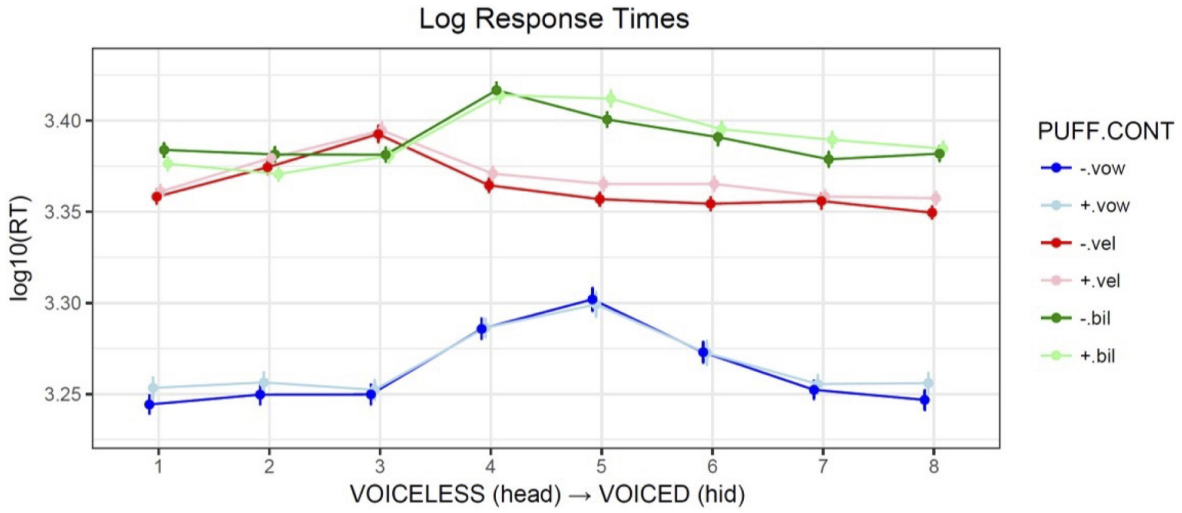


FIGURE 5
Comparison of mean $\log_{10}$ response times averaged across participants, by PUFF (+airflow vs. −airflow), CONTinuum type (VOWel, VELar, BILabial), and continuum STEP. Error bars show the standard error of the mean.

and quality of the auditory signal (minimal effects at endpoints, maximal effects near the ambiguous crossover point of the VOT continua). While the presence of an air puff did result in more voiceless responses, these acted to shift the existing perceptual boundary rather than overriding it; in other words they did not uniformly increase voiceless responses at every continuum step. This suggests that aero-tactile sensation was processed as a potential additional cue for disambiguation of voiceless from voiced sounds, but weighted by relevance and the degree of ambiguity, as an instance of true multi-sensory integration.

TABLE 6 Output of LMM predicting $\log_{10}$ response times from PUFF, CONTinuum, and stimulus STEP along the continuum.

	Coefficients	*t*-value	*P*-value	Significance
STEP4	0.04363	6.248	0.000	***
STEP5	0.06053	8.575	0.000	***
STEP6	0.03054	4.246	0.000	***
CONTvel	0.1026	14.307	0.000	***
CONTbil	0.1275	18.295	0.000	***
+PUFF:CONTbil	−0.01717	−2.163	0.031	*
STEP3:CONTvel	0.02652	3.248	0.001	**
STEP4:CONTvel	−0.03783	−4.552	0.000	***
STEP5:CONTvel	−0.06216	−7.539	0.000	***
STEP6:CONTvel	−0.03499	−4.247	0.000	***
STEP8:CONTvel	−0.01404	−1.729	0.084	
STEP5:CONTbil	−0.04550	−5.550	0.000	***
STEP6:CONTbil	−0.02480	−3.022	0.003	**
STEP7:CONTbil	−0.01666	−2.049	0.040	*
+PUFF:STEP5:CONTbil	0.03088	2.759	0.006	**
+PUFF:STEP6:CONTbil	0.0214	1.911	0.056	
+PUFF:STEP7:CONTbil	0.02473	2.210	0.027	*

The baseline represents −PUFF at STEP1 on the Vowel continuum. Only significant values are shown. Pseudo-R^2 for this model (comparison of fitted vs. observed values) is 0.447. (***$p < 0.001$; **$p < 0.01$; *$p < 0.05$).

Although participants were not instructed to answer as quickly as possible, analysis of response times did reveal significant differences between continua and within continua. The intermediate steps of the continua, that is, the ambiguous stimuli between the two endpoints, were the hardest for participants to categorize, as expected. This was suggested by the longer response times associated with these steps, for all three continua, as longer response times generally indicate a greater cognitive load (e.g., DeLeeuw and Mayer, 2008). For the two VOT continua in general response times were slower than the vowel control baseline. It is important to note that the response times for the VOT continua did not show a uniform response to air puffs, shown most clearly by the bilabial continuum. As illustrated in **Figure 5** and shown by the results in **Table 6**, air puffs had a *facilitatory* effect at the voiceless end of the continuum (encoded by the negative + PUFF:CONTbil interaction; $t = -2.2^{*}$); i.e., responses were faster with puffs. This effect was likely caused by the complementary nature of the added information (cf. Whalen, 1984). Conversely, air puffs at the voiced end of the continuum had an *inhibitory* effect [encoded by the positive + PUFF:STEP:CONTbil interaction for steps 5 ($t = 2.8^{**}$), 6 ($t = 1.9\cdot$), and 7 ($t = 2.2$)]. In this case, the added information was incongruent. As no overall main effect of PUFF was observed, these results are inconsistent with Massaro's position that listeners respond to air puff stimuli unimodally; rather, the pattern of results indicates that an air puff cue is evaluated together with the concurrent audio stimulus and weighted by the ambiguity of the latter.

We have mentioned, in the Introduction, that evidence for multisensory integration has been used to argue in favor of certain approaches for the objects of speech perception. The argument was that if non-acoustic information, tactile in the current case, is an integral part of the process of speech perception, the objects of speech perception cannot be auditory, or at least not exclusively auditory. A counter argument that has been discussed in the literature is that the association with visual and other sensory information may be learned, that is, associated by experience with the auditory primitives (e.g., Massaro, 1987; Diehl and Kluender, 1989; Kluender, 1994). Rosenblum (2008a) offers a few arguments against learned association: first, multisensory integration has been shown in pre-linguistic infants (Rosenblum et al., 1997). Second, multisensory integration has been shown to operate at an early stage of online perception, before phonetic categorization and possibly before phonetic feature extraction (Summerfield, 1987; Green, 1998; Rosenblum and Gordon, 2001; Choi et al., 2021). The evidence for multisensory integration at an early stage of speech processing is consistent with evidence for multisensory integration in other domains [for discussion see Shimojo and Shams (2001), Stoffregen and Bardy (2001); but see Remez et al. (1998)]. Multisensory integration has been shown in contexts where participants had no speech experience associated with the task (Fowler and Dekle, 1991). However, in the experiment conducted by Fowler and Dekle the participants were aware of the task and thus it is not clear that this is indeed a counter-argument for learned association.

Based on the evidence cited above, Rosenblum (2008a) argues that the objects of speech perception are modality-neutral. Specifically, he argues for gestural objects that have spatial and temporal dimensions but are not specified along any sensory dimension. According to this view the sensory dimensions are the medium through which perceivers recover the gestures, and the objects of speech perception themselves are of a higher order than just auditory, visual or tactile. The idea is that perception is sensitive to underlying gestural primitives instantiated in any modality. This view, which is consistent with Direct Realism, Motor Theory and Articulatory Phonology, is supported by the cited evidence for the automaticity and ubiquity of multisensory integration. However, it is not the only view that is consistent with such evidence. It may be the case that the objects of speech perception do have a sensory content, but they are specified for more than one modality. That is, it may be the case that they are not just auditory, but multimodal in nature. The evidence presented here suggests that tactile information is considered during the perception of speech [and see as well Bruderer et al. (2015) and Choi et al. (2021)]. However, it does not rule out the possibility that the integration of the additional tactile modality operates in later stages of online perception.

The lack of an obvious connection between distal aero-tactile stimulation and speech perception in the current experiment contrasts with the direct somatosensory link posited by Ito et al. (2009). In their experiment they determined that perception of vowels is affected by deforming the skin on the face of the participant in the same way the skin moves when these vowels are produced. Crucially, deformations applied orthogonal to the up and down directions used in the production of these vowels had no effect. This kind of direct link between somatosensory stimulus and speech perception is not reflected in the current study, as air puffs were applied on the back of hand of the participants, a location that does not typically relate directly to the tactile sensation of aspiration during the production of stop consonants. Nonetheless, the results presented here confirm that aero-tactile stimulation can also shift perception, though only when the cue is relevant (vowel perception was unaffected) and the primary VOT cues are ambiguous. In both the air puff and skin pull studies then, tactile information affected speech perception only when the cues applied were congruent with the ones expected in production of the perceived sounds.

In addition to addressing Massaro's critique against Gick and Derrick (2009) and Derrick and Gick (2013), and providing evidence for integration of auditory and tactile input in the perception of speech, the current work extends the work of Gick and Derrick in two additional ways. First, rather than a between-subject design, here a within-subject design was used in which each participant served as their own control. Thus, the comparison between the perception of the VOT continua with and without tactile stimuli was done within participant, and not across groups of participants. This allowed a direct comparison between the responses of the same individual to the same auditory stimuli with and without aero-tactile stimulus. Second, a vowel continuum was used as a control. Since aero-tactile sensation is hypothesized not to be relevant for distinguishing /ɛ/ from /ɪ/, effects observed on the VOT continua but not on the vowel continuum shows that the obtained results were not just an artifact of puffs alone, but rather a context-sensitive effect, indicating a true multi-sensory phenomenon. Moreover, since this was a within-subject design, the comparison between the VOT continuum and the vowel continuum was done within participant. That is, the participants that heard vowel blocks were sensitive to the effect of aero-tactile stimulation when the acoustic stimuli were taken from a VOT continuum, and at the same time showed no such sensitivity when the acoustic stimuli were taken from a vowel continuum. As discussed above, these results are consistent with Ito et al. (2009) showing that while tactile cues can indeed modulate perception, they do so only when congruent with the production contrast being disambiguated.

While statistically significant, the effect of puffs found in this study was not observed for all the participants, similar to other studies of multimodal integration. Population estimates of audiovisual integration susceptibility vary widely and range between 26 and 98% of the tested population

(Nath and Beauchamp, 2012). In the current study, between two thirds (in the bilabial continuum) and three quarters (in the velar continuum) of the participants showed susceptibility to puffs in their responses. These clear majorities contrast with participants who showed some effect of puff on their response to the vowel continuum (about half), though of these shifts, none were significant. The absence of effect on the VOT continua for some of the participants may stem from lack of sufficient statistical power, given the small size of the effect and further division of the data into participant-sized bins, though for most of the participants a significant effect was found even after the division of the data. Finally, it is possible that some of the participants were not affected by the aero-tactile stimuli because of the relatively low airflow (5 SLPM), in comparison to the average airflow of voiceless stop consonants in CV syllables (about 56 SLPM, Isshiki and Ringel, 1964). Although the puff detection test has confirmed that these participants have felt the puff, it is possible that they did not interpret it as related to aspiration since the airflow was inconsistent with the typical airflow of speech.

The current study did not test the length of the integration window, as it did not vary the relative timing of the auditory stimuli and the tactile stimuli. However, it has been shown previously that this window operates asymmetrically. Derrick et al. (2009) and Gick et al. (2010) found for audio-tactile stimuli that integration extends to 200 ms when air puff follows audio but only 50 ms when air puff precedes audio. Bicevskis (2015) studied visuo-tactile integration by presenting participants with video of faces producing the syllables /pa/ and /ba/, without an air puff, or accompanied by an air puff occurring synchronously with the visual stimuli or at different timings, up to 300 ms before and after the stop release. Bicevskis found that the integration window for visuo-tactile stimuli is also asymmetric: when an air puff followed visual stimuli the integration window extended to 300 ms, but when it preceded visual stimuli the integration window only extended to 100 ms. These windows extend farther than the audiovisual integration window reported by Munhall et al. (1996) for McGurk phenomena (0–180 ms) and also Van Wassenhove et al. (2007) (−30 to 170 ms) but exhibit the same properties of asymmetry. The asymmetry appears to be ordered by the relative speed by which each modality is processed: visual input is processed more slowly than auditory (Molholm et al., 2002), and tactile sensation is also slower than audition. Munhall et al. (1996) suggest that knowledge of the natural world may play a role in validating the range over which integration is permitted to occur; e.g., thunder is expected to follow lightning, and air turbulence is typically heard before it is felt. Thus, relative timings of potential speech cues that violate these expectations are potentially less likely to be integrated.

The tolerance for asynchrony in multimodal integration differs from that observed for parsing the acoustic signal alone. For example, work by Remez and colleagues confirms that individual tones in sinewave speech are not separate streams needing integration but are instead necessarily tightly timed (within 50 ms) in order to provide speech information Remez et al. (2008, 2010). Similarly, if a non-speech "chirp" in a duplex paradigm precedes a speech third formant (F3) by more than 50 ms, the non-speech percept generally "captures" the F3, leaving the ambiguous base as the percept (Whalen and Liberman, 1996). For multimodal integration, the tolerances are greater, presumably due to the need to buffer separately acquired channels with differing inherent timescales for perception.

The limited activation of speech percepts by the puffs themselves further argues for an integrative rather than a unimodal biasing process. In audiovisual integration, it is clear that both channels can convey the speech signal at greater than chance levels independently, though not to the same degree (phonemes > visemes), and their respective weighting in combination can vary with ambient factors such as background noise (Vatikiotis-Bateson et al., 1998). Because the tactile sensation of puffs alone has insufficient bandwidth to convey anything like the full speech signal, its potential effects are limited to those restricted cases where acoustic ambiguity lowers the threshold for such cues to become relevant in producing an integrated percept.

The mismatch in bandwidth capacity, processing speeds and tolerance for asynchrony suggests that some form of perceptual buffering exists for each contributing modality, which is then weighted to form the composite percept (Rosenblum et al., 2017). But although we have observed an effect of distal aero-tactile stimulation on speech perception, we have not provided an explanation for why the phenomenon occurs. Numerous studies have shown that listeners can make use of all available information, "parsing" it into plausible percepts (Fowler and Smith, 1986; Fowler and Brown, 2000) and rejecting components that do not parse as being simultaneous non-speech (Xu et al., 1997). Multimodal integration indicates that such speech parsing goes beyond the acoustic signal to include all aspects of the production (Liberman and Mattingly, 1985; cf. MacDonald, 2018). There is considerable evidence that much of this integration can occur before much, if any, experience has been attained [supported by Bruderer et al. (2015)]. Still, if familiarity plays a role in the uptake of multisensory information, the use of tactile (puff) information is puzzling. It is possible that close proximity of the hand to the mouth during the babbling phase of language acquisition might develop a learned association between aspiration and tactile sensation felt there. Similarly, such association may also arise from exposure as children to speech produced by others who are in close proximity to them. Hall (1966) defined four spaces encircling every person. The most inner space, the intimate space, is characterized as the spaces closest to the body, up to 45 cm away from it. This is a space reserved for sexual partners and children. This distance is sufficiently short for aspirated stops to be felt on the skin of a child or a partner. Children are also

found in close proximity to others during social interaction with their peers: Aiello and Jones (1971) studied the proxemic behavior of children ages 6–8 and found that the mean distance between children during social interaction differed by sex and sub-culture, but overall ranged between 5.3 and 13.5 inches, a distance sufficiently short for aspirated stops to be felt on the skin. Aiello and De Carlo Aiello (1974) found that personal space grows bigger as children grow older, suggesting that the chance of being exposed to felt aspiration at younger age is larger than it is in conversations at later stages of life. Because such stimulation would not be particularly localized to a single point of contact, the association between aspiration and tactile sensation could then eventually be generalized to any skin location, consistent with Gick and Derrick (2009) and Derrick and Gick (2013) who show that air puffs affect VOT perception when the point of contact is the neck or even the ankle. They also show that not just any tactile stimulus produces the effect, as tapping the skin at the same location as delivered air puffs did not affect perception, and this selective response suggests some type of learned link between aspiration and air puffs rather than a general tactile effect. However, while the pathway to acquiring an association between VOT aspiration and the tactile sensation specific to feeling its effect on the skin is speculative, the results from Gick and Derrick (2009) and this confirmatory study indicate that such an association is real. Once available, tactile information joins other potential cues (visual, lexical, etc.) available for exploitation by language users to disambiguate the speech signal.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by Yale Human Research Protection Program. The participants provided their written informed consent to participate in this study.

Author contributions

DG constructed the experimental setup and collected the data. MT wrote the controlling software and performed data analysis. All authors designed the research, wrote the manuscript, and approved the submitted version.

Funding

This research had been supported by National Institutes of Health (NIH) Grant DC-002717 to Haskins Laboratories.

Acknowledgments

The authors are grateful to Dr. Evyatar Shaulsky for his assistance with setting up and maintaining the air-puffing system.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abramson, A. S., and Whalen, D. H. (2017). Voice Onset Time (VOT) at 50: theoretical and practical issues in measuring voicing distinctions. *J. Phonetics* 63, 75–86. doi: 10.1016/j.wocn.2017.05.002

Aiello, J. R., and De Carlo Aiello, T. (1974). The development of personal space: proxemic behavior of children 6 through 16. *Hum. Ecol.* 2, 177–189. doi: 10.1007/BF01531420

Aiello, J. R., and Jones, S. E. (1971). Field study of the proxemic behavior of young school children in three subcultural groups. *J. Pers. Soc. Psychol.* 19:351. doi: 10.1037/h0031433

Alsius, A., Paré, M., and Munhall, K. G. (2017). Forty years after hearing lips and seeing voices: the McGurk effect revisited. *Multis. Res.* 31, 111–144. doi: 10.1163/22134808-00002565

Arnold, P., and Hill, F. (2001). Bisensory augmentation: a speechreading advantage when speech is clearly audible and intact. *Br. J. Psychol.* 92, 339–355. doi: 10.1348/000712601162220

Banati, R. B., Goerres, G., Tjoa, C., Aggleton, J. P., and Grasby, P. (2000). The functional anatomy of visual-tactile integration in man: a study using positron emission tomography. *Neuropsychologia* 38, 115–124.

Bates, D., Machler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Bernstein, L. E., Demorest, M. E., Coulter, D. C., and O'Connell, M. P. (1991). Lipreading sentences with vibrotactile vocoders: performance of normal-hearing and hearing-impaired subjects. *J. Acoust. Soc. Am.* 90, 2971–2984. doi: 10.1121/1.401771

Bicevskis, K. (2015). *Visual-Tactile Integration and Individual Differences in Speech Perception*. Vancouver, BC: University of British Columbia. Unpublished MA Thesis.

Browman, C. P., and Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology* 6, 201–251. doi: 10.1017/S0952675700001019

Browman, C. P., and Goldstein, L. (1992). Articulatory phonology: an overview. *Phonetica* 49, 155–180. doi: 10.1159/000261913

Browman, C. P., and Goldstein, L. M. (1986). Towards an articulatory phonology. *Phonology* 3, 219–252. doi: 10.1017/S0952675700000658

Bruderer, A. G., Danielson, D. K., Kandhadai, P., and Werker, J. F. (2015). Sensorimotor influences on speech perception in infancy. *Proc. Natl. Acad. Sci. U.S.A.* 112, 13531–13536. doi: 10.1073/pnas.1508631112

Burnham, D., and Dodd, B. (1996). "Auditory-visual speech perception as a direct process: the McGurk effect in infants and across languages," in *Speechreading by Humans and Machines: Models, Systems and Applications*, eds D. Stork and M. Hennecke (Berlin: Springer), 103–114. doi: 10.1007/978-3-662-13015-5_7

Byrd, D. (1993). 54,000 American stops. *UCLA Work. Papers Phonetics* 83, 97–116.

Calvert, G. A. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cereb. Cortex* 11, 1110–1123. doi: 10.1093/cercor/11.12.1110

Choi, D., Dehaene-Lambertz, G., Peña, M., and Werker, J. F. (2021). Neural indicators of articulator-specific sensorimotor influences on infant speech perception. *Proc. Natl. Acad. Sci. U.S.A.* 118:e2025043118. doi: 10.1073/pnas.2025043118

Colonius, H., and Diederich, A. (2004). Multisensory interaction in saccadic reaction time: a time-window-of-integration model. *J. Cogn. Neurosci.* 16, 1000–1009. doi: 10.1162/0898929041502733

Cooper, A. M. (1991). *An Articulatory Account of Aspiration in English*. New Haven, CT: Yale University. Unpublished Ph.D. Thesis.

DeLeeuw, K. E., and Mayer, R. E. (2008). A comparison of three measures of cognitive load: evidence for separable measures of intrinsic, extraneous, and germane load. *J. Educ. Psychol.* 100:223. doi: 10.1037/0022-0663.100.1.223

Derrick, D., and Gick, B. (2013). Aerotactile integration from distal skin stimuli. *Multisens. Res.* 26, 405–416. doi: 10.1163/22134808-00002427

Derrick, D., Anderson, P., Gick, B., and Green, S. (2009). Characteristics of air puffs produced in English "pa": experiments and simulations. *J. Acoust. Soc. Am.* 125, 2272–2281. doi: 10.1121/1.3081496

Diehl, R. L., and Kluender, K. R. (1989). On the objects of speech perception. *Ecol. Psychol.* 1, 121–144. doi: 10.1207/s15326969eco0102_2

Diehl, R. L., Lotto, A. J., and Holt, L. L. (2004). Speech perception. *Annu. Rev. Psychol.* 55, 149–179. doi: 10.1146/annurev.psych.55.090902.142028

Eagleman, D. M. (2008). Human time perception and its illusions. *Curr. Opin. Neurobiol.* 18, 131–136. doi: 10.1016/j.conb.2008.06.002

Eagleman, D. M., and Holcombe, A. O. (2002). Causality and the perception of time. *Trends Cogn. Sci.* 6, 323–325. doi: 10.1016/S1364-6613(02)01945-9

Fastl, H., and Zwicker, E. (2006). *Psychoacoustics: Facts and Models*. Berlin: Springer Verlag. doi: 10.1007/978-3-540-68888-4

Fowler, C. A. (1981). Production and perception of coarticulation among stressed and unstressed vowels. *J. Speech Lang. Hear. Res.* 24, 127–139. doi: 10.1044/jshr.2401.127

Fowler, C. A. (1984). Segmentation of coarticulated speech in perception. *Percept. Psychophys.* 36, 359–368. doi: 10.3758/BF03202790

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *J. Acoust. Soc. Am.* 99, 1730–1741. doi: 10.1121/1.415237

Fowler, C. A., and Dekle, D. J. (1991). Listening with eye and hand: cross-modal contributions to speech perception. *J. Exp. Psychol. Hum. Percept. Perform.* 17:816. doi: 10.1037/0096-1523.17.3.816

Fowler, C., and Brown, J. (2000). Perceptual parsing of acoustic consequences of velum lowering from information for vowels. *Percept. Psychophys.* 62, 21–32. doi: 10.3758/BF03212058

Fowler, C., and Smith, M. (1986). "Speech perception as "vector analysis": an approach to the problems of segmentation and invariance," in *Invariance and Variability in Speech Processes*, eds J. Perkell and D. Klatt (Hillsdale, NJ: Lawrence Erlbaum Associates), 123–136.

Galantucci, B., Fowler, C. A., and Goldstein, L. (2009). Perceptuomotor compatibility effects in speech. *Attent. Percept. Psychophys.* 71, 1138–1149. doi: 10.3758/APP.71.5.1138

Geers, A., and Brenner, C. (1994). Speech perception results: audition and lipreading enhancement. *Volta Rev.* 96, 97–108.

Gick, B., and Derrick, D. (2009). Aero-tactile integration in speech perception. *Nature* 462, 502–504. doi: 10.1038/nature08572

Gick, B., Ikegami, Y., and Derrick, D. (2010). The temporal window of audio-tactile integration in speech perception. *J. Acoust. Soc. Am.* 128, EL342–EL346. doi: 10.1121/1.3505759

Gick, B., Jóhannsdóttir, K. M., Gibraiel, D., and Mühlbauer, J. (2008). Tactile enhancement of auditory and visual speech perception in untrained perceivers. *J. Acoust. Soc. Am.* 123, EL72–EL76. doi: 10.1121/1.2884349

Goldstein, L., and Fowler, C. A. (2003). "Articulatory phonology: a phonology for public language use," in *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*, eds N. O. Schiller, and A. S. Meyer (Berlin: Mouton de Gruyter), 159–207.

Grant, K. W., and Seitz, P.-F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *J. Acoust. Soc. Am.* 108, 1197–1208. doi: 10.1121/1.1288668

Green, K. (1998). "The use of auditory and visual information during phonetic processing: implications for theories of speech perception," in *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech*, eds B. Dodd and R. Campbell (Hove: Psychology Press), 3–25.

Haggard, P., Clark, S., and Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nat. Neurosci.* 5, 382–385. doi: 10.1038/nn827

Hall, E. T. (1966). *The Hidden Dimension*. Garden City, NY: Doubleday.

Hardison, D. M. (2007). "The visual element in phonological perception and learning," in *Phonology in Context*, ed. M. Pennington (London: Palgrave Macmillan), 135–158. doi: 10.1057/9780230625396_6

Harrar, V., and Harris, L. R. (2008). The effect of exposure to asynchronous audio, visual, and tactile stimulus combinations on the perception of simultaneity. *Exp. Brain Res.* 186, 517–524. doi: 10.1007/s00221-007-1253-0

Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. *Nat. Rev. Neurosci.* 8, 393–402. doi: 10.1038/nrn2113

Isshiki, N., and Ringel, R. (1964). Air flow during the production of selected consonants. *J. Speech Hear. Res.* 7, 233–244. doi: 10.1044/jshr.0703.233

Isshiki, N., and von Leden, H. (1964). Hoarseness: aerodynamic studies. *Arch. Otolaryngol.* 80, 206–213. doi: 10.1001/archotol.1964.00750040212020

Ito, T., Tiede, M., and Ostry, D. J. (2009). Somatosensory function in speech perception. *Proc. Natl. Acad. Sci. U.S.A.* 106, 1245–1248. doi: 10.1073/pnas.0810063106

Jaeger, J. J. (1978). "Speech aerodynamics and phonological universals," in *Proceedings of the 4th Annual Meeting of the Berkeley Linguistics Society* (Berkeley, CA), 312–329. doi: 10.3765/bls.v4i0.2221

Kaiser, A. R., Kirk, K. I., Lachs, L., and Pisoni, D. B. (2003). Talker and lexical effects on audiovisual word recognition by adults with cochlear implants. *J. Speech Lang. Hear. Res.* 46, 390–404. doi: 10.1044/1092-4388(2003/032)

Klatt, D. H. (1979). Speech perception: a model of acoustic–phonetic analysis and lexical access. *J. Phonetics* 7, 279–312. doi: 10.1016/S0095-4470(19)31059-9

Kluender, K. R. (1994). "Speech perception as a tractable problem in cognitive science," in *Handbook of Psycholinguistics*, ed. M. Gernsbacher (San Diego, CA: Academic Press), 173–217.

Kuznetsova, A., Brockhoff, P., and Christensen, R. (2017). lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* 82, 1–26. doi: 10.18637/jss.v082.i13

Lachs, L., Pisoni, D. B., and Kirk, K. I. (2001). Use of audiovisual information in speech perception by prelingually deaf children with cochlear implants: a first report. *Ear Hear.* 22:236. doi: 10.1097/00003446-200106000-00007

Lee, J.-T., Bollegala, D., and Luo, S. (2019). ""Touching to see" and "seeing to feel": robotic cross-modal sensory data generation for visual-tactile perception," in *Proceedings of the 2019 International Conference on Robotics and Automation (ICRA)* (Montreal, QC: IEEE), 4276–4282. doi: 10.1109/ICRA.2019.8793763

Lefcheck, J., and Sebastian Casallas, J. (2014). *R-Squared for Generalized Linear Mixed-Effects Models*. Available online at: https://github.com/jslefche/rsquared.glmm (accessed September, 2016).

Liberman, A. M., and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition* 21, 1–36. doi: 10.1016/0010-0277(85)90021-6

Liberman, A. M., and Whalen, D. H. (2000). On the relation of speech to language. *Trends Cogn. Sci.* 4, 187–196. doi: 10.1016/S1364-6613(00)01471-6

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychol. Rev.* 74:431. doi: 10.1037/h0020279

Lisker, L., and Abramson, A. S. (1964). A cross-language study of voicing in initial stops: acoustical measurements. *Word* 20, 384–422. doi: 10.1080/00437956.1964.11659830

Lisker, L., and Abramson, A. S. (1967). Some effects of context on voice onset time in English stops. *Lang. Speech* 10, 1–28. doi: 10.1177/002383096701000101

MacDonald, J. (2018). Hearing lips and seeing voices: the origins and development of the 'McGurk Effect'and reflections on audio–visual speech perception over the last 40 years. *Multisens. Res.* 31, 7–18. doi: 10.1163/22134808-00002548

MacLeod, A., and Summerfield, Q. (1990). A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise: rationale, evaluation, and recommendations for use. *Br. J. Audiol.* 24, 29–43. doi: 10.3109/03005369009077840

Massaro, D. (1987). "Speech perception by ear and eye," in *Hearing by Eye: The Psychology of Lip-Reading*, eds B. Dodd and R. Campbell (Hillsdale, NJ: Lawrence Erlbaum), 53–83.

Massaro, D. (2009). *Caveat emptor*: the meaning of perception and integration in speech perception. *Nat. Prec.* 1–1. doi: 10.1038/npre.2009.4016.1

Massaro, D. W., Cohen, M. M., Gesi, A., Heredia, R., and Tsuzaki, M. (1993). Bimodal speech perception: an examination across languages. *J. Phonet.* 21, 445–478. doi: 10.1016/S0095-4470(19)30230-X

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748. doi: 10.1038/264746a0

Mills, A. (1987). "The development of phonology in the blind child," in *Hearing by Eye: The Psychology of Lip Reading*, eds B. Dodd and R. Campbell (Hillsdale, NJ: Lawrence Erlbaum), 145–162.

Miyazaki, M., Yamamoto, S., Uchida, S., and Kitazawa, S. (2006). Bayesian calibration of simultaneity in tactile temporal order judgment. *Nat. Neurosci.* 9, 875–877. doi: 10.1038/nn1712

Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E., and Foxe, J. J. (2002). Multisensory auditory–visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Cogn. Brain Res.* 14, 115–128. doi: 10.1016/S0926-6410(02)00066-6

Munhall, K. G., Gribble, P., Sacco, L., and Ward, M. (1996). Temporal constraints on the McGurk effect. *Percept. Psychophys.* 58, 351–362. doi: 10.3758/BF03206811

Nakagawa, S., and Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods Ecol. Evol.* 4, 133–142. doi: 10.1111/j.2041-210x.2012.00261.x

Nath, A. R., and Beauchamp, M. S. (2012). A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *Neuroimage* 59, 781–787. doi: 10.1016/j.neuroimage.2011.07.024

Penfield, W., and Rasmussen, T. (1950). *The Cerebral Cortex of Man; A Clinical Study of Localization of Function*. New York, NY: Macmillan.

Purcell, D. W., and Munhall, K. G. (2006). Compensation following real-time manipulation of formants in isolated vowels. *J. Acoust. Soc. Am.* 119, 2288–2297. doi: 10.1121/1.2173514

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Randolph, M. (1989). *Syllable-Based Constraints on Properties of English Sounds*. Cambridge, MA: MIT. Unpublished Ph.D. thesis.

Reed, C. M., Durlach, N. I., Braida, L. D., and Schultz, M. C. (1989). Analytic study of the tadoma method: effects of hand position on segmental speech perception. *J. Speech Lang. Hear. Res.* 32, 921–929. doi: 10.1044/jshr.3204.921

Reisberg, D., McLean, J., and Goldfield, A. (1987). "Easy to hear but hard to understand: a lip-reading advantage with intact auditory stimuli," in *Hearing by Eye: The Psychology of Lip-Reading*, eds B. Dodd and R. Campbell (Hillsdale, NJ: Lawrence Erlbaum), 97–114.

Remez, R. E., Fellowes, J. M., Pisoni, D. B., Goh, W. D., and Rubin, P. E. (1998). Multimodal perceptual organization of speech: evidence from tone analogs of spoken utterances. *Speech Commun.* 26, 65–73.

Remez, R. E., Ferro, D. F., Dubowski, K. R., Meer, J., Broder, R. S., and Davids, M. L. (2010). Is desynchrony tolerance adaptable in the perceptual organization of speech? *Attent. Percept. Psychophys.* 72, 2054–2058. doi: 10.3758/BF03196682

Remez, R. E., Ferro, D. F., Wissig, S. C., and Landau, C. A. (2008). Asynchrony tolerance in the percpetual organization of speech. *Psychon. Bull. Rev.* 15, 861–865. doi: 10.3758/PBR.15.4.861

Rosen, S. M., and Howell, P. (1981). Plucks and bows are not categorically perceived. *Percept. Psychophys.* 30, 156–168. doi: 10.3758/BF03204474

Rosenblum, L. D. (2008a). Speech perception as a multimodal phenomenon. *Curr. Direct. Psychol. Sci.* 17, 405–409. doi: 10.1111/j.1467-8721.2008.00615.x

Rosenblum, L. D. (2008b). "Primacy of multimodal speech perception," in *The Handbook of Speech Perception*, eds D. B. Pisoni and R. E. Remez (Malden, MA: Blackwell), 51–78. doi: 10.1002/9780470757024.ch3

Rosenblum, L. D., and Gordon, M. S. (2001). The generality of specificity: some lessons from audiovisual speech. *Behav. Brain Sci.* 24, 239. doi: 10.1017/S0140525X01503945

Rosenblum, L. D., Dias, J. W., and Dorsi, J. (2017). The supramodal brain: implications for auditory perception. *J. Cogn. Psychol.* 29, 65–87. doi: 10.1080/20445911.2016.1181691

Rosenblum, L. D., Schmuckler, M. A., and Johnson, J. A. (1997). The McGurk effect in infants. *Percept. Psychophys.* 59, 347–357. doi: 10.3758/BF03211902

Shimojo, S., and Shams, L. (2001). Sensory modalities are not separate modalities: plasticity and interactions. *Curr. Opin. Neurobiol.* 11, 505–509. doi: 10.1016/S0959-4388(00)00241-5

Sparks, D. W., Kuhl, P. K., Edmonds, A. E., and Gray, G. P. (1978). Investigating the MESA (Multipoint Electrotactile Speech Aid): the transmission of segmental features of speech. *J. Acoust. Soc. Am.* 63, 246–257. doi: 10.1121/1.381720

Spence, C., and Bayne, T. (2015). "Is consciousness multisensory?," in *Perception and Its Modalities*, eds D. Stokes, M. Matthen, and S. Biggs (Oxford: Oxford University Press), 95–132. doi: 10.1093/acprof:oso/9780199832798.003.0005

Stein, B. E., and Stanford, T. R. (2008). Multisensory integration: current issues from the perspective of the single neuron. *Nat. Rev. Neurosci.* 9, 255–266. doi: 10.1038/nrn2331

Stein, B. E., Stanford, T. R., and Rowland, B. A. (2009). The neural basis of multisensory integration in the midbrain: its organization and maturation. *Hear. Res.* 258, 4–15. doi: 10.1016/j.heares.2009.03.012

Stetson, C., Cui, X., Montague, P. R., and Eagleman, D. M. (2006). Motor-sensory recalibration leads to an illusory reversal of action and sensation. *Neuron* 51, 651–659. doi: 10.1016/j.neuron.2006.08.006

Stevens, K. N. (1981). "Constraints imposed by the auditory system on the properties used to classify speech sounds: data from phonology, acoustics, and psychoacoustics, in *Advances in Psychology*, eds T. Myers, J. Laver, and J. Anderson (Amsterdam: North-Holland). doi: 10.1016/S0166-4115(08)60179-X

Stevens, K. N. (1989). On the quantal nature of speech. *J. Phonetics* 17, 3–45. doi: 10.1016/S0095-4470(19)31520-7

Stevens, K. N. (2000). *Acoustic Phonetics*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/1072.001.0001

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *J. Acoust. Soc. Am.* 111, 1872–1891. doi: 10.1121/1.1458026

Stoffregen, T. A., and Bardy, B. G. (2001). On specification and the senses. *Behav. Brain Sci.* 24, 195–213. doi: 10.1017/S0140525X01003946

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212–215. doi: 10.1121/1.1907309

Summerfield, Q. (1987). "Some preliminaries to a comprehensive account of audio-visual speech perception," in *Hearing by Eye: The Psychology of Lip-Reading*, eds B. Dodd and R. Campbell (Hillsdale, NJ: Lawrence Erlbaum), 3–52.

Van Wassenhove, V., Grant, K. W., and Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia* 45, 598–607. doi: 10.1016/j.neuropsychologia.2006.01.001

Vatikiotis-Bateson, E., Eigsti, I.-M., Yano, S., and Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Percept.Psychophys.* 60, 926–940. doi: 10.3758/BF03211929

Whalen, D. H. (1984). Subcategorical phonetic mismatches slow phonetic judgments. *Percept. Psychophys.* 35, 49–64. doi: 10.3758/BF03205924

Whalen, D., and Liberman, A. M. (1996). Limits on phonetic integration in duplex perception. *Percept. Psychophys.* 58, 857–870. doi: 10.3758/BF03205488

Xu, Y., Liberman, A. M., and Whalen, D. (1997). On the immediacy of phonetic perception. *Psychol. Sci.* 8, 358–362. doi: 10.1111/j.1467-9280.1997.tb00425.x

Yi, H. G., Leonard, M. K., and Chang, E. F. (2019). The encoding of speech sounds in the superior temporal gyrus. *Neuron* 102, 1096–1110. doi: 10.1016/j.neuron.2019.04.023

Yun, S., Choi, J.-Y., and Shattuck-Hufnagel, S. (2020). A landmark-cue-based approach to analyzing the acoustic realizations of American English intervocalic flaps. *J. Acoust. Soc. Am.* 147, EL471–EL477. doi: 10.1121/10.0001345

frontiers | Frontiers in Human Neuroscience

TYPE Brief Research Report
PUBLISHED 22 July 2022
DOI 10.3389/fnhum.2022.894676

Check for updates

OPEN ACCESS

EDITED BY
Xing Tian,
New York University Shanghai, China

REVIEWED BY
Akira Toyomura,
Gunma University, Japan
Andrew L. Bowers,
University of Arkansas, United States

*CORRESPONDENCE
Emily O. Garnett
emilyog@umich.edu

SPECIALTY SECTION
This article was submitted to Speech and Language, a section of the journal Frontiers in Human Neuroscience

RECEIVED 12 March 2022
ACCEPTED 27 June 2022
PUBLISHED 22 July 2022

CITATION
Garnett EO, Chow HM, Limb S, Liu Y and Chang S-E (2022) Neural activity during solo and choral reading: A functional magnetic resonance imaging study of overt continuous speech production in adults who stutter.
Front. Hum. Neurosci. 16:894676.
doi: 10.3389/fnhum.2022.894676

COPYRIGHT
© 2022 Garnett, Chow, Limb, Liu and Chang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Neural activity during solo and choral reading: A functional magnetic resonance imaging study of overt continuous speech production in adults who stutter

Emily O. Garnett[1]*, Ho Ming Chow[1,2], Sarah Limb[1], Yanni Liu[1] and Soo-Eun Chang[1]

[1]Michigan Medicine, Department of Psychiatry, University of Michigan, Ann Arbor, MI, United States, [2]Department of Communication Sciences and Disorders, University of Delaware, Newark, DE, United States

Previous neuroimaging investigations of overt speech production in adults who stutter (AWS) found increased motor and decreased auditory activity compared to controls. Activity in the auditory cortex is heightened, however, under fluency-inducing conditions in which AWS temporarily become fluent while synchronizing their speech with an external rhythm, such as a metronome or another speaker. These findings suggest that stuttering is associated with disrupted auditory motor integration. Technical challenges in acquiring neuroimaging data during continuous overt speech production have limited experimental paradigms to short or covert speech tasks. Such paradigms are not ideal, as stuttering primarily occurs during longer speaking tasks. To address this gap, we used a validated spatial ICA technique designed to address speech movement artifacts during functional magnetic resonance imaging (fMRI) scanning. We compared brain activity and functional connectivity of the left auditory cortex during continuous speech production in two conditions: solo (stutter-prone) and choral (fluency-inducing) reading tasks. Overall, brain activity differences in AWS relative to controls in the two conditions were similar, showing expected patterns of hyperactivity in premotor/motor regions but underactivity in auditory regions. Functional connectivity of the left auditory cortex (STG) showed that within the AWS group there was increased correlated activity with the right insula and inferior frontal area during choral speech. The AWS also exhibited heightened connectivity between left STG and key regions of the default mode network (DMN) during solo speech. These findings indicate possible interference by the DMN during natural, stuttering-prone speech in AWS, and that enhanced coordination between auditory and motor regions may support fluent speech.

KEYWORDS

stuttering, fMRI, functional connectivity, speech fluency, continuous speech, default mode network, auditory motor integration

Introduction

Robust connectivity and interactions among cortical auditory and speech-motor brain areas provide the basis for speech production. Auditory-motor integration is crucial for fluent speech, which is disrupted in disorders such as developmental stuttering. Stuttering affects 1% of the population and manifests as frequent involuntary interruptions in the speech flow such as repetitions (i.e., sound/syllable repetitions) and dysrhythmic phonations (i.e., blocks and prolongation of sound/syllables). Decades of behavioral and neuroimaging research have offered accounts of inefficient or disrupted auditory motor integration in adults who stutter (AWS). Key findings come from several lines of research that center on the application or modulation of sensory input during speech production, including reduced motor adaptation in response to auditory perturbations (Cai et al., 2012, 2014; Daliri et al., 2018; Daliri and Max, 2018) and near-elimination of speech disfluencies under rhythmic pacing or delayed auditory feedback conditions (Barber, 1940; Azrin et al., 1968; Hutchinson and Norris, 1977; Stager et al., 1997; Toyomura et al., 2011). Further evidence for disrupted auditory motor integration comes from studies using neuroimaging methods such as functional magnetic resonance imaging (fMRI) that show aberrant brain activity and/or connectivity among speech and auditory brain areas (for review see Chang et al., 2019). In particular, findings from the neuroimaging literature include *overactivation* of cortical speech motor regions, particularly in the right hemisphere, but *decreased* activation in auditory regions, during speech production, both of which become attenuated under fluency induced conditions or following intensive fluency training (Braun, 1997; Stager et al., 2003; Brown et al., 2005; Toyomura et al., 2011, 2015; Budde et al., 2014). Together, these findings suggest that disruption in auditory-motor integration may have a negative impact on generating fluent speech.

The auditory system plays a crucial role in speech production (Guenther and Hickok, 2015) reflected in traditional and as well as newer neurocomputational models. For example, according to the dual route speech processing model (Hickok, 2012), the ventral and dorsal streams work in parallel to integrate sound into both meaning and action, respectively, utilizing the very nature of a speech target, the auditory signal, as a corrective speech output tool. The Directions into Velocities of Articulators [(DIVA); (Bohland and Guenther, 2006; Tourville and Guenther, 2011] model consists of a feedforward control system, which generates already established motor commands guiding speech production, and a feedback control system that provides online detection of production errors by comparing the incoming auditory signal to the expected auditory signal. These auditory targets are represented in the auditory state map in posterior auditory cortex. The superior temporal gyrus (STG) is part of this auditory feedback control system that through an inverse mapping process transforms the auditory target to motor commands in the motor areas using robust structural connections to the ventral motor cortex *via* the feedback control map in right ventral premotor cortex. Importantly, as these systems develop and become established, speech production is achieved primarily through a strong feedforward system, with less of a role of the auditory cortex and feedback control (Bohland and Guenther, 2006; Tourville and Guenther, 2011; Kearney and Guenther, 2019).

Indeed, several theoretical perspectives on stuttering have hypothesized that stuttering is due to an over-reliance on auditory feedback (Max et al., 2004; Civier et al., 2010, 2013), particularly as a result of impaired feedforward control mechanisms, although others propose that the issue arises in the feedback control system itself (e.g., (Max and Daliri, 2019). Recently, using the GODIVA model, Chang and Guenther (2020) proposed that the key impairment underlying stuttering is in the feedforward system, specifically in the cortico-basal ganglia loop associated with initiating speech motor programs, similarly proposed also by others in the field (Maguire et al., 2002, 2004; Alm, 2004; Chang and Zhu, 2013; Civier et al., 2013). Importantly, the auditory system can influence motor behavior specifically through these *corticostriatal* projections (Znamenskiy and Zador, 2013). As noted in Chang and Guenther (2020), auditory feedback of self-generated speech may not match the target auditory pattern for a speech sound due, for example, to minor articulation errors. In this case, this mismatch between expected and actual sensorimotor context may impair crucial initiation commands by the basal ganglia, leading to stuttering. In this context, inhibiting auditory feedback of one's own speech to avoid detection of minor errors during production may help reduce the mismatch and allow the basal ganglia to generate initiation signal to allow fluent speech. Such an account is consistent with one of the most commonly reported neuroimaging findings in AWS, namely decreased activation in auditory regions during speech tasks. Conversely, auditory activity in AWS becomes comparable or even exceeds levels observed in non-stuttering adults under (or after) fluency inducing conditions such as choral speech or fluency shaping (Fox et al., 1996; Ingham, 2003; Stager et al., 2003; De Nil et al., 2008; Toyomura et al., 2011). Such observations suggest that studies delving further into the mechanisms by which fluency-inducing conditions modulate brain activity in speech-motor and auditory regions may lead not only to a better general understanding of the biological basis of stuttering but may also inform current treatment strategies.

Fluency inducing conditions include speaking in unison with another person, metronome-timed speech, singing, masking, or listening to transformed sensory (auditory) feedback of one's own voice (Andrews et al., 1982; Bloodstein and Ratner, 2008; Frankford et al., 2021). Such techniques have several factors in common. First, the effects are robust but temporary (Kalinowski and Saltuklaroglu, 2003). Second,

they typically involve an external pacing component. In choral speech, this is represented by the other speaker's reading pattern and pace. The person who stutters then speaks in unison with the other speaker, which drastically reduces their stuttering. In paced speech the external component is represented by a metronome, for example, and the person who stutters matches the timing of their own speech (typically at the syllable or word level) to the beat of the metronome, again resulting in perceptually fluent speech. One proposed account for this "rhythm effect" (Barber, 1940; Azrin et al., 1968; Stager et al., 1997; Toyomura et al., 2011, 2015; Davidow, 2014; Frankford et al., 2021) is that stuttering stems from an inefficient or disrupted *internal* timing mechanism, whereby the addition of an external rhythm allows speech production to bypass the faulty internal mechanism and proceed using the external pace (Alm, 2004; Etchell et al., 2014). Under these external pacing conditions speech production proceeds fluently, resulting from better matching between expected and actual incoming sensory input. With the improved speech timing, the feedforward control mechanism can guide speech production, rather than over-relying on the feedback control system (Civier et al., 2013).

Third, fluency inducing conditions seem to reduce brain activity differences observed during stutter-prone speech. Namely, speaking under conditions that involve external pacing results in increased left frontotemporal activation, and reduced motor hyperactivity, including in the right frontal opercular areas (De Nil et al., 2003; Neumann et al., 2005; Giraud, 2008; Kell et al., 2009; Toyomura et al., 2011, 2015). In particular, STG consistently shows increased activity under fluency inducing conditions suggesting that this region plays an integral role in facilitating fluent speech in people who stutter.

One critical limitation of the aforementioned research findings is that the studies primarily used covert speech, single words, and/or short phrases as speech production tasks while capturing functional brain activity. The use of truncated speech, often incorporating sparse scanning paradigms, were required due to the motion artifacts associated with continuous speech that severely affects the fMRI signal. These paradigms are limited because stuttering typically does not occur on single words; sentence-level or longer utterances are needed to capture brain activity patterns that differentiate stutterers from non-stutterers. Moreover, single word tasks may not fully elucidate brain areas involved in fluency inducing conditions.

To address this gap in the field, we used a validated fMRI artifact removal technique designed specifically for continuous speech production studies to explore brain activity in AWS during continuous natural speech and under fluency inducing conditions. This technique effectively removes speech-related movement artifacts in fMRI data, allowing us to capture brain activity patterns during overt, continuous speech production (AbdulSabur et al., 2014; Xu et al., 2014). In this study, we examined brain activity during choral reading (fluency-inducing) and solo reading (prone to disfluencies) in AWS. Among the potential fluency-inducing conditions, we chose to use choral speech for the following reasons. First, past neuroimaging investigations examining brain activity differences during fluent and induced fluent speech had primarily involved reading and choral speech conditions (Fox et al., 1996, Braun et al., 1997; Fox et al., 2000; Ingham et al., 2012). One reason for this is that metronome or other similarly paced conditions could lead to unnatural sounding speech. Second, we were concerned about the possible interaction between the regular pulse sounds of the scanner and the rhythmic sounds of the metronome. Third, designing a condition that controlled for the auditory feedback of the metronome to be applied during the solo condition was also challenging. Finally, our primary aim was to examine how brain activity patterns and functional connectivity of the auditory cortex differ between an induced fluency condition that involves external rhythmic stimuli (choral reading) and a condition that relies on the speaker's internal timing ability (solo reading).

Guided by previous findings, we hypothesized that relative to controls, fluency-induced speech in AWS would be associated with increased activity in the auditory regions including posterior STG, and reduced hyperactivity in motor cortical regions including the IFG, premotor, and motor cortical areas. We further hypothesized that compared to natural speech, fluency-induced speech would be associated with greater functional connectivity between left STG and speech motor areas in AWS. Although we primarily focused on the neurophysiological effects of choral reading, we also examined the behavioral effects, namely the effectiveness of choral reading in reducing stuttering. We therefore expected that choral reading would lead to a greater reduction in amount of stuttering compared to solo reading; however, it is likely that stuttering will occur only rarely in either reading condition due to the masking effects of the scanner noise. Such effects, however, are constant across the solo and choral conditions, and would not preclude investigation of the primary research question, which was to examine the effects of rhythmic pacing that would be provided by the choral and not the solo condition.

Materials and methods

Participants

Thirty-one adults participated in this study, 15 AWS (4F) and 16 adults (4F) who did not stutter (controls). Detailed demographic information can be found in **Supplementary Table 1**. All participants were native English speakers who reported no speech, language, hearing, cognitive, or psychiatric disorders, other than stuttering for the AWS group. Groups did not differ significantly in age or expressive or receptive language. The AWS group reported slightly higher years of education (M = 14.8) than the control group

(M = 13.43; p = 0.04). Stuttering severity was obtained by certified speech-language pathologists (SLPs) using the Stuttering Severity Instrument (SSI-4; Riley, 2009), and ranged from very mild to very severe based on SSI-4 composite scores. The protocol was approved by the Institutional Review Boards of the University of Michigan Medical School. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

Study design

Participants were recruited as part of a larger within-subject double-blind study that investigated the effects of transcranial direct current stimulation (tDCS) paired with speech fluency training on brain activity (Garnett et al., 2019). In the present study, only the fMRI scans acquired before stimulation were analyzed, to eliminate possible influences of tDCS or speech fluency training.

Stimuli and task

Stimuli were recordings of short paragraphs spoken by a native English male speaker in a neutral tone at approximately 155 words per minute (Chow et al., 2014, 2015). The passages were assessed to be at a 7th grade Flesch-Kincaid Grade Level. The current study used 16 unique paragraphs, each 30 s in length. Additionally, a second recording was created in which each paragraph was played backwards to be used during solo reading (see below). Therefore, there were 32 paragraphs in total (16 forward, 16 backward). These were divided into 4 unique sets to correspond to the 4 runs of the fMRI experiment. Each run began with a 14-s fixation cross, followed by eight 30-s trials, one paragraph per trial. Within each run, participants read 4 paragraphs under "solo reading" and the same 4 paragraphs under "choral reading" conditions, in an alternating fashion. During choral reading, participants read the paragraph shown on the screen while matching their reading pace with an audio recording of the same passage presented *via* MRI-compatible earphones. During solo reading, participants read the passage naturally while a recording of the passage was played backwards via the earphones. The reversed speech recording was used to match the level of external auditory feedback delivered between the solo and chorus conditions, while not inducing speech pacing for the solo condition.

Each trial consisted of a brief instruction screen lasting 3 s that indicated if the subject should "read alone" (solo reading) or "read together" (choral reading) when the next paragraph appeared on the screen. Following the instructions, there was a 3 s fixation cross, after which the paragraph appeared on the screen for 30 s. There was 16 s of fixation cross at the end of the run. Each run lasted approximately 7 min. See **Supplementary Figure 1** for an example trial. Prior to the fMRI session, participants completed practice trials with corrective feedback. Participants wore over-the-ear headphones as well as ear plugs during scanning. Additionally, a noise canceling microphone was placed close to the mouth to capture participants' speech, and a flexible camera was placed over the participants' mouth to separately capture video during speech.

Functional magnetic resonance imaging parameters, processing, and analysis

Functional magnetic resonance imaging acquisition

The fMRI data were acquired using a 3T GE MRI scanner (MR 750). A standard echoplanar (EPI) pulse sequence was used, with the following parameters: repetition time (TR) = 2 s; echo time (TE) = 30 ms, flip angle = 90°, in-plane resolution = 3.4 × 3.4 mm; 37 interleaved sagittal slices; slice thickness = 4 mm, acceleration factor = 2. In addition, high-resolution structural images were acquired at the beginning of each scanning session using spoiled gradient-recalled acquisition in steady state (SPGR) imaging (TR = 12.236 ms, TE = 5.184 ms, flip angle = 15°, resolution = 1 × 1 × 1mm).

Denoising speech-related movement artifacts

SPM12[1] was used for fMRI data preprocessing and statistical analysis unless specified otherwise. For each participant, functional images were corrected for differences in slice acquisition timings. Anatomical scans and functional volumes were co-registered to the first volume of the first scan using rigid body rotation. Functional scans were concatenated and de-noised using a strategy detailed in our previous publication (Xu et al., 2014). This fMRI denoising technique uses spatial independent component analysis (sICA) to decompose the functional images into a number of independent components and automatically identify and remove noise components based on their spatial patterns (Xu et al., 2014). This technique has been validated using positron emission tomography (PET) and has been demonstrated to be able to remove fMRI artifacts associated with continuous speech production (AbdulSabur et al., 2014; Xu et al., 2014). Because PET is less susceptible to motion artifacts, it is considered the "gold standard" for studying the neural processing of speech production. Xu et al. (2014) study showed that sICA denoising method can effectively remove artifacts associated with speech production and that the results of de-noised fMRI and PET were comparable. De-noised functional scans were normalized to MNI space using DARTEL and spatially smoothed with a 6 mm FWHM kernel (Ashburner, 2007).

Task-based functional magnetic resonance imaging

We performed two separate task-based fMRI analyses. First, we compared brain activity *between groups* in each reading

1 https://www.fil.ion.ucl.ac.uk/spm/software/spm12/

condition by examining the contrast *AWS - Control* in the solo reading condition (Section "Brain activity in adults who stutter compared to controls during solo reading") and the same contrast separately in the choral reading condition (Section "Brain activity in adults who stutter compared to controls during choral reading"). This contrast provides information about potential between-group differences in each of the two reading conditions. Here we expected to observe significant differences between AWS and controls during solo reading, but more similar patterns of activity during choral reading.

Second, we compared brain activity between conditions *within each group* by examining the contrast *Choral – Solo* within the AWS group (Section "Brain activity during choral compared to solo reading in adults who stutter") and separately within the control group (Section "Brain activity during choral compared to solo reading in controls"). This contrast provides information about how brain activity differs between choral and solo reading in each participant group. For example, one might expect minimal or no differences between choral and solo reading in the control group because there will be no induced fluency effect in this group (as they do not stutter). Conversely, one might expect to see more differences between choral and solo reading in the AWS group, however, given the strong fluency-inducing effect of choral reading on stuttering frequency.

Each participant's preprocessed data was analyzed using a general linear model (GLM) implemented in SPM12. Reading conditions (choral and solo) were modeled with separate regressors. Individual beta estimates were entered into group level analysis. Statistical threshold was set at voxel-wise $p = 0.001$ and cluster size of 19, corresponding to $p < 0.05$ corrected, using AFNI 3dClustSim (version 17.2.13) with non-Gaussian auto-correlation function (-acf option) (Cox et al., 2017).

Task-based functional connectivity

Previous research has reported induced fluency (rhythmic) conditions are associated with heightened STG activity particularly in the left hemisphere in AWS (Salmelin et al., 1998; Stager et al., 2003; Toyomura et al., 2011). Using a seed-based functional connectivity analysis, we asked whether induced fluency during choral reading (relative to solo reading) was associated with increased functional connectivity between left STG and speech motor areas. For this analysis, we selected a left STG seed region with peak coordinates of −50, −34, 18 as reported in Toyomura et al. (2011). In that study, left STG was identified as the key region that showed increased activity in AWS during fluency induced conditions (exceeding activity levels seen in controls) but significantly reduced activity during solo reading. In this analysis we examined areas across the whole brain that showed significantly different functional connectivity the left STG seed region (Section "Functional connectivity").

For each subject, functional images corresponding to each reading condition were separated and concatenated. Time-series for each condition were band-passed filtered with cutoff frequency at 0.03 to 0.2 Hz. Time-series of the seed region was extracted by averaging voxels in a sphere of 5 mm radius at the coordinates. Pearson's correlation coefficients were calculated between the time-series of the seed region and the time series of each voxel in the whole brain and converted to Fisher's z-scores. The individual maps were analyzed using GLM. Using AFNI 3dClustSim, a voxel wise height threshold of $p = 0.01$ and a cluster size of 70 was considered significant, corresponding to a corrected $p < 0.05$.

Speech rate, loudness, and stuttering frequency during scanning

Speech rate in syllables per second (SPS) was calculated by dividing the number of syllables by total speaking time. Speech rate and %SS were calculated separately for solo and choral reading passages. To assess any differences in loudness between choral and solo reading, a trained study team member blinded to condition and study objectives listened to each passage and rated it from 1 (quietest) to 5 (loudest) separately for each subject. That is, loudness ratings were completed within each subject rather than across subjects, as speaking volume naturally varied across participants. The study team member was blind as to the type of reading passage (solo or choral). This analysis was completed for AWS and control groups.

This analysis was completed for both AWS and control groups. A certified SLP with expertise in stuttering and disfluency analysis listened to the recordings of each passage for each AWS participant to determine stuttering frequency. The SLP was blinded to the condition (choral or solo reading). The total number of syllables was noted and marked for the presence or absence of stuttering, defined as dysrhythmic phonations (prolongations, blocks) and whole word or part-word repetitions. Percent stuttered syllables (%SS) was calculated for the AWS group only, as no participants in the control group stuttered.

Results

Speech rate, loudness, and stuttering frequency

Supplementary Table 2 shows between group differences in speech rate and loudness during solo and choral reading. Loudness ratings for two controls and two AWS, speech rate for one control and one AWS, and disfluency rates for one AWS were unable to be calculated due to poor audio recording quality. There were no significant differences between the AWS and control groups in syllables per second (SPS) or loudness in either reading condition (all p values > 0.078; see **Supplementary Material** for details).

Within-group comparisons showed no significant differences in speech rate in solo (M = 3.60) compared to choral reading (M = 3.64) in the AWS group (p = 0.777). The control group on average spoke significantly faster during solo reading (M = 3.85) than choral reading (M = 3.68, $p < 0.001$; **Supplementary Table 3**). Both groups spoke slightly but significantly louder in solo reading compared to choral reading (**Supplementary Table 3**).

In the AWS group, stuttering frequency as measured by percent stuttered syllables (%SS) was comparable in the choral (M = 1.36%) and solo (M = 1.8%) reading conditions (p = 0.777; **Supplementary Table 4**). However, closer inspection of the individual subjects showed that the subject with the highest SSI score showed a dramatic decrease in %SS during choral reading but maintained a high rate of stuttering during the solo condition. Consequently, we repeated this analysis after excluding this subject. Results showed that %SS was significantly greater in the choral reading condition (M = 1.42%) than the solo reading condition (M = 0.42%; $p < 0.001$). Given that 3% is an often-used threshold to determine stuttering status, the %SS in both conditions was well below this number. The effect of the scanner noise during speech is likely to have had a strong influence on induced fluency. Therefore, the %SS difference between the two conditions is not considered to be meaningful.

Task-based activation

Brain activity in adults who stutter compared to controls during solo reading

We first compared brain activity between groups during solo reading (**Table 1** and **Figure 1A**). Compared to controls, AWS exhibited heightened activity in right precentral gyrus, bilateral supplementary motor area (SMA), and left middle temporal gyrus (MTG). In contrast, AWS exhibited decreased activity in left cerebellum, occipital/lingual gyri, right cuneus, and bilateral STG.

Brain activity in adults who stutter compared to controls during choral reading

During choral reading AWS exhibited heightened activity compared to controls in right precentral gyrus, as well as left middle temporal, SMA, and STG/Insular gyri (**Table 1** and **Figure 1B**). Decreased activity for AWS relative to controls was found in left hemisphere lingual, middle occipital, and STG, as well as right cuneus, cerebellum, and posterior cingulate.

Brain activity during choral compared to solo reading in adults who stutter

The AWS group exhibited heightened activity for choral compared to solo reading in left angular gyrus, middle frontal gyrus (MFG), right STG/SMG in the area of SPT, right middle cingulate, and bilateral superior frontal gyrus (SFG). Decreased activity during choral relative to solo reading was found in right precuneus, cingulate gyrus, MFG/SFG, and insula, left IFG, and the cerebellar declive. See **Supplementary Table 5** (top panel) for details of cluster sizes, coordinates, and test statistics, and **Supplementary Figure 2A** for activity patterns.

Brain activity during choral compared to solo reading in controls

Controls exhibited heightened activity during choral relative to solo reading in left anterior cingulate and cerebellar crus I, right middle cingulate (extends into L), and bilateral angular gyrus and MFG. Controls exhibited decreased activity during choral relative to solo reading in left cerebellar crus II, STG, precuneus, and insula extending into the caudate, right SFG/MFG, IFG/insula, SMA, and MFG, and bilateral superior parietal lobe. See **Supplementary Table 5** (bottom panel) for details of cluster sizes, coordinates, and test statistics, and **Supplementary Figure 2B** for activity patterns.

Functional connectivity

Functional connectivity analyses were conducted within each group for choral versus solo reading. Results showed that AWS exhibited increased connectivity between left STG and right insula in the IFG (and left IFG detected sub-threshold; **Figure 2**) during choral versus solo reading condition. On the other hand, functional connectivity of left STG was significantly increased in the bilateral angular gyri and precuneus for AWS during the solo condition relative to the choral reading condition (**Table 2** and **Figure 2**). In the control group, functional connectivity of the left STG did not differ significantly between the two speech conditions (not shown).

Discussion

A major aim of this study was to investigate how brain activity patterns in the auditory cortex during continuous speech production differ between AWS and controls. Overall, group differences in brain activity patterns observed in each condition were largely similar, showing the expected pattern in AWS of heightened activity in motor areas (right hemisphere premotor cortex and SMA) but decreased activity in auditory regions previously reported as neural signatures associated with stuttering (Brown et al., 2005; Belyk et al., 2015; Neef et al., 2015). In this way, the current results partially support our hypothesis that choral reading would attenuate the aberrant motor and auditory activity during speech in AWS relative to controls; however, these activity pattern differences were subtle. For example, compared to controls, AWS exhibited heightened bilateral SMA during solo reading, but only left SMA showed this pattern during choral reading. Additionally, solo reading was associated with decreased activity in bilateral STG, yet during choral reading

TABLE 1 Group differences in solo reading (left panel) and choral reading (right panel).

Solo reading						Choral reading					
Region	**x**	**y**	**z**	***t***	**Voxels**	**Region**	**x**	**y**	**z**	***t***	**Voxels**
AWS > Controls						***AWS > Controls***					
Precentral (R)	54	3	36	5.02	37	Precentral (R)	54	3	36	5.59	42
SMA (L)	−3	18	48	4.89	31	MTG (L)	−51	−63	3	5.66	35
SMA (R)	9	18	63	4.77	24	SMA (L)	−3	18	48	4.68	30
MTG (L)	−51	−63	3	5.02	21	STG/INS (L)	−45	−36	15	5.37	21
AWS < Controls						***AWS < Controls***					
Lingual (L)	−27	−93	−18	−6.75	119	Lingual (L)	−27	−93	−18	−7.36	192
Cuneus (R)	15	−93	0	−5.52	50	Cuneus (R)	15	−93	0	−5.72	48
STG/HG (L)	−45	−21	3	−6.02	36	STG/HG (L)	−45	−21	3	−5.76	31
Cerebellar Lobules I-IV (L)	−6	−51	−6	−4.9	34	MOC (L)	−24	−93	12	−4.61	24
MOC (L)	−24	−93	12	−4.75	34	Cerebellum (Crus I) (R)	27	−87	−18	−4.97	23
STG/HG (R)	51	−15	3	−4.38	25	Posterior cingulate/cuneus (R)	9	−72	6	−4.26	19

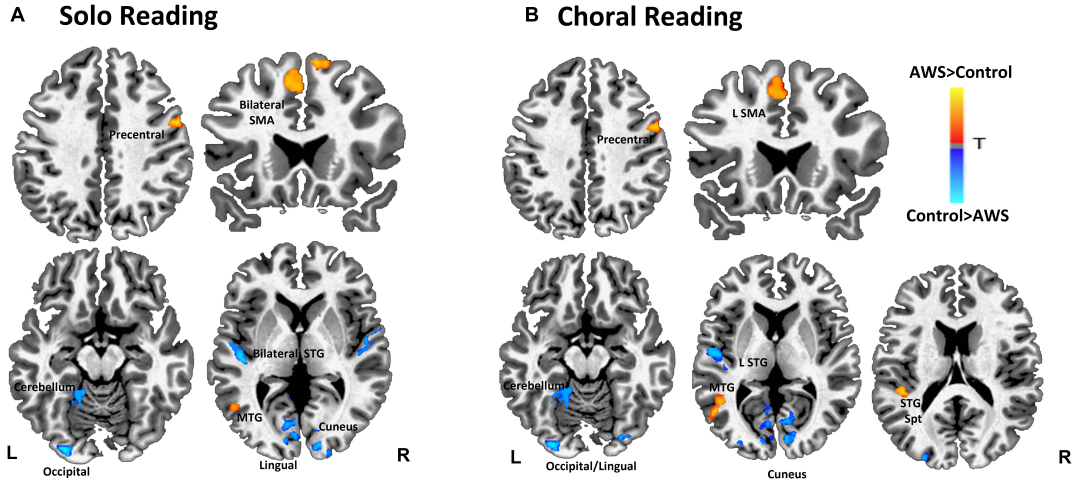


FIGURE 1
Contrast between adults who stutter (AWS) and control groups during solo reading **(A)** and choral reading **(B)**. Warmer colors represent with significantly greater activity for AWS compared to controls. Statistical map has a threshold at $p < 0.05$ (corrected).

only left STG showed decreased auditory activity in AWS relative to controls.

When directly comparing choral and solo reading in AWS and controls separately, the induced fluency condition in AWS was associated with a pattern of greater brain activity in areas in bilateral STG/SMG, angular gyrus, and MFG, regions linked to the executive control network. The solo condition was associated with greater brain activity in areas linked to the cingulo-opercular network in AWS, which supports maintaining sustained goal oriented cognitive control. In contrast, controls exhibited greater activity in anterior and middle cingulate in the choral relative to the solo condition. Choral reading was also associated with decreased activity in bilateral parietal regions as well as several important motor speech regions including SMA and the IFG/insular area.

An interesting convergent finding in the task-based fMRI activity was that induced fluency seemed to be associated with increased activity in a cluster in posterior STG bordering the temporal parietal junction showed increased activity for AWS relative to controls. This cluster (using both peak [−45, −36, 15] and center of mass [−43.0, −37.3, 15.1] coordinates) falls within the range of coordinates for area Spt (Buchsbaum et al., 2001, 2005; Hickok et al., 2003, 2009, 2011) arguably a crucial region for auditory motor integration. In their SFC model, Hickok et al. (2011) refer to stuttering as reflecting "noisy" mapping between auditory and motor systems resulting in inaccurate predictions and, consequently, inaccurate corrective commands [see also (Max et al., 2004)]. Moreover, a significant cluster in the vicinity of Spt in the *right* hemisphere was also significantly increased during choral relative to solo speech in the AWS

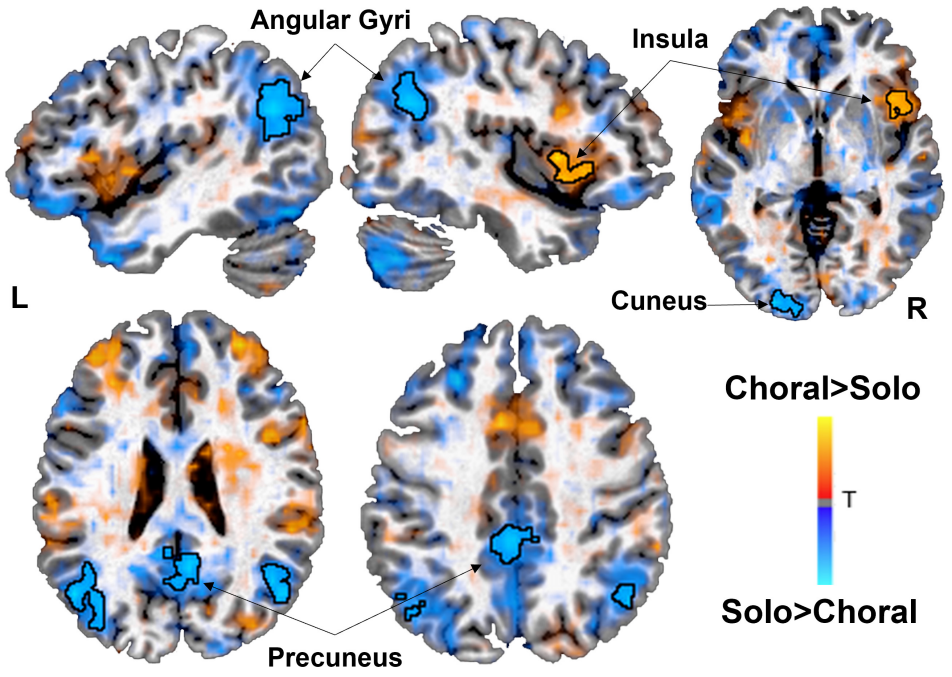


FIGURE 2
Brain areas showing significant functional connectivity with the left superior temporal gyrus (STG) seed during choral (warm) vs. solo (cold) reading in adults who stutter (AWS). During choral (fluency induced) reading, the left auditory region showed increased correlated activity patterns with the right insula/inferior frontal gyrus. During solo (stuttering prone) reading, the left auditory region increased functional connectivity with the bilateral angular gyri, cuneus, and precuneus, areas that are part of the default mode network. There were no significant functional connectivity findings in the control group. Statistical map has a threshold at $p < 0.05$ (corrected).

group. Thus, heightened activity in left Spt in AWS relative to controls during choral reading may indicate improved mapping between auditory and motor systems.

Though the task-based activity contrast analyses did not reveal substantial differences, a clear difference emerged between choral and solo reading based on the functional connectivity analysis. Here we examined brain areas showing significantly correlated activity with that of a left STG region previously linked to heightened activity in AWS during a rhythmic speaking task (Toyomura et al., 2011). A novel finding was that during solo speech in AWS, there was heightened connectivity between left STG and key regions of the default mode network (DMN), including bilateral angular gyri and precuneus. Defined based on its correlated activity at rest, DMN is often associated with mind wandering, prospection, theory of mind, and autobiographical memory (Buckner et al., 2008; Spreng et al., 2009). The DMN shows anti-correlations with task-positive networks such as those supporting attention, executive control, and somatomotor functions (Lee et al., 2012). It has also been suggested that performing fluid, automatic motor tasks characteristic of well-learned and skilled movements can break down when attention is focused inwardly to oneself (linked to DMN) versus outwardly toward a movement target (linked to motor networks) (Wulf and Lewthwaite, 2016). Accordingly, efficiently switching from DMN to sensorimotor networks might be expected to support fluent speech production. The significantly increased functional connectivity between DMN-linked regions and left STG that was only present in AWS during solo reading may indicate possible interference of the DMN during natural, continuous speech in AWS (Sonuga-Barke and Castellanos, 2007).

This notion is supported by our previous work showing that children who stutter exhibit aberrant connectivity between DMN and speech and attention networks, and in particular that anomalous connectivity involving DMN predicted persistent

TABLE 2 Regions showing significant functional connectivity in AWS using an *a priori* determined left superior temporal gyrus (LSTG) seed (Toyomura et al., 2011).

Region	x	y	z	*t*	Voxels
Choral* > *Solo					
Insula (R)	42	6	3	4.7	78
Solo* > *Choral					
Precuneus	6	−54	18	−4.3	197
Angular Gyrus (L)	−48	−68	27	−4.4	187
Angular Gyrus (R)	51	−66	24	−4.2	108
Cuneus/middle occ. (L)	−12	−99	0	−4.3	72

stuttering (Chang et al., 2018). Specifically, in that study, connectivity between the somatomotor network (SMN) and the DMN was one of the inter-network connectivity differences that predicted stuttering status. In particular, STG within the SMN showed heightened connectivity with a number of DMN nodes. The SMN on the other hand also showed aberrant connectivity with the attention networks (dorsal and ventral attention networks). Persistence in stuttering was found to be predicted primarily through intra- and inter-network connectivity involving the DMN and its connections to attention and executive control networks. In the present study, the AWS group is by definition a group of adults with persistent stuttering. Interference from the DMN has also been implicated in other neurodevelopmental disorders besides stuttering. For example, in adults with ADHD, hyperactivity of DMN has been shown regardless of task (Cortese et al., 2012), supporting the default mode interference hypothesis (Sonuga-Barke and Castellanos, 2007). For stuttering, hyperconnectivity between the DMN and SMN may reflect heightened internal focus on one's speech that leads to de-automatized speech patterns that are prone to breakdown. Well-learned motor tasks are performed optimally when focus is on the movement goal (externally focused attention), rather than when excessive inward attention is paid to one's articulators (which can lead to movement breakdown, and "choking" as documented in athletes under pressure). Supporting this notion, some past reports have shown that stuttering could be reduced in dual task conditions where working memory and attention were manipulated during speaking tasks (Eichorn et al., 2016, 2019). Such dual tasking effects on speech were present regardless of working memory load, suggesting that a general attention allocation away from speaking might be sufficient to increase fluency in speakers who stutter. This may mean that if stuttering speakers can better disengage their somatomotor networks from DMN, better fluency might be achieved. Because the present study did not systematically examine inter-network connectivity between DMN and task positive networks including SMN, however, these interpretations in the context of the present results are speculative and will need to be confirmed in future studies.

During choral relative to solo speech, AWS exhibited increased functional connectivity between left STG and right insula extending into IFG. This finding is partially in line with a recent study investigating the effects of an intensive fluency shaping treatment program on neurofunctional reorganization (Korzeczek et al., 2021). In that study, the intervention strengthened connectivity involving *a priori* defined hubs with a sensorimotor integration network, in particular between left IFG and right pSTG. Right frontal areas have also been associated with feedback control in the DIVA model: if there is a mismatch between expected and actual sensory feedback, the feedback control map in the right frontal/ventral premotor cortex issues an error signal. During auditory and somatosensory perturbation experiments (Tourville et al., 2008; Golfinopoulos et al., 2011), compensation for the perturbations was associated with an increase in right lateralized frontal activity. Therefore, it is possible that corrective actions to motor plans, which can be found during compensatory movements during perturbation and during induced fluency conditions like choral speech, is reflected by increased communication between temporal and frontal regions. More research is needed to examine specific roles of bilateral IFG and STG in stuttering, their functional connectivity with other regions during normal and induced fluency conditions, and how these change as a result of treatment or natural recovery.

Turning briefly to the speech patterns exhibited by AWS during scanning, the results were not completely in line with our hypotheses. We expected the choral reading condition to significantly decrease stuttering to a greater degree than the solo reading condition, but we found the opposite pattern. Importantly, however, *both reading conditions* showed very little stuttering, less than 2%. One potential explanation is that at times, the process of attempting to speak in unison with the recording during choral reading resulted in speech "adjustments" such as slowing a specific sound in order to stay in pace with the recording. Although we did not calculate %SS for the control group as a whole, we tested this hypothesis by having a study team member blinded to group assignment listen and calculate %SS for three control subjects. A similar phenomenon was observed in these three control subjects, none of whom stuttered. Therefore, we speculate that the apparent higher %SS in the choral reading condition was an artifact of attempts at pacing with the audio recording.

Both AWS and control groups spoke louder during solo compared to choral reading. This finding is potentially consistent with the Lombard effect, an innate tendency to speak louder in noisy environments (Lombard, 1911). However, such an account is not so straightforward, given that the overall auditory environment (i.e., scanner noise, bone-conduction, presence and loudness of auditory feedback in the headphones) was comparable in both solo and choral reading conditions. Although we cannot rule out that participants expended greater speech effort during the solo condition in an attempt to hear their own voice more clearly, it does not appear to differ between AWS and controls in this study. It is also possible that the increased loudness during solo reading reflects attempts to "ignore" the reversed speech being played. While this also cannot be ruled out, participants were specifically instructed to speak at approximately the same pace during solo reading as they did during the choral reading condition, so as to remain engaged in overt speech for the same amount of time (i.e., for the 30 s that the text appeared on the screen). In this way, they could not simply tune out the reversed speech or their speech rate would have differed wildly between conditions, as they would likely have reverted to speaking at their natural rate. When examining the speech rate in syllables per second (SPS), while the controls spoke somewhat faster during the solo condition compared to the choral condition, the AWS did not, nor were

there significant differences between groups in SPS in either solo or choral reading.

Limitations

This is the first report comparing brain activity during continuous solo and choral reading in AWS captured with fMRI and using advanced de-noising techniques. Despite some strengths, several important limitations exist. Our sample size was modest and may have contributed to observing overall similar activation between reading conditions, which was seen even at the individual subject level. On the other hand, the differences observed were in line with expectations of reduced motor hyperactivity and increased auditory activity. Because of the novelty of our task and the sICA denoising method, it is difficult to directly compare the current results with those reported in previous studies.

For our functional connectivity analysis, we chose the STG peak showing the greatest change in Toyomura et al. (2011) which was found in the Rhythm vs. Solo contrast. We note their Rhythm condition consisted of metronome-paced speech, which differs from the fluency-inducing condition in the present study. Therefore, it is possible that using different seeds might reveal greater differences in activity patterns between natural and fluent speech, which should be explored in future studies. Nevertheless, our results support the view that increased sensorimotor integration – as evidenced by our induced fluency choral reading condition – is associated with improved neural communication between auditory and motor regions.

Conclusion

This study leveraged an advanced fMRI de-noising method to allow us to investigate brain activity patterns during continuous speech in adults who stutter and controls under choral and solo reading conditions. Overall, brain activity differences between AWS relative to controls in the two conditions were similar, showing expected patterns of hyperactivity in premotor/motor regions but underactivity in auditory regions. Functional connectivity of left STG showed that within the AWS group there was increased correlated activity with the right insula during choral speech, as well as heightened connectivity with regions of DMN during solo speech. These findings suggest that induced fluency conditions specifically modulated brain activity in the AWS group. Further, they indicate possible interference by the DMN during natural, stuttering-prone speech in AWS, and that enhanced coordination between auditory and motor regions may support fluent speech. These findings have clinical implications for designing interventions that involve fluency-inducing conditions to treat stuttering.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by the Institutional Review Board of the University of Michigan Medical School (IRBMED). The patients/participants provided their written informed consent to participate in this study.

Author contributions

EG, S-EC, and HC contributed to the concept and design of the study. EG and HC collected to the data. EG, S-EC, HC, YL, and SL contributed to the analysis and interpretation of results and drafted the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was funded by the American Speech-Language-Hearing Foundation (S-EC), National Institute on Deafness and Other Communication Disorders (HC; R21DC015853), and the Matthew K. Smith Stuttering Research Fund (S-EC).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnhum.2022.894676/full#supplementary-material

References

AbdulSabur, N. Y., Xu, Y., Liu, S., Chow, H. M., Baxter, M., Carson, J., et al. (2014). Neural correlates and network connectivity underlying narrative production and comprehension: a combined fMRI and PET study. *Cortex* 57, 107–127. doi: 10.1016/j.cortex.2014.01.017

Alm, P. A. (2004). Stuttering and the basal ganglia circuits: a critical review of possible relations. *J. Commun. Disord.* 37, 325–369. doi: 10.1016/j.jcomdis.2004.03.001

Andrews, G., Howie, P. M., Dozsa, M., and Guitar, B. E. (1982). Stuttering. *J. Speech Lang. Hear. Res.* 25, 208–216. doi: 10.1044/jshr.2502.208

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage* 38, 95–113. doi: 10.1016/j.neuroimage.2007.07.007

Azrin, N., Jones, R. J., and Flye, B. (1968). A Synchronization Effect and Its Application to Stuttering by a Portable Apparatus1. *J. Appl. Behav. Anal.* 1, 283–295. doi: 10.1901/jaba.1968.1-283

Barber, V. (1940). Studies in the Psychology of Stuttering, XVI. *J. Speech Disord.* 5, 29–42. doi: 10.1044/jshd.0501.29

Belyk, M., Kraft, S. J., and Brown, S. (2015). Stuttering as a trait or state—An ALE meta-analysis of neuroimaging studies. *Eur. J. Neurosci.* 41, 275–284. doi: 10.1111/ejn.12765

Bloodstein, O., and Ratner, N. B. (2008). *A Handbook On Stuttering*. New York, NY: Thomson Delmar Learning.

Bohland, J. W., and Guenther, F. H. (2006). An fMRI investigation of syllable sequence production. *NeuroImage* 32, 821–841. doi: 10.1016/j.neuroimage.2006.04.173

Braun, A. (1997). Altered patterns of cerebral activity during speech and language production in developmental stuttering. An H2(15)O positron emission tomography study. *Brain* 120, 761–784. doi: 10.1093/brain/120.5.761

Braun, A. R., Balkin, T. J., Wesenten, N. J., Carson, R. E., Varga, M., Baldwin, P., et al. (1997). Regional cerebral blood flow throughout the sleep-wake cycle. An H2(15)O PET study. *Brain* 120(Pt 7), 1173–1197. doi: 10.1093/brain/120.7.1173

Brown, S., Ingham, R. J., Ingham, J. C., Laird, A. R., and Fox, P. T. (2005). Stuttered and fluent speech production: an ALE meta-analysis of functional neuroimaging studies. *Hum. Brain Mapp.* 25, 105–117. doi: 10.1002/hbm.20140

Buchsbaum, B. R., Hickok, G., and Humphries, C. (2001). Role of left posterior superior temporal gyrus in phonological processing for speech perception and production. *Cogn. Sci.* 25, 663–678. doi: 10.1207/s15516709cog2505_2

Buchsbaum, B. R., Olsen, R. K., Koch, P., and Berman, K. F. (2005). Human Dorsal and Ventral Auditory Streams Subserve Rehearsal-Based and Echoic Processes during Verbal Working Memory. *Neuron* 48, 687–697. doi: 10.1016/j.neuron.2005.09.029

Buckner, R. L., Andrews-Hanna, J. R., and Schacter, D. L. (2008). The Brain's Default Network: anatomy, Function, and Relevance to Disease. *Ann. N.Y. Acad. Sci.* 1124, 1–38. doi: 10.1196/annals.1440.011

Budde, K. S., Barron, D. S., and Fox, P. T. (2014). Stuttering, induced fluency, and natural fluency: a hierarchical series of activation likelihood estimation meta-analyses. *Brain Lang.* 139, 99–107. doi: 10.1016/j.bandl.2014.10.002

Cai, S., Beal, D. S., Ghosh, S. S., Guenther, F. H., and Perkell, J. S. (2014). Impaired timing adjustments in response to time-varying auditory perturbation during connected speech production in persons who stutter. *Brain Lang.* 129, 24–29. doi: 10.1016/j.bandl.2014.01.002

Cai, S., Beal, D. S., Ghosh, S. S., Tiede, M. K., Guenther, F. H., and Perkell, J. S. (2012). Weak Responses to Auditory Feedback Perturbation during Articulation in Persons Who Stutter: evidence for Abnormal Auditory-Motor Transformation. *PLoS One* 7:e41830. doi: 10.1371/journal.pone.0041830

Chang, S.-E., and Guenther, F. H. (2020). Involvement of the Cortico-Basal Ganglia-Thalamocortical Loop in Developmental Stuttering. *Front. Psychol.* 10:3088. doi: 10.3389/fpsyg.2019.03088

Chang, S.-E., and Zhu, D. C. (2013). Neural network connectivity differences in children who stutter. *Brain* 136, 3709–3726. doi: 10.1093/brain/awt275

Chang, S.-E., Angstadt, M., Chow, H. M., Etchell, A. C., Garnett, E. O., Choo, A. L., et al. (2018). Anomalous network architecture of the resting brain in children who stutter. *J. Fluen. Disord.* 55, 46–67. doi: 10.1016/j.jfludis.2017.01.002

Chang, S.-E., Garnett, E. O., Etchell, A., and Chow, H. M. (2019). Functional and Neuroanatomical Bases of Developmental Stuttering: current Insights. *Neuroscientist* 25, 566–582. doi: 10.1177/1073858418803594

Chow, H. M., Mar, R. A., Xu, Y., Liu, S., Wagage, S., and Braun, A. R. (2014). Embodied Comprehension of Stories: interactions between Language Regions and Modality-specific Neural Systems. *J. Cogn. Neurosci.* 26, 279–295. doi: 10.1162/jocn_a_00487

Chow, H. M., Mar, R. A., Xu, Y., Liu, S., Wagage, S., and Braun, A. R. (2015). Personal experience with narrated events modulates functional connectivity within visual and motor systems during story comprehension. *Hum. Brain Mapp.* 36, 1494–1505. doi: 10.1002/hbm.22718

Civier, O., Bullock, D., Max, L., and Guenther, F. H. (2013). Computational modeling of stuttering caused by impairments in a basal ganglia thalamo-cortical circuit involved in syllable selection and initiation. *Brain Lang.* 126, 263–278. doi: 10.1016/j.bandl.2013.05.016

Civier, O., Tasko, S. M., and Guenther, F. H. (2010). Overreliance on auditory feedback may lead to sound/syllable repetitions: simulations of stuttering and fluency-inducing conditions with a neural model of speech production. *J. Fluen. Disord.* 35, 246–279. doi: 10.1016/j.jfludis.2010.05.002

Cortese, S., Kelly, C., Chabernaud, C., Proal, E., Di Martino, A., Milham, M. P., et al. (2012). Toward Systems Neuroscience of ADHD: a Meta-Analysis of 55 fMRI Studies. *Am. J. Psychiatry* 169, 1038–1055. doi: 10.1176/appi.ajp.2012.11101521

Cox, R. W., Chen, G., Glen, D. R., Reynolds, R. C., and Taylor, P. A. (2017). FMRI Clustering in AFNI: false-Positive Rates Redux. *Brain Connect.* 7, 152–171. doi: 10.1089/brain.2016.0475

Daliri, A., and Max, L. (2018). Stuttering adults' lack of pre-speech auditory modulation normalizes when speaking with delayed auditory feedback. *Cortex* 99, 55–68. doi: 10.1016/j.cortex.2017.10.019

Daliri, A., Wieland, E. A., Cai, S., Guenther, F. H., and Chang, S.-E. (2018). Auditory-motor adaptation is reduced in adults who stutter but not in children who stutter. *Dev. Sci.* 21:e12521. doi: 10.1111/desc.12521

Davidow, J. H. (2014). Systematic studies of modified vocalization: the effect of speech rate on speech production measures during metronome-paced speech in persons who stutter: speech rate and speech production measures during metronome-paced speech in PWS. *Int. J. Lang. Commun. Disord.* 49, 100–112. doi: 10.1111/1460-6984.12050

De Nil, L. F., Kroll, R. M., Lafaille, S. J., and Houle, S. (2003). A positron emission tomography study of short- and long-term treatment effects on functional brain activation in adults who stutter. *J. Fluen. Disord.* 28, 357–380. doi: 10.1016/j.jfludis.2003.07.002

De Nil, L., Beal, D. S., Lafaille, S. J., Kroll, R. M., Crawley, A. P., and Gracco, V. L. (2008). The effects of simulated stuttering and prolonged speech on the neural activation patterns of stuttering and nonstuttering adults. *Brain Lang.* 107, 114–123. doi: 10.1016/j.bandl.2008.07.003

Eichorn, N., Marton, K., Schwartz, R. G., Melara, R. D., and Pirutinsky, S. (2016). Does Working Memory Enhance or Interfere With Speech Fluency in Adults Who Do and Do Not Stutter? Evidence From a Dual-Task Paradigm. *J. Speech Lang. Hear. Res.* 59, 415–429. doi: 10.1044/2015_JSLHR-S-15-0249

Eichorn, N., Pirutinsky, S., and Marton, K. (2019). Effects of different attention tasks on concurrent speech in adults who stutter and fluent controls. *J. Fluen. Disord.* 61:105714. doi: 10.1016/j.jfludis.2019.105714

Etchell, A. C., Johnson, B. W., and Sowman, P. F. (2014). Behavioral and multimodal neuroimaging evidence for a deficit in brain timing networks in stuttering: a hypothesis and theory. *Front. Hum. Neurosci.* 8:467. doi: 10.3389/fnhum.2014.00467

Fox, P. T., Ingham, R. J., Ingham, J. C., Hirsch, T. B., Downs, J. H., Martin, C., et al. (1996). A PET study of the neural systems of stuttering. *Nature* 382, 158–162. doi: 10.1038/382158a0

Fox, P. T., Ingham, R. J., Ingham, J. C., Zamarripa, F., Xiong, J. H., and Lancaster, J. L. (2000). Brain correlates of stuttering and syllable production: a PET performance-correlation analysis. *Brain* 123, 1985–2004. doi: 10.1093/brain/123.10.1985

Frankford, S. A., Heller Murray, E. S., Masapollo, M., Cai, S., Tourville, J. A., Nieto-Castañón, A., et al. (2021). The Neural Circuitry Underlying the "Rhythm Effect" in Stuttering. *J. Speech Lang. Hear. Res.* 64, 2325–2346. doi: 10.1044/2021_JSLHR-20-00328

Garnett, E. O., Chow, H. M., Choo, A. L., and Chang, S.-E. (2019). Stuttering Severity Modulates Effects of Non-invasive Brain Stimulation in Adults Who Stutter. *Front. Hum. Neurosci.* 13:411. doi: 10.3389/fnhum.2019.00411

Giraud, A. (2008). Severity of dysfluency correlates with basal ganglia activity in persistent developmental stuttering. *Brain Lang.* 104, 190–199. doi: 10.1016/j.bandl.2007.04.005

Golfinopoulos, E., Tourville, J. A., Bohland, J. W., Ghosh, S. S., Nieto-Castanon, A., and Guenther, F. H. (2011). FMRI investigation of unexpected somatosensory

feedback perturbation during speech. *NeuroImage* 55, 1324–1338. doi: 10.1016/j.neuroimage.2010.12.065

Guenther, F. H., and Hickok, G. (2015). "Chapter 9—Role of the auditory system in speech production," in *Handbook of Clinical Neurology*, eds M. J. Aminoff, F. Boller, and D. F. Swaab (Amsterdam: Elsevier), 161–175. doi: 10.1016/B978-0-444-62630-1.00009-3

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nat. Rev. Neurosci.* 13, 135–145. doi: 10.1038/nrn3158

Hickok, G., Buchsbaum, B., Humphries, C., and Muftuler, T. (2003). Auditory–Motor Interaction Revealed by fMRI: speech, Music, and Working Memory in Area Spt. *J. Cogn. Neurosci.* 15, 673–682. doi: 10.1162/jocn.2003.15.5.673

Hickok, G., Houde, J., and Rong, F. (2011). Sensorimotor Integration in Speech Processing: computational Basis and Neural Organization. *Neuron* 69, 407–422. doi: 10.1016/j.neuron.2011.01.019

Hickok, G., Okada, K., and Serences, J. T. (2009). Area Spt in the Human Planum Temporale Supports Sensory-Motor Integration for Speech Processing. *J. Neurophysiol.* 101, 2725–2732. doi: 10.1152/jn.91099.2008

Hutchinson, J. M., and Norris, G. M. (1977). The differential effect of three auditory stimuli on the frequency of stuttering behaviors. *J. Fluen. Disord.* 2, 283–293. doi: 10.1016/0094-730X(77)90032-8

Ingham, R. J. (2003). Brain imaging and stuttering: some reflections on current and future developments. *J. Fluen. Disord.* 28, 411–420. doi: 10.1016/j.jfludis.2003.08.003

Ingham, R. J., Grafton, S. T., Bothe, A. K., and Ingham, J. C. (2012). Brain activity in adults who stutter: similarities across speaking tasks and correlations with stuttering frequency and speaking rate. *Brain Lang.* 122, 11–24. doi: 10.1016/j.bandl.2012.04.002

Kalinowski, J., and Saltuklaroglu, T. (2003). Choral speech: the amelioration of stuttering via imitation and the mirror neuronal system. *Neurosci. Biobehav. Rev.* 27, 339–347. doi: 10.1016/S0149-7634(03)00063-0

Kearney, E., and Guenther, F. H. (2019). Articulating: the neural mechanisms of speech production. *Lang. Cogn. Neurosci.* 34, 1214–1229. doi: 10.1080/23273798.2019.1589541

Kell, C. A., Neumann, K., von Kriegstein, K., Posenenske, C., von Gudenberg, A. W., Euler, H., et al. (2009). How the brain repairs stuttering. *Brain* 132, 2747–2760. doi: 10.1093/brain/awp185

Korzeczek, A., Primaßin, A., Wolff von Gudenberg, A., Dechent, P., Paulus, W., Sommer, M., et al. (2021). Fluency shaping increases integration of the command-to-execution and the auditory-to-motor pathways in persistent developmental stuttering. *NeuroImage* 245:118736. doi: 10.1016/j.neuroimage.2021.118736

Lee, M. H., Hacker, C. D., Snyder, A. Z., Corbetta, M., Zhang, D., Leuthardt, E. C., et al. (2012). Clustering of Resting State Networks. *PLoS One* 7:e40370. doi: 10.1371/journal.pone.0040370

Lombard, É (1911). Le signe de l'élévation de la voix. *Ann. Mal. Loreille Larynx* XXXVII, 101–109.

Maguire, G. A., Riley, G. D., and Yu, B. P. (2002). A neurological basis of stuttering? *Lancet Neurol.* 1:407. doi: 10.1016/S1474-4422(02)00217-X

Maguire, G. A., Yu, B. P., Franklin, D. L., and Riley, G. D. (2004). Alleviating stuttering with pharmacological interventions. *Expert Opin. Pharmacother.* 5, 1565–1571. doi: 10.1517/14656566.5.7.1565

Max, L., and Daliri, A. (2019). Limited Pre-Speech Auditory Modulation in Individuals Who Stutter: data and Hypotheses. *J. Speech Lang. Hear. Res.* 62, 3071–3084. doi: 10.1044/2019_JSLHR-S-CSMC7-18-0358

Max, L., Guenther, F. H., Gracco, V. L., Ghosh, S. S., and Wallace, M. E. (2004). Unstable or Insufficiently Activated Internal Models and Feedback-Biased Motor Control as Sources of Dysfluency: a Theoretical Model of Stuttering. *Contemp. Issues Commun. Sci. Disord.* 31, 105–122. doi: 10.1044/cicsd_31_S_105

Neef, N. E., Anwander, A., and Friederici, A. D. (2015). The Neurobiological Grounding of Persistent Stuttering: from Structure to Function. *Curr. Neurol. Neurosci. Rep.* 15:63. doi: 10.1007/s11910-015-0579-4

Neumann, K., Preibisch, C., Euler, H. A., von Gudenberg, A. W., Lanfermann, H., Gall, V., et al. (2005). Cortical plasticity associated with stuttering therapy. *J. Fluen. Disord.* 30, 23–39. doi: 10.1016/j.jfludis.2004.12.002

Riley, G. D. (2009). *Stuttering severity instrument for children and adults (SSI-4)* (4th Edn.). Austin, TX: Pro-Ed.

Salmelin, R., Schnitzler, A., Schmitz, F., Jäncke, L., Witte, O. W., and Freund, H.-J. (1998). Functional organization of the auditory cortex is different in stutterers and fluent speakers. *NeuroReport* 9, 2225–2229. doi: 10.1097/00001756-199807130-00014

Sonuga-Barke, E. J. S., and Castellanos, F. X. (2007). Spontaneous attentional fluctuations in impaired states and pathological conditions: a neurobiological hypothesis. *Neurosci. Biobehav. Rev.* 31, 977–986. doi: 10.1016/j.neubiorev.2007.02.005

Spreng, R. N., Mar, R. A., and Kim, A. S. N. (2009). The Common Neural Basis of Autobiographical Memory, Prospection, Navigation, Theory of Mind, and the Default Mode: a Quantitative Meta-analysis. *J. Cogn. Neurosci.* 21, 489–510. doi: 10.1162/jocn.2008.21029

Stager, S. V., Denman, D. W., and Ludlow, C. L. (1997). Modifications in Aerodynamic Variables by Persons Who Stutter Under Fluency-Evoking Conditions. *J. Speech Lang. Hear. Res.* 40, 832–847. doi: 10.1044/jslhr.4004.832

Stager, S. V., Jeffries, K. J., and Braun, A. R. (2003). Common features of fluency-evoking conditions studied in stuttering subjects and controls: an H215O PET study. *J. Fluen. Disord.* 28, 319–336. doi: 10.1016/j.jfludis.2003.08.004

Tourville, J. A., and Guenther, F. H. (2011). The DIVA model: a neural theory of speech acquisition and production. *Lang. Cogn. Process.* 26, 952–981. doi: 10.1080/01690960903498424

Tourville, J. A., Reilly, K. J., and Guenther, F. H. (2008). Neural mechanisms underlying auditory feedback control of speech. *NeuroImage* 39, 1429–1443. doi: 10.1016/j.neuroimage.2007.09.054

Toyomura, A., Fujii, T., and Kuriki, S. (2011). Effect of external auditory pacing on the neural activity of stuttering speakers. *NeuroImage* 57, 1507–1516. doi: 10.1016/j.neuroimage.2011.05.039

Toyomura, A., Fujii, T., and Kuriki, S. (2015). Effect of an 8-week practice of externally triggered speech on basal ganglia activity of stuttering and fluent speakers. *NeuroImage* 109, 458–468. doi: 10.1016/j.neuroimage.2015.01.024

Wulf, G., and Lewthwaite, R. (2016). Optimizing performance through intrinsic motivation and attention for learning: the OPTIMAL theory of motor learning. *Psychon. Bull. Rev.* 23, 1382–1414. doi: 10.3758/s13423-015-0999-9

Xu, Y., Tong, Y., Liu, S., Chow, H. M., AbdulSabur, N. Y., Mattay, G. S., et al. (2014). Denoising the speaking brain: toward a robust technique for correcting artifact-contaminated fMRI data under severe motion. *NeuroImage* 103, 33–47. doi: 10.1016/j.neuroimage.2014.09.013

Znamenskiy, P., and Zador, A. M. (2013). Corticostriatal neurons in auditory cortex drive decisions during auditory discrimination. *Nature* 497, 482–485. doi: 10.1038/nature12077

frontiers | Frontiers in Human Neuroscience

TYPE Original Research
PUBLISHED 28 July 2022
DOI 10.3389/fnhum.2022.859731

Check for updates

OPEN ACCESS

EDITED BY
Xing Tian,
New York University Shanghai, China

REVIEWED BY
Fuyin Yang,
Shanghai Jiao Tong University, China
Chao Yan,
East China Normal University, China

*CORRESPONDENCE
Sonja A. Kotz
sonja.kotz@maastrichtuniversiy.nl

SPECIALTY SECTION
This article was submitted to Speech and Language, a section of the journal Frontiers in Human Neuroscience

RECEIVED 21 January 2022
ACCEPTED 29 June 2022
PUBLISHED 28 July 2022

CITATION
Johnson JF, Belyk M, Schwartze M, Pinheiro AP and Kotz SA (2022) Hypersensitivity to passive voice hearing in hallucination proneness. *Front. Hum. Neurosci.* 16:859731. doi: 10.3389/fnhum.2022.859731

COPYRIGHT
© 2022 Johnson, Belyk, Schwartze, Pinheiro and Kotz. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Hypersensitivity to passive voice hearing in hallucination proneness

Joseph F. Johnson[1], Michel Belyk[2], Michael Schwartze[1], Ana P. Pinheiro[3] and Sonja A. Kotz[1,4]*

[1]Department of Neuropsychology and Psychopharmacology, University of Maastricht, Maastricht, Netherlands, [2]Department of Psychology, Edge Hill University, Ormskirk, United Kingdom, [3]Faculdade de Psicologia, Universidade de Lisboa, Lisbon, Portugal, [4]Department of Neuropsychology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

Voices are a complex and rich acoustic signal processed in an extensive cortical brain network. Specialized regions within this network support voice perception and production and may be differentially affected in pathological voice processing. For example, the experience of hallucinating voices has been linked to hyperactivity in temporal and extra-temporal voice areas, possibly extending into regions associated with vocalization. Predominant self-monitoring hypotheses ascribe a primary role of voice production regions to auditory verbal hallucinations (AVH). Alternative postulations view a generalized perceptual salience bias as causal to AVH. These theories are not mutually exclusive as both ascribe the emergence and phenomenology of AVH to unbalanced top-down and bottom-up signal processing. The focus of the current study was to investigate the neurocognitive mechanisms underlying predisposition brain states for emergent hallucinations, detached from the effects of inner speech. Using the temporal voice area (TVA) localizer task, we explored putative hypersalient responses to passively presented sounds in relation to hallucination proneness (HP). Furthermore, to avoid confounds commonly found in in clinical samples, we employed the Launay-Slade Hallucination Scale (LSHS) for the quantification of HP levels in healthy people across an experiential continuum spanning the general population. We report increased activation in the right posterior superior temporal gyrus (pSTG) during the perception of voice features that positively correlates with increased HP scores. In line with prior results, we propose that this right-lateralized pSTG activation might indicate early hypersensitivity to acoustic features coding speaker identity that extends beyond own voice production to perception in healthy participants prone to experience AVH.

KEYWORDS

temporal voice area (TVA), voice perception, hallucination proneness, functional magnetic brain imaging (fMRI), neuroimaging, salience account

Introduction

The human voice is a complex signal that carries rich information. This allows the listener not only to identify linguistic messages but also who speaks and how something is said (Belin et al., 2004; Lavan et al., 2019). Some individuals experience auditory verbal hallucinations (AVH), in which they perceive voices in the absence of a corresponding incoming voice signal (Bentall, 1990; Anthony, 2004; Brookwell et al., 2013). Experience of AVH is a key symptom of schizophrenia (Bauer et al., 2011; Larøi et al., 2012; Hugdahl and Sommer, 2018). Yet, it is also reported in multiple other psychiatric, developmental, and neurological disorders (Van Os et al., 2000; Reininghaus et al., 2016; Waters and Fernyhough, 2017; Rollins et al., 2019; Zhuo et al., 2019) and in a minority of otherwise healthy people (Beavan et al., 2011; Linscott and Van Os, 2013; McGrath et al., 2015). Variability in AVH phenomenology exists within and across brain disorders (Stephane et al., 2003; Jones, 2010) and between clinical and non-clinical voice hearers (Daalman et al., 2011; Larøi et al., 2012; Johns et al., 2014; Baumeister et al., 2017). However, hallucinated voices commonly carry information regarding the identity or emotion of a perceived speaker (Stephane et al., 2003; Larøi and Woodward, 2007; Badcock and Chhabra, 2013; McCarthy-Jones et al., 2014), therefore involving a wide range of cortical areas in a voice perception network. Multiple cognitive theories have been proposed delineating the emergence and phenomenology of AVH (Jones, 2010; Ćurčić-Blake et al., 2017; Rollins et al., 2019). One long standing model considers hallucinations as the misattribution of self-generated input to an outside source (Feinberg, 1978). In terms of AVH, signals from voice production cortical regions during inner speech are misperceived as hearing someone else speak (Allen et al., 2007a; Jones and Fernyhough, 2007a,b; Swiney and Sousa, 2014; Gregory, 2016). Recently, competing theories have gained traction, claiming that the initiation of hallucinations does not require motor activity while they are, at their core, misperceived sensations from the environment (e.g., Ford and Mathalon, 2019; Thakkar et al., 2021).

The selection and processing of sensory inputs from the environment relevant to learning, adaptation, or behavioral responses involves multiple regions and distributed networks across the brain. The role of salience attribution within this integrated system provides the necessary trigger to shift processing from a state of rest to active sensation and perception (Menon and Uddin, 2010; Menon, 2011; Palaniyappan and Liddle, 2012; Uddin, 2015). According to this framework, increased auditory cortex activation associated with AVH can be ascribed to a bottom-up hypersensitivity, or salience bias, toward irrelevant sounds. The modulation and over-weighting of top-down predictions may influence this salience bias as well as guide the system to perceive what it expects in meaningless unimodal and multimodal stimuli (Friston, 2005, 2012; Fletcher and Frith, 2009; Deneve and Jardri, 2016; Jardri et al., 2016; Leptourgos et al., 2017). Since voice signals in humans are inherently salient to human listeners, they may be particularly implicated in hypersensitive responses leading to false perceptions. Furthermore, for those who experience AVH, the engagement of brain regions controlling inner speech signals, memory retrieval, and emotion may then guide the phenomenology of the perceived speech in terms of content and speaker-related features (Waters et al., 2012). Abnormal salience processing has been strongly linked to positive symptoms of schizophrenia (Miyata, 2019).

Researching the contribution of these mechanisms to AVH in non-clinical samples may be particularly useful as it avoids potential confounds seen in clinical populations such as medication, age of onset, and duration of symptoms that may affect brain structure and function (Verdoux and van Os, 2002; Kelleher et al., 2010; Kelleher and Cannon, 2011). This perspective is in line with the experiential continuum of psychosis (Johns, 2005; Beavan et al., 2011; Larøi et al., 2012; de Leede-Smith and Barkus, 2013; Johns et al., 2014; Zhuo et al., 2019), whereby functional variability in the mechanisms serving perception across the population account for the spectrum of normal experience, vivid perceptions and imagery, sub-clinical forms of hallucinations, and those seen in full-blown psychosis. The revised Launay-Slade Hallucination Scale (LSHS) is as a measure of perceptual experience and beliefs associated with vivid daydreams, thoughts, imagery, and those related to false perceptions such as visual and auditory hallucinations (Larøi and Van Der Linden, 2005). The LSHS provides a measure of hallucination proneness (HP), where higher scores signify increasing abnormality in perceptual experience and beliefs, including true hallucinations. Although individual items from the LSHS can be used to identify the prevalence of AVH (e.g., Kompus et al., 2015), HP itself is not a measure of risk for psychosis.

Two critical factors have been incorporated into the formulation of our hypotheses. First, differential brain activity may indicate abnormal voice processing as a predisposition for false perceptions, i.e., activation patterns similar to those during hallucinations. Second, the localization of reported changes in brain responses may indicate a specific stage within hierarchical voice processing at which this predisposition manifests. To date, no consensus has been empirically established regarding a trait-based association between hallucinations and brain responses to the voice. For example, when presented with voices, patients who commonly experience hallucinations display decreases (Copolov et al., 2003), increases (Martí-Bonmatí et al., 2007; Parellada et al., 2008; Escartí et al., 2010), or no activation differences in voice selective temporal regions (Woodruff et al., 1997; Simons et al., 2010). Such inconsistency is likely due to methodological

heterogeneity (Bohlken et al., 2017). For example, these studies differed in terms of stimulus type, stimulus content, and the inclusion of a non-hallucinating patient control group. Moreover, patients with chronic hallucinations can experience spontaneous AVH during scanning (Jardri et al., 2011; Kühn and Gallinat, 2012; van Lutterveld et al., 2013; Zmigrod et al., 2016), which may even be unintentionally elicited by tasks (e.g., Copolov et al., 2003; Parellada et al., 2008). Although this hallucinatory state elicits brain activity in voice perception regions, simultaneous external voice input during AVH results in a paradoxical net activity decrease (Kompus et al., 2011; Hugdahl and Sommer, 2018).

The localization of changes in functional brain activity within the voice processing network can be particularly informative in determining how HP may arise. Within the upper bank and lateral regions of the temporal lobe, voice signals are processed hierarchically along a pathway composed of multiple functional subsystems or components (Belin et al., 2004; Pernet et al., 2015; Zhang et al., 2021). The engagement of these temporal voice areas (TVA) starts with the evaluation of low-level acoustic features in the posterior superior temporal gyrus (STG), an area specialized in processing spectro-temporal properties of complex sounds (Griffiths and Warren, 2004; Warren J. D. et al., 2005; Warren J. E. et al., 2005). Further processing occurs along hemispherically specialized pathways, with linguistic features predominantly in the left and paralinguistic (i.e., speaker-related information) in the right side of the brain (Belin et al., 2000; Formisano et al., 2008). However, some stimuli such as emotional vocalizations contain both speaker-and speech-relevant information and involve bilateral processing of separate features in the signal (Schirmer and Kotz, 2006). Importantly, AVH often contain marked paralinguistic information about speaker identity or emotion (Larøi and Woodward, 2007; Larøi et al., 2012; McCarthy-Jones et al., 2014). In non-clinical voice hearers, however, the degree of perceived emotional valence is less prominent (Daalman et al., 2012; de Boer et al., 2016). Speaker-related feature processing operates along a multi-stage hierarchy in the right temporal cortex along a posterior to anterior gradient (Nakamura et al., 2001, Belin and Zatorre, 2003; von Kriegstein et al., 2003, von Kriegstein and Giraud, 2004). The TVA localizer is a widely used fMRI task which reliably identifies activation peaks localized in the bilateral anterior, middle, and posterior superior temporal cortex (Pernet et al., 2015). By comparing voice to non-voice activation in response to passively heard sounds, regions of interest (ROI) can be defined for further investigation. Using ROIs produced by this task, we predicted HP-related early sensitivity to low-level voice features to be isolated to the posterior STG ROI. Alternatively, changes to voice processing in the anterior direction of the right STG might indicate an abnormal salience bias for identity or emotion associated with an increasing propensity to hallucinate.

Methods

Participants

Twenty-six participants took part in this study, recruited through the SONA system and social media channels at Maastricht University, Netherlands. Participants were provided with informed consent and offered university study credit for compensation. Exclusion criteria included any history of psychotic disorder, neurological impairment, history of drug dependence or abuse, and traumatic brain injury. Participants were screened for MRI safety and reported no metal implants, claustrophobia, or pregnancy. Furthermore, all participants reported no known hearing deficits. Robust statistics using the interquartile range rule for participant age revealed one outlier (Rousseeuw and Hubert, 2011), leading to the exclusion of the dataset from further analysis. Of the resulting 25 individuals (17 female), the average age was 20.92 years (SD 3.95; range 18–32). The Ethical Review Committee of the Faculty of Psychology and Neuroscience at Maastricht University (ERCPN-176_08_02_2017) approved this study.

Hallucination proneness

The revised LSHS was employed as a self-report measure of HP (Larøi and Van Der Linden, 2005). The questionnaire consists of 16 items targeting tactile, sleep-related, visual, and auditory modalities of psychosis-like experience as well as vivid thoughts and daydreaming. Responses were given using a five-point Likert scale, measuring the extent to which each statement applied to them. The sum of all responses equated to an overall HP measure. Furthermore, to investigate the exclusivity of auditory-only items, subscores of three items were summed to produce a composite score (Larøi et al., 2004; Larøi and Van Der Linden, 2005).

Voice area fMRI-localizer task

Voice selective cortical brain regions were identified using a standard fMRI-localizer task (Belin et al., 2000). This widely used tool reliably probes activity across three bilateral peaks in the superior temporal gyrus (e.g., Pernet et al., 2015), often designated as anterior, middle, and posterior temporal voice areas (TVA). Furthermore, many studies applying this task have reported extra-temporal voice regions, such as the inferior frontal cortex (IFC). The voice area localizer consists of 20 vocal (V) and 20 non-vocal (NV) trials. Additionally, 20 silence (S) trials are included allowing relaxation of the hemodynamic response to auditory stimuli. The voice condition is composed of human speech (words, syllables, or sentence excerpts) and non-speech voices produced by male and female speakers of different

ages (7 babies, 12 adults, 23 children, and 5 elderly). This broad selection of voice stimuli allows for the probing and inclusion of functionally diverse regions of TVA. Conversely, the non-voice condition includes environmental (natural and animal) and man-made (e.g., cars, alarm clocks, instrumental music) sounds. Sound clips are presented at a standard 70 db volume (for a detailed report of the included sounds and recording duration, amplitude, and frequency see Pernet et al., 2015). Trials were presented in a pseudorandom order, each with a duration of eight seconds. With a two second inter-trial interval, the total run time of the task was 10 min.

FMRI data acquisition

Scanning was conducted using a Siemens 3T Magnetom Prisma Fit equipped with a 32-channel head coil (Siemens Healthcare, Erlangen, Germany), at the Scannexus facilities (Maastricht, Netherlands). Structural whole-brain T1-weighted images were acquired with a single-shot echoplanar imaging (EPI) sequence [field of view (FOV) 256 mm; 192 axial slices; 1 mm slice thickness; 1 mm × 1 mm × 1 mm voxel size; repetition time (TR) of 2250 ms; echo-time (TE) 2.21 ms]. For the functional localizer task, T2-weighted EPI scans were collected (FOV 208 mm; 60 axial slices; 2 mm slice thickness; 2 mm × 2 mm × 2 mm voxel size; TE 30 ms; flip angle = 77°). To reduce scanner noise interference, auditory stimuli were presented *via* S14 MR-compatible earphones, fitted with foam earplugs (Sensimetrics Corporation). Furthermore, to provide relative silence during playback of auditory stimuli, a long inter-acquisition-interval was adopted where time between consecutive acquisition was delayed, resulting in a TR of 10 s. The delayed TR was timed to allow a 2,000 ms acquisition period during peak activation in the auditory cortex (Belin et al., 1999; Hall et al., 1999).

Data pre-processing and analysis

Pre-processing of the TVA localizer blood-oxygen-level-dependent (BOLD) signal was conducted in SPM12 (Wellcome Department of Cognitive Neurology, London, United Kingdom). A standard pipeline was applied using slice timing correction, realignment and unwarping, segmentation, normalization to standard (MNI) space (Fonov et al., 2009), and 8 mm isotropic Gaussian kernel full width at half maximum (FWHM) smoothing. Analysis followed a two-level procedure in which contrast estimates were first determined as fixed effects at the level of individual participants then modeled as random effects at the level of the sample. Contrast estimates were computed on BOLD data to assess voice sensitivity (V > NV) and sensitivity to environmental sounds (NV > S) for each participant. A first-level fixed-effects GLM analysis for the conjunction analysis [(V > NV) ∩ (V > S)] was computed to localize the temporal voice areas. A second-level random-effects analysis tested for group-level significance and determined the ROIs for parameter extraction. Contrast estimates of V > S and NV > S were then used to contrast voice with non-voice activity, corrected for baseline, in the subsequent hypothesis-driven ROI analysis to investigate the correlation of voice-preferential TVA activity compared to HP. Contrast estimates were extracted from a 5 mm radius of the center coordinates from each region of peak activity produced in the TVA-localizer using the SPM MARSbar toolbox (Brett et al., 2002). Pearson's correlation analysis using bootstrapping (5000 samples) and bias-corrected confidence intervals was then employed to test for significant relationships between the sensitivity of the voice ROIs and HP measures.

Results

Hallucination proneness

For the HP composite score (possible maximum score of 80), the mean self-reported rating was 25.20 (SD 10.47; range 0–42). The HP auditory subscale mean score (possible maximum score 15) was 3.92 (SD 2.74; range 0–11). To test for normality of the distribution of demographics and HP across the sample, Shapiro–Wilk tests were conducted. Both total LSHS (0.948, df = 25, p = 0.229) and auditory subscale (0.928, df = 25, p = 0.078) were not different from normal. A moderately strong correlation was also found between LSHS auditory subscale and non-auditory item totals (r = 0.457, df = 25, p = 0.019).

Voice area localizer

The fMRI localizer task produced 5 clusters covering bilateral lateral temporal cortices, bilateral inferior frontal gyri, and the right precentral gyrus (preCG) (**Table 1** and **Figure 1**). Within each bilateral temporal cortex "voice patch," peak activity localizations were distinguished in three distinct regions: posterior (pSTG), middle (mSTG), and anterior STG (aSTG). These regions correspond to the expected divisions of the TVA localizer (Pernet et al., 2015).

FMRI correlation

Correlational tests were performed between contrast estimates representing voice preference [(V > S) > (NV > S)] observed in each TVA-ROI with both the composite HP score and the auditory subscore of the LSHS. All thresholds for significance were Bonferroni-adjusted for multiple comparisons using (p < 0.025). Only the right pSTG reached statistical

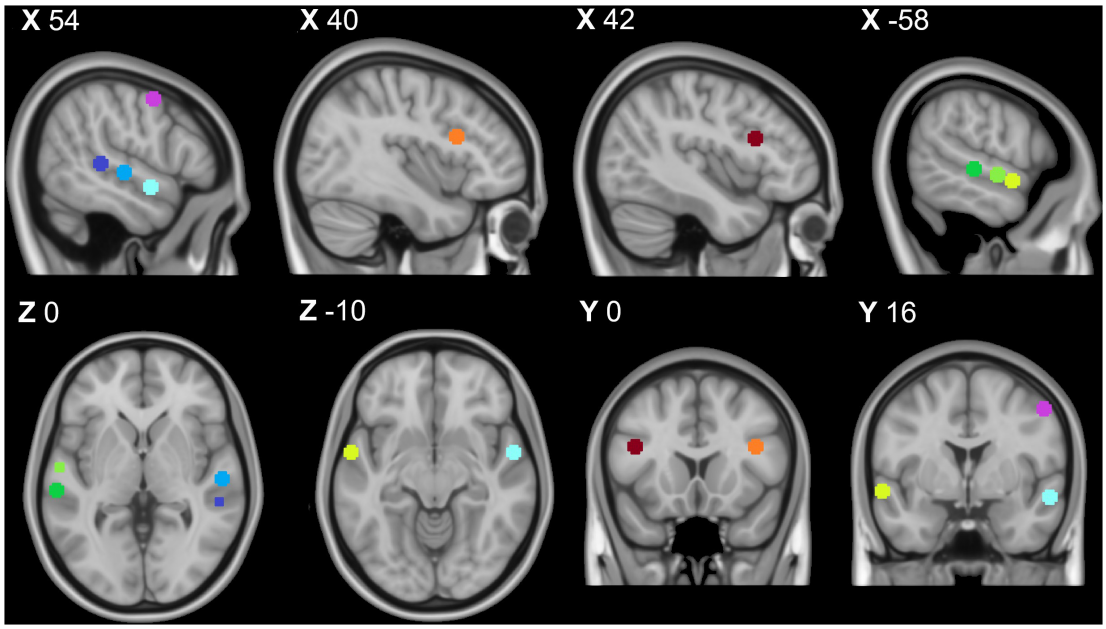


FIGURE 1
Temporal voice area fMRI localizer task results: Purple = right premotor cortex, dark blue = right posterior temporal gyrus, middle blue = right middle temporal gyrus, light blue = right anterior temporal gyrus, orange = right inferior frontal cortex, dark green = left posterior superior temporal gyrus, middle green = left middle superior temporal gyrus, light green = left anterior superior temporal gyrus, and red = left inferior temporal cortex. All coordinates listed in MNI space (x,y,z). This image was created using the FSL toolbox fsleyes (McCarthy, 2022).

TABLE 1 Results from temporal voice area fMRI localizer task.

Cluster #	Hem.	Label	BA	x	y	z	Cluster-Level p-FDR	Peak-Level p-FDR	Size (voxels)
1	L	mSTG	22	−58	−10	−4	1.6782E-17	1.4637E-09	4145
		pSTG	22	−60	−26	0		1.4637E-09	
		aSTG	22	−58	0	−8		1.3575E-08	
2	R	mSTG	22	56	−18	−2	2.0689E-17	1.4637E-09	4010
		aSTG	22	56	0	−12		1.6043E-08	
		pSTG	22	54	−34	4		1.6043E-08	
3	R	pMC	6	52	2	48	0.0049	4.1457E-05	285
4	L	IFC	44	−42	16	22	0.0383	0.0018	142
5	R	IFC	44	40	16	22	0.0227	0.0302	180

Hem, hemisphere; (a/m/p) STG, (anterior/middle/posterior) superior temporal gyrus; pMC, premotor cortex; IFC, inferior frontal cortex; BA, Brodmann's Area; p-FDR, false discovery rate corrected *p*-value (threshold = 0.05). All coordinates listed in MNI space (x, y, z).

significance (r = 0.470, df = 25, p = 0.020) (Table 2 and Figure 2). *Post hoc* correlation analyses were run to assess the relative contributions of both voice (V > S) and non-voice (NV > S) contrasts to correlational analyses (see detailed results in Supplementary Material). We conducted these analyses in order to rule out a general hypersensitivity of temporal cortex activity non-specific to the conditions of interest probed by the conjunction analysis. No significant correlations with HP were found in any ROI for voice (V > S), however, a significant negative correlation was reported in the right IFC for non-voice (V > S) sensitive activity (r = −0.614, df = 25, p = 0.001).

Discussion

The current study investigated whether a measure of abnormal perceptual experience (HP) in a non-clinical sample is associated with variability in the functional brain responses of the temporal cortex regions serving detecting and processing of voice signals. Considering the well-established roles of specific voice sensitive regions of the cerebral cortex, we aimed to determine if this putative relationship would be limited to specific subprocesses in hierarchical voice perception. As hypothesized, activity for voice versus non-voice processing correlated positively with HP only in the pSTG, a region

TABLE 2 Voice preference response [(Voice > Silence) > (Non-voice > Silence)] correlation with hallucination proneness results.

		ROI			LSHS		LSHS-Auditory	
Hem.	**Label**	**μ**	**SD**	**CI (95%)**	***r***	***p***	***r***	***p***
L	aSTG	1.189	0.479	0.203–0.434	0.120	0.576	0.178	0.406
	mSTG	1.505	0.586	0.997–1.380	−0.237	0.267	−0.024	0.915
	pSTG	1.511	0.560	1.271–1.740	−0.058	0.791	0.055	0.797
R	aSTG	1.019	0.452	0.838–1.200	0.266	0.208	0.165	0.440
	mSTG	1.295	0.515	1.089–1.501	−0.177	0.408	−0.033	0.882
	pSTG	1.213	0.406	1.051–1.375	0.470	*0.020	0.276	0.192
R	pMC	0.625	0.447	0.446–0.804	0.087	0.685	−0.103	0.635
L	IFC	0.319	0.288	0.204–0.434	−0.048	0.827	−0.025	0.911
R	IFC	0.293	0.323	0.164–0.422	0.231	0.277	0.134	0.534

ROI, region of interest; (a/m/p) STG, (anterior/middle/posterior) superior temporal gyrus; pMC, premotor cortex; IFC, inferior frontal cortex; μ, mean activation from contrast; SD, standard deviation; LSHS, Launay-Slade Hallucination Proneness scale; LSHS-Auditory, subset of 3 auditory items, *r* = correlation coefficient, Bonferroni-corrected significance level (*$p < 0.025$).

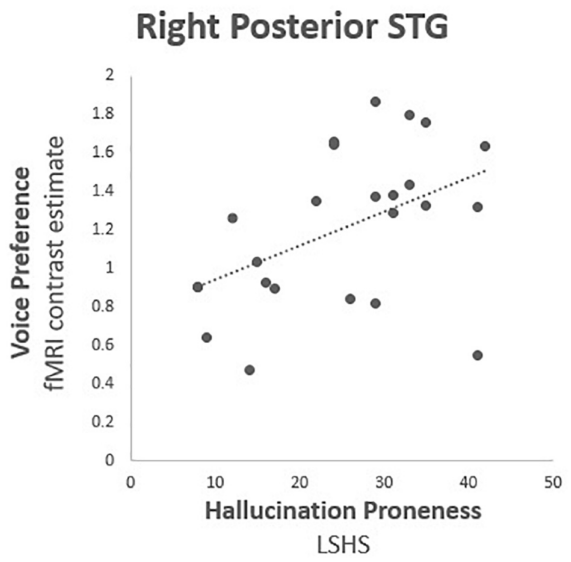


FIGURE 2
Hallucination proneness fMRI correlation analysis results: Right posterior superior temporal gyrus (BA 22; MNI 54, −34, 4), Voice preference = contrast estimate [(Voice > Silence) > (Non-voice > Silence)], LSHS = Launay Slade Hallucination Proneness scale. Correlation coefficient $r = 0.470$, $df = 25$, $p = 0.020$.

associated with the early processing of low-level acoustic features in complex auditory signals (i.e., Griffiths and Warren, 2004; Warren J. D. et al., 2005; Warren J. E. et al., 2005). Furthermore, this finding was restricted to the right hemisphere and therefore is likely linked to the processing of paralinguistic voice information (Belin et al., 2000; Formisano et al., 2008). Additionally, *post hoc* analysis revealed a negative correlation with HP in the right IFC for non-voice versus silence. Together, these findings may confirm that as the propensity to hallucinate increases, right posterior temporal lobe voice hypersensitivity increases and is accompanied by a decreased prefrontal response to non-vocal environmental sounds.

Hallucination proneness and hypersensitivity

Multiple neurocognitive mechanisms underlying hallucinations have been proposed. Most commonly, these theories have focused on describing the emergence and phenomenology of pathological voice hearing in patients with psychotic disorders such as schizophrenia (Allen et al., 2008; Hugdahl, 2015; Ćurčić-Blake et al., 2017). The most influential models describe atypical increases in brain activity in cortical voice regions. The current investigation was approached from the perspective of perceptual salience models claiming a central role of hypersensitivity to irrelevant sensory stimuli in auditory

regions (Menon and Uddin, 2010; Menon, 2011; Palaniyappan and Liddle, 2012; Uddin, 2015). Conversely, prominent self-monitoring models of hallucinatory experience describe increased activity as the result of insufficient suppression of sensory cortices during inner speech (Frith and Done, 1988; Weiss and Heckers, 1999; Tracy and Shergill, 2006; Allen et al., 2007a, 2008; Jones and Fernyhough, 2007b). According to this theory, the activation of speech production regions is required for the emergence of AVH. However, the current results demonstrate that variability in voice processing cortical regions in relation to HP exists without motor activity.

It is possible that theories proposing divergent involvement of speech production and perception mechanisms in AVH may be not mutually exclusive. Experiences of people who hallucinate are diverse. As theories of HP become more specific and concrete, they may become less well aligned with the phenomenology of the hallucinator. Therefore, hallucinatory experience might be best characterized by multiple subtypes, to which specific theories might apply better than others (Jones, 2010). For example, models describing the phenomenology of voice hearing ascribe the top-down contribution of intrusive memories and thoughts to the quality of false perception experiences (Hugdahl, 2015; Upthegrove et al., 2016; Bohlken et al., 2017; Ćurčić-Blake et al., 2017). A core abnormality in brain function central to the emergence of false perceptions likely rests in the interactive process of top-down predictions and bottom-up sensory input (Allen et al., 2008; Hugdahl, 2009, 2015; Kowalski et al., 2021). Regarding perceptual salience, bottom-up hypersensitivity to sensory input is congruent with established computation neuroscience accounts of predictive coding in false perceptions (Sterzer et al., 2018). Here, weighted top-down predictions and bottom-up explanations of sensation interact along a hierarchical network, constantly updating *via* Bayesian inference to form the most reliable percept (Friston, 2005, 2012; Fletcher and Frith, 2009; Feldman and Friston, 2010; Hohwy, 2017). When internal prediction signals are weighted too strongly, one "senses what they expect." Moreover, when the top-down input is too strong, the threshold for active perception may be reached under minimal sensory input. However, the self-monitoring theory posits a delayed or absent prediction signal resulting in increased activation of sensory cortical regions and is therefore in apparent conflict with the former account (Corlett et al., 2019; Leptourgos and Corlett, 2020). These expectations could operate on separate time scales, at different levels of the information processing hierarchy, or simply serve two different functions in hallucinations (Thakkar et al., 2021).

The role of perceptual salience in a multistage process leading to false perceptions has gathered substantial support in functional neuroimaging. Namely, research into large-scale functional brain networks has provided a resting-state hypothesis, outlining brain states serving as a predisposition for hallucinations, including voice hearing (Northoff and Qin, 2011; Northoff, 2014). While at rest, activation of the salience network, under conditions of irrelevant stimuli, may interrupt the Default Mode Network (DMN) and engage active sensory processing (Alderson-Day et al., 2015, 2016; Schmidt et al., 2015). The salience network therefore operates as a switch between the DMN and central executive network and how attention is directed toward incoming sensations, constituting a triple network model (TMN) subserving the advent of hallucinatory experience (Menon, 2011). Although we did not acquire behavioral data from the participants with ratings of perceived salience while listening to stimuli during scanning, we suggest that the change in brain activity that we observed in the right pSTG is indicative of the TMN in response to voice stimuli.

Hierarchical voice network processing

Voices are processed along a series of bilateral voice patches in the posterior, middle, and anterior STG. These temporal voice areas are reliably identified by a standardized TVA localizer task (Pernet et al., 2015). Participants with greater HP displayed increased right pSTG activation in response to vocal stimuli. Activity in this region may reflect sensitivity to low-level acoustic features during early stages of voice processing (Griffiths and Warren, 2004; Warren J. D. et al., 2005; Warren J. E. et al., 2005). Furthermore, the pSTG is not specialized for voice processing *per se*, and likely plays a broader role in extracting spectro-temporal acoustic features from complex sounds, of which voices are an example. However, activation in these regions preferentially responds to salient stimuli, such as voices, over and above other similarly complex environmental sounds (Pernet et al., 2007).

In terms of the salience hypothesis for hallucinatory experience, the assignment of salience to irrelevant, neutral, events must be considered in terms of the paralinguistic factors which may be involved. Indeed, the phenomenology of AVH is often marked by prominent paralinguistic features in the identity and emotional valence of the hallucinated speaker (Stephane et al., 2003; Larøi and Woodward, 2007; Badcock and Chhabra, 2013; McCarthy-Jones et al., 2014). Individuals who experience hallucinations often express difficulty in discerning the identity of veridical voices. For example, in schizophrenia patients who experience hallucinations, there is a bias to externalize voices to another person (Johns et al., 2001; Allen et al., 2007b; Mechelli et al., 2007; Pinheiro et al., 2016, 2017). Likewise, severity of AVH in patients is increasingly altered by emotional processing (Rossell and Boundy, 2005; Shea et al., 2007; Alba-Ferrara et al., 2013; Tseng et al., 2013). The role of salience may be influential in perceptions of speaker identity, as misattributions are more prevalent for emotional stimuli (Ditman and Kuperberg, 2005; Costafreda et al., 2008; Pinheiro et al., 2016, 2017). However, the effects of emotional valence in perceiving voice identity for people prone to false

perceptions of voices has not shown clear consensus (i.e., Brookwell et al., 2013). Comparisons of AVH severity in patients with schizophrenia with judgments of speaker identity have indicated an increasing proneness to externalize voices with negative content (Allen et al., 2004; Pinheiro et al., 2016). In non-clinical groups, the involvement of salient emotional features in voices is less clear. For example, higher levels of HP in the general population are not associated with atypical evaluation of emotional valence in words or vocalizations (Pinheiro et al., 2019). However, it has been indicated that non-clinical individuals prone to voice hearing require stronger emotional information to consider a stimulus as emotional (Amorim et al., 2021) or may allocate similar attention to voices irrespective of their emotional salience (Castiajo and Pinheiro, 2021). Future research is required into how variability in perceived salience of speaker-related features may affect processing in the hierarchical voice network and, in particular, how posterior STG activity related to HP may be influenced.

In addition to the TVA findings, the localizer task often provides a subset of extra-temporal regions indicating an extended voice processing network (Pernet et al., 2015). In our sample, extra-temporal peak activations were ascribed to bilateral inferior frontal and right hemisphere premotor cortex. Prefrontal involvement of the left IFC is commonly found in voice perception, with different subregions serving various functions. For example, the pars orbitalis is involved in processing semantic and emotional information (Belyk et al., 2017). Here, the left IFC peak was found in Broca's area, which has been theorized to represent mirror neuron activity which may be useful in guiding conversational turn-taking (Rizzolatti and Craighero, 2004; Grafton and Hamilton, 2007; Kilner et al., 2007). Likewise, precentral motor regions are involved in the perception and production of speech (Wilson et al., 2004; Pulvermüller et al., 2006; Cheung et al., 2016). This could explain speech production region activity sometimes reported during AVH (Jardri et al., 2011; Kühn and Gallinat, 2012; Zmigrod et al., 2016). However, self-monitoring theories take this as evidence for top-down inner speech signals guiding the perceived hallucinatory voice. Notably, transcranial direct-current stimulation targeting a fronto-parietal sensorimotor network is an effective treatment for the alleviation of AVH in patients with schizophrenia (Yang et al., 2019). In our *post hoc* analysis, the right IFC ROI shows an intriguing negative correlation to HP, however, only for non-voice sounds. The right IFC may serve a role in salience processing, for example in recognizing salient cues in voice signals (Johnstone et al., 2006; Bestelmeyer et al., 2012; Charest et al., 2013; Johns et al., 2015; Johnson et al., 2021). Additionally, this area shares a high functional integration with temporal regions serving voice perception and may assist successful voice recognition (Aglieri et al., 2018). Although this finding is difficult to interpret on its own, it may indicate a decrease in salience attribution for environmental sounds during a voice perception task. This may indicate not only an HP-related salience bias affecting the sensitivity of cortical responses to voice sounds, but also a general bias away from non-voice sounds between hypersalient responses to intermittent voice stimuli.

Limitations and recommendations

We identify a number of limitations within the current study and provide suggestions for future research. First, although the use of the established TVA localizer task facilitated the testing of our hypotheses regarding an early hypersensitivity to voice sounds, it did not preclude further investigation into how more complex stages of the voice processing hierarchy may relate to HP. Specifically, BOLD responses from this task are averaged across the trials containing different types of voice stimuli. This implies that signals extracted from ROIs serving different functional roles in voice processing, e.g., emotion or identity, do not represent the processing of specific features, but rather constitute a generalized voice detection signal. Second, in this study, behavioral measures of perceived stimulus salience were not collected. Therefore, interpretations of a salience bias attributed to increased functional brain responses cannot be directly linked to the subjective perception of the participants. Third, participants in the current study were sampled from a relatively homogenous sample of university students, similar in age, ethnicity, and cultural backgrounds. Due to the uneven distribution of environmental risk factors for psychotic symptoms throughout the population (Johns and van Os, 2001, DeRosse and Karlsgodt, 2015; Baumeister et al., 2017), our sample may unintentionally capture a set of protective factors. To address these limitations in future studies, we suggest a two-step procedure using a novel task that systematically varies paralinguistic voice features. This may allow investigations into how hierarchical processing downstream of initial HP-related hypersensitivity may influence responses to the perceived emotion or identity of the speaker. Furthermore, behavioral appraisals of perceived salience may be included to compare fMRI response patterns and HP scores. Finally, subsequent research may benefit from an increased sample size and diversity, including a structured collection of additional demographic data and associated environmental risk factors as possible covariates for HP-related brain changes.

Conclusion

We observed that HP is positively correlated with increased activation in the right pSTG in response to passively heard voices. This suggests a hypersensitivity associated with a propensity to hallucinate in a region of the brain which extracts low-level acoustic features from complex auditory signals. The right pSTG comprises the early processing of

voice signals along the paralinguistic information pathway of the cortical voice processing network. We propose that this increases activity in response to voices represents a perceptual salience bias as a precursor for the emergence of hallucinations. This interpretation is in line with functional network models that posit abnormal engagement of a salience network during irrelevant stimulus exposure as the underlying neurocognitive mechanism of false perceptions. Furthermore, the current findings conflict with self-monitoring accounts of inner speech models that propose a critical role of voice production regions in the inception of AVH. We have demonstrated that HP is associated with right pSTG activation driven by external auditory signals. Although we do not reject self-monitoring accounts, we suggest that a state of cortical hypersensitivity to irrelevant sensory input may be the first step in the emergence of a hallucinatory experience, possibly followed by the influence of top-down signals such as inner speech, memory, and thought that together contribute to the phenomenology of AVH.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by the Ethical Review Committee of the Faculty of Psychology and Neuroscience at Maastricht University. The patients/participants provided their written informed consent to participate in this study.

Author contributions

JJ conceptualized and carried out experiment, performed the analyses, and wrote manuscript with input from all authors. MB verified the analytical methods. SK, AP, MS, and MB conceptualized and interpreted the results. SK and AP provided the original idea for project and secured funding. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by Fundação para a Ciência e a Tecnologia, Grant/Award Number: PTDC/MHC-PCN/0101/2014 and BIAL Foundation, Grant/Award Number: BIAL 238/16.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnhum.2022.859731/full#supplementary-material

References

Aglieri, V., Chaminade, T., Takerkart, S., and Belin, P. (2018). Functional connectivity within the voice perception network and its behavioural relevance. *Neuroimage* 183, 356–365. doi: 10.1016/j.neuroimage.2018.08.011

Alba-Ferrara, L., De Erausquin, G. A., Hirnstein, M., Weis, S., and Hausmann, M. (2013). Emotional prosody modulates attention in schizophrenia patients with hallucinations. *Front. Hum. Neurosci.* 7:59. doi: 10.3389/fnhum.2013.00059

Alderson-Day, B., Diederen, K., Fernyhough, C., Ford, J. M., Horga, G., Margulies, D. S., et al. (2016). Auditory hallucinations and the brain's resting-state networks: findings and methodological observations. *Schizophr. Bull.* 42, 1110–1123. doi: 10.1093/schbul/sbw078

Alderson-Day, B., McCarthy-Jones, S., and Fernyhough, C. (2015). Hearing voices in the resting brain: a review of intrinsic functional connectivity research on auditory verbal hallucinations. *Neurosci. Biobehav. Rev.* 55, 78–87. doi: 10.1016/j.neubiorev.2015.04.016

Allen, P. P., Johns, L. C., Fu, C. H., Broome, M. R., Vythelingum, G. N., and McGuire, P. K. (2004). Misattribution of external speech in patients with hallucinations and delusions. *Schizophr. Res.* 69, 277–287. doi: 10.1016/j.schres.2003.09.008

Allen, P., Aleman, A., and Mcguire, P. K. (2007a). Inner speech models of auditory verbal hallucinations: evidence from behavioural and neuroimaging studies. *Int. Rev. Psychiatry* 19, 407–415. doi: 10.1080/09540260701486498

Allen, P., Amaro, E., Fu, C. H., Williams, S. C., Brammer, M. J., Johns, L. C., et al. (2007b). Neural correlates of the misattribution of speech in schizophrenia. *Br. J. Psychiatry* 190, 162–169. doi: 10.1192/bjp.bp.106.025700

Allen, P., Larøi, F., McGuire, P. K., and Aleman, A. (2008). The hallucinating brain: a review of structural and functional neuroimaging studies of hallucinations. *Neurosci. Biobehav. Rev.* 32, 175–191. doi: 10.1016/j.neubiorev.2007.07.012

Amorim, M., Roberto, M. S., Kotz, S. A., and Pinheiro, A. P. (2021). The perceived salience of vocal emotions is dampened in non-clinical auditory verbal hallucinations. *Cogn. Neuropsychiatry* 27, 169–182. doi: 10.1080/13546805.2021.1949972

Anthony, D. (2004). The cognitive neuropsychiatry of auditory verbal hallucinations: an overview. *Cogn. Neuropsychiatry* 9, 107–123. doi: 10.1080/13546800344000183

Badcock, J. C., and Chhabra, S. (2013). Voices to reckon with: perceptions of voice identity in clinical and non-clinical voice hearers. *Front. Hum. Neurosci.* 7:114. doi: 10.3389/fnhum.2013.00114

Bauer, S. M., Schanda, H., Karakula, H., Olajossy-Hilkesberger, L., Rudaleviciene, P., Okribelashvili, N., et al. (2011). Culture and the prevalence of hallucinations in schizophrenia. *Compr. Psychiatry* 52, 319–325. doi: 10.1016/j.comppsych.2010.06.008

Baumeister, D., Sedgwick, O., Howes, O., and Peters, E. (2017). Auditory verbal hallucinations and continuum models of psychosis: a systematic review of the healthy voice-hearer literature. *Clin. Psychol. Rev.* 51, 125–141. doi: 10.1016/j.cpr.2016.10.010

Beavan, V., Read, J., and Cartwright, C. (2011). The prevalence of voice-hearers in the general population: a literature review. *J. Ment. Health* 20, 281–292. doi: 10.3109/09638237.2011.562262

Belin, P., and Zatorre, R. J. (2003). Adaptation to speaker's voice in right anterior temporal lobe. *Neuroreport* 14, 2105–2109. doi: 10.1097/00001756-200311140-00019

Belin, P., Fecteau, S., and Bedard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends Cogn. Sci.* 8, 129–135. doi: 10.1016/j.tics.2004.01.008

Belin, P., Zatorre, R. J., Hoge, R., Evans, A. C., and Pike, B. (1999). Event-related fMRI of the auditory cortex. *Neuroimage* 10, 417–429. doi: 10.1006/nimg.1999.0480

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., and Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature* 403, 309–312. doi: 10.1038/35002078

Belyk, M., Brown, S., Lim, J., and Kotz, S. A. (2017). Convergence of semantics and emotional expression within the IFG pars orbitalis. *Neuroimage* 156, 240–248. doi: 10.1016/j.neuroimage.2017.04.020

Bentall, R. P. (1990). The illusion of reality: a review and integration of psychological research on hallucinations. *Psychol. Bull.* 107:82. doi: 10.1037/0033-2909.107.1.82

Bestelmeyer, P. E., Latinus, M., Bruckert, L., Rouger, J., Crabbe, F., and Belin, P. (2012). Implicitly perceived vocal attractiveness modulates prefrontal cortex activity. *Cereb. Cortex* 22, 1263–1270. doi: 10.1093/cercor/bhr204

Bohlken, M. M., Hugdahl, K., and Sommer, I. E. C. (2017). Auditory verbal hallucinations: neuroimaging and treatment. *Psychol. Med.* 47, 199–208. doi: 10.1017/S003329171600115X

Brett, M., Anton, J., Valabregue, R., and Poline, J. (2002). Region of interest analysis using an SPM toolbox [abstract]. *Paper Presented at the 8th International Conference on Functional Mapping of the Human Brain*, Sendai.

Brookwell, M. L., Bentall, R. P., and Varese, F. (2013). Externalizing biases and hallucinations in source-monitoring, self-monitoring and signal detection studies: a meta-analytic review. *Psychol. Med.* 43, 2465–2475. doi: 10.1017/S0033291712002760

Castiajo, P., and Pinheiro, A. P. (2021). Acoustic salience in emotional voice perception and its relationship with hallucination proneness. *Cogn. Affect. Behav. Neurosci.* 21, 412–425. doi: 10.3758/s13415-021-00864-2

Charest, I., Pernet, C., Latinus, M., Crabbe, F., and Belin, P. (2013). Cerebral processing of voice gender studied using a continuous carryover fMRI design. *Cereb. Cortex* 23, 958–966. doi: 10.1093/cercor/bhs090

Cheung, C., Hamilton, L. S., Johnson, K., and Chang, E. F. (2016). The auditory representation of speech sounds in human motor cortex. *Elife* 5:e12577. doi: 10.7554/eLife.12577

Copolov, D. L., Seal, M. L., Maruff, P., Ulusoy, R., Wong, M. T., Tochon-Danguy, H. J., et al. (2003). Cortical activation associated with the experience of auditory hallucinations and perception of human speech in schizophrenia: a PET correlation study. *Psychiatry Res. Neuroimaging* 122, 139–152. doi: 10.1016/S0925-4927(02)00121-X

Corlett, P. R., Horga, G., Fletcher, P. C., Alderson-Day, B., Schmack, K., and Powers, A. R. III (2019). Hallucinations and strong priors. *Trends Cogn. Sci.* 23, 114–127. doi: 10.1016/j.tics.2018.12.001

Costafreda, S. G., Brébion, G., Allen, P., McGuire, P. K., and Fu, C. H. (2008). Affective modulation of external misattribution bias in source monitoring in schizophrenia. *Psychol. Med.* 38, 821–824. doi: 10.1017/S0033291708003243

Ćurčić-Blake, B., Ford, J. M., Hubl, D., Orlov, N. D., Sommer, I. E., Waters, F., et al. (2017). Interaction of language, auditory and memory brain networks in auditory verbal hallucinations. *Prog. Neurobiol.* 148, 1–20. doi: 10.1016/j.pneurobio.2016.11.002

Daalman, K., Boks, M. P., Diederen, K. M., de Weijer, A. D., Blom, J. D., Kahn, R. S., et al. (2011). The same or different? A phenomenological comparison of auditory verbal hallucinations in healthy and psychotic individuals. *J. Clin. Psychiatry* 72, 320–325. doi: 10.4088/JCP.09m05797yel

Daalman, K., Verkooijen, S., Derks, E. M., Aleman, A., and Sommer, I. E. C. (2012). The influence of semantic top-down processing in auditory verbal hallucinations. *Schizophr. Res.* 139, 82–86. doi: 10.1016/j.schres.2012.06.005

de Boer, J. N., Heringa, S. M., van Dellen, E., Wijnen, F. N. K., and Sommer, I. E. C. (2016). A linguistic comparison between auditory verbal hallucinations in patients with a psychotic disorder and in nonpsychotic individuals: not just what the voices say, but how they say it. *Brain Lang.* 162, 10–18. doi: 10.1016/j.bandl.2016.07.011

de Leede-Smith, S., and Barkus, E. (2013). A comprehensive review of auditory verbal hallucinations: lifetime prevalence, correlates and mechanisms in healthy and clinical individuals. *Front. Hum. Neurosci.* 7:367. doi: 10.3389/fnhum.2013.00367

Deneve, S., and Jardri, R. (2016). Circular inference: mistaken belief, misplaced trust. *Curr. Opin. Behav. Sci.* 11, 40–48. doi: 10.1016/j.cobeha.2016.04.001

DeRosse, P., and Karlsgodt, K. H. (2015). Examining the psychosis continuum. *Curr. Behav. Neurosci. Rep.* 2, 80–89. doi: 10.1007/s40473-015-0040-7

Ditman, T., and Kuperberg, G. R. (2005). A source-monitoring account of auditory verbal hallucinations in patients with schizophrenia. *Harv. Rev. Psychiatry* 13, 280–299. doi: 10.1080/10673220500326391

Escartí, M. J., de la Iglesia-Vayá, M., Martí-Bonmatí, L., Robles, M., Carbonell, J., Lull, J. J., et al. (2010). Increased amygdala and parahippocampal gyrus activation in schizophrenic patients with auditory hallucinations: an fMRI study using independent component analysis. *Schizophr. Res.* 117, 31–41. doi: 10.1016/j.schres.2009.12.028

Feinberg, I. (1978). Efference copy and corollary discharge: implications for thinking and its disorders. *Schizophr. Bull.* 4:636. doi: 10.1093/schbul/4.4.636

Feldman, H., and Friston, K. (2010). Attention, uncertainty, and free-energy. *Front. Hum. Neurosci.* 4:215. doi: 10.3389/fnhum.2010.00215

Fletcher, P. C., and Frith, C. D. (2009). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nat. Rev. Neurosci.* 10, 48–58. doi: 10.1038/nrn2536

Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C. R., and Collins, D. L. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *Neuroimage* 47:S102. doi: 10.1016/S1053-8119(09)70884-5

Ford, J. M., and Mathalon, D. H. (2019). Efference copy, corollary discharge, predictive coding, and psychosis. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* 4, 764–767. doi: 10.1016/j.bpsc.2019.07.005

Formisano, E., De Martino, F., Bonte, M., and Goebel, R. (2008). " Who" is saying" what"? Brain-based decoding of human voice and speech. *Science* 322, 970–973. doi: 10.1126/science.1164318

Friston, K. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. B Biol. Sci.* 360, 815–836.

Friston, K. (2012). Prediction, perception and agency. *Int. J. Psychophysiol.* 83, 248–252. doi: 10.1016/j.ijpsycho.2011.11.014

Frith, C. D., and Done, D. J. (1988). Towards a neuropsychology of schizophrenia. *Br. J. Psychiatry* 153, 437–443. doi: 10.1192/bjp.153.4.437

Grafton, S. T., and Hamilton, A. F. D. C. (2007). Evidence for a distributed hierarchy of action representation in the brain. *Hum. Mov. Sci.* 26, 590–616. doi: 10.1016/j.humov.2007.05.009

Gregory, D. (2016). Inner speech, imagined speech, and auditory verbal hallucinations. *Rev. Philos. Psychol.* 7, 653–673. doi: 10.1007/s13164-015-0274-z

Griffiths, T. D., and Warren, J. D. (2004). What is an auditory object? *Nat. Rev. Neurosci.* 5, 887–892. doi: 10.1038/nrn1538

Hall, D. A., Haggard, M. P., Akeroyd, M. A., Palmer, A. R., Summerfield, A. Q., Elliott, M. R., et al. (1999). "Sparse" temporal sampling in auditory fMRI. *Hum. Brain Mapp.* 7, 213–223. doi: 10.1002/(SICI)1097-019319997:3<213::AID-HBM5<3.0.CO;2-N

Hohwy, J. (2017). Priors in perception: top-down modulation, Bayesian perceptual learning rate, and prediction error minimization. *Conscious. Cogn.* 47, 75–85. doi: 10.1016/j.concog.2016.09.004

Hugdahl, K. (2009). "Hearing voices": auditory hallucinations as failure of top-down control of bottom-up perceptual processes. *Scand. J. Psychol.* 50, 553–560. doi: 10.1111/j.1467-9450.2009.00775.x

Hugdahl, K. (2015). Auditory hallucinations: a review of the ERC "VOICE" project. *World J. Psychiatry* 5:193. doi: 10.5498/wjp.v5.i2.193

Hugdahl, K., and Sommer, I. E. (2018). Auditory verbal hallucinations in schizophrenia from a levels of explanation perspective. *Schizophr. Bull.* 44, 234–241. doi: 10.1093/schbul/sbx142

Jardri, R., Hugdahl, K., Hughes, M., Brunelin, J., Waters, F., Alderson-Day, B., et al. (2016). Are hallucinations due to an imbalance between excitatory and inhibitory influences on the brain? *Schizophr. Bull.* 42, 1124–1134. doi: 10.1093/schbul/sbw075

Jardri, R., Pouchet, A., Pins, D., and Thomas, P. (2011). Cortical activations during auditory verbal hallucinations in schizophrenia: a coordinate-based meta-analysis. *Am. J. Psychiatry* 168, 73–81. doi: 10.1176/appi.ajp.2010.09101522

Johns, L. C. (2005). Hallucinations in the general population. *Curr. Psychiatry Rep.* 7, 162–167. doi: 10.1006/nimg.2002.1132

Johns, L. C., and van Os, J. (2001). The continuity of psychotic experiences in the general population. *Clin. Psychol. Rev.* 21, 1125–1141. doi: 10.1016/S0272-7358(01)00103-9

Johns, A. B., Farrall, A. J., Belin, P., and Pernet, C. R. (2015). Hemispheric association and dissociation of voice and speech information processing in stroke. *Cortex* 71, 232–239. doi: 10.1016/j.cortex.2015.07.004

Johns, L. C., Kompus, K., Connell, M., Humpston, C., Lincoln, T. M., Longden, E., et al. (2014). Auditory verbal hallucinations in persons with and without a need for care. *Schizophr. Bull.* 40(Suppl. 4), S255–S264. doi: 10.1093/schbul/sbu005

Johns, L. C., Rossell, S., Frith, C., Ahmad, F., Hemsley, D., Kuipers, E., et al. (2001). Verbal self-monitoring and auditory verbal hallucinations in patients with schizophrenia. *Psychol. Med.* 31, 705–715. doi: 10.1017/S0033291701003774

Jones, S. R. (2010). Do we need multiple models of auditory verbal hallucinations? Examining the phenomenological fit of cognitive and neurological models. *Schizophr. Bull.* 36, 566–575. doi: 10.1093/schbul/sbn129

Jones, S. R., and Fernyhough, C. (2007a). Thought as action: inner speech, self-monitoring, and auditory verbal hallucinations. *Conscious. Cogn.* 16, 391–399. doi: 10.1016/j.concog.2005.12.003

Jones, S. R., and Fernyhough, C. (2007b). Neural correlates of inner speech and auditory verbal hallucinations: a critical review and theoretical integration. *Clin. Psychol. Rev.* 27, 140–154. doi: 10.1016/j.cpr.2006.10.001

Johnson, J. F., Belyk, M., Schwartze, M., Pinheiro, A. P., and Kotz, S. A. (2021). Expectancy changes the self-monitoring of voice identity. *Eur. J. Neurosci.* 53, 2681–2695. doi: 10.1111/ejn.15162

Johnstone, T., Van Reekum, C. M., Oakes, T. R., and Davidson, R. J. (2006). The voice of emotion: an FMRI study of neural responses to angry and happy vocal expressions. *Soc. Cogn. Affect. Neurosci.* 1, 242–249. doi: 10.1093/scan/nsl027

Kelleher, I., and Cannon, M. (2011). Psychotic-like experiences in the general population: characterizing a high-risk group for psychosis. *Psychol. Med.* 41, 1–6. doi: 10.1017/S0033291710001005

Kelleher, I., Jenner, J. A., and Cannon, M. (2010). Psychotic symptoms in the general population–an evolutionary perspective. *Br. J. Psychiatry* 197, 167–169. doi: 10.1192/bjp.bp.109.076018

Kilner, J. M., Friston, K. J., and Frith, C. D. (2007). Predictive coding: an account of the mirror neuron system. *Cogn. Process.* 8, 159–166. doi: 10.1007/s10339-007-0170-2

Kompus, K., Løberg, E. M., Posserud, M. B., and Lundervold, A. J. (2015). Prevalence of auditory hallucinations in Norwegian adolescents: results from a population-based study. *Scand. J. Psychol.* 56, 391–396. doi: 10.1111/sjop.12219

Kompus, K., Westerhausen, R., and Hugdahl, K. (2011). The "paradoxical" engagement of the primary auditory cortex in patients with auditory verbal hallucinations: a meta-analysis of functional neuroimaging studies. *Neuropsychologia* 49, 3361–3369. doi: 10.1016/j.neuropsychologia.2011.08.010

Kowalski, J., Aleksandrowicz, A., Dąbkowska, M., and Gawęda, Ł (2021). Neural correlates of aberrant salience and source monitoring in schizophrenia and at-risk mental states—a systematic review of fMRI studies. *J. Clin. Med.* 10:4126. doi: 10.3390/jcm10184126

Kühn, S., and Gallinat, J. (2012). Quantitative meta-analysis on state and trait aspects of auditory verbal hallucinations in schizophrenia. *Schizophr. Bull.* 38, 779–786. doi: 10.1093/schbul/sbq152

Larøi, F., and Van Der Linden, M. (2005). Nonclinical participants' reports of hallucinatory experiences. *Can. J. Behav. Sci.* 37:33. doi: 10.1037/h0087243

Larøi, F., and Woodward, T. S. (2007). Hallucinations from a cognitive perspective. *Harv. Rev. Psychiatry* 15, 109–117. doi: 10.1080/10673220701401993

Larøi, F., Marczewski, P., and Van der Linden, M. (2004). Further evidence of the multi-dimensionality of hallucinatory predisposition: factor structure of a modified version of the Launay-Slade hallucinations scale in a normal sample. *Eur. Psychiatry* 19, 15–20. doi: 10.1016/S0924-9338(03)00028-2

Larøi, F., Sommer, I. E., Blom, J. D., Fernyhough, C., Ffytche, D. H., Hugdahl, K., et al. (2012). The characteristic features of auditory verbal hallucinations in clinical and nonclinical groups: state-of-the-art overview and future directions. *Schizophr. Bull.* 38, 724–733. doi: 10.1093/schbul/sbs061

Lavan, N., Burton, A. M., Scott, S. K., and McGettigan, C. (2019). Flexible voices: identity perception from variable vocal signals. *Psychonom. Bull. Rev.* 26, 90–102. doi: 10.3758/s13423-018-1497-7

Leptourgos, P., and Corlett, P. R. (2020). Embodied predictions, agency, and psychosis. *Front. Big Data* 3:27. doi: 10.3389/fdata.2020.00027

Leptourgos, P., Denève, S., and Jardri, R. (2017). Can circular inference relate the neuropathological and behavioral aspects of schizophrenia? *Curr. Opin. Neurobiol.* 46, 154–161. doi: 10.1016/j.conb.2017.08.012

Linscott, R. J., and Van Os, J. (2013). An updated and conservative systematic review and meta-analysis of epidemiological evidence on psychotic experiences in children and adults: on the pathway from proneness to persistence to dimensional expression across mental disorders. *Psychol. Med.* 43:1133. doi: 10.1017/S0033291712001626

Martí-Bonmatí, L., Lull, J. J., García-Martí, G., Aguilar, E. J., Moratal-Pérez, D., Poyatos, C., et al. (2007). Chronic auditory hallucinations in schizophrenic patients: MR analysis of the coincidence between functional and morphologic abnormalities. *Radiology* 244, 549–556. doi: 10.1148/radiol.2442060727

McCarthy, P. (2022). *FSLeyes (1.4.0)*. Zenodo. doi: 10.5281/zenodo.6511596

McCarthy-Jones, S., Trauer, T., Mackinnon, A., Sims, E., Thomas, N., and Copolov, D. L. (2014). A new phenomenological survey of auditory hallucinations: evidence for subtypes and implications for theory and practice. *Schizophr. Bull.* 40, 231–235. doi: 10.1093/schbul/sbs156

McGrath, J. J., Saha, S., Al-Hamzawi, A., Alonso, J., Bromet, E. J., Bruffaerts, R., et al. (2015). Psychotic experiences in the general population: a cross-national analysis based on 31 261 respondents from 18 countries. *JAMA Psychiatry* 72, 697–705. doi: 10.1001/jamapsychiatry.2015.0575

Mechelli, A., Allen, P., Amaro, E. Jr., Fu, C. H., Williams, S. C., Brammer, M. J., et al. (2007). Misattribution of speech and impaired connectivity in patients with auditory verbal hallucinations. *Hum. Brain Mapp.* 28, 1213–1222. doi: 10.1002/hbm.20341

Menon, V. (2011). Large-scale brain networks and psychopathology: a unifying triple network model. *Trends Cogn. Sci.* 15, 483–506. doi: 10.1016/j.tics.2011.08.003

Menon, V., and Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain Struct. Funct.* 214, 655–667. doi: 10.1007/s00429-010-0262-0

Miyata, J. (2019). Toward integrated understanding of salience in psychosis. *Neurobiol. Dis.* 131:104414. doi: 10.1016/j.nbd.2019.03.002

Nakamura, K., Kawashima, R., Sugiura, M., Kato, T., Nakamura, A., Hatano, K., et al. (2001). Neural substrates for recognition of familiar voices: a PET study. *Neuropsychologia* 39, 1047–1054. doi: 10.1016/S0028-3932(01)00037-9

Northoff, G. (2014). Are auditory hallucinations related to the brain's resting state activity? A 'neurophenomenal resting state hypothesis'. *Clin. Psychopharmacol. Neurosci.* 12:189. doi: 10.9758/cpn.2014.12.3.189

Northoff, G., and Qin, P. (2011). How can the brain's resting state activity generate hallucinations? A 'resting state hypothesis' of auditory verbal hallucinations. *Schizophr. Res.* 127, 202–214. doi: 10.1016/j.schres.2010.11.009

Palaniyappan, L., and Liddle, P. F. (2012). Does the salience network play a cardinal role in psychosis? An emerging hypothesis of insular dysfunction. *J. Psychiatry Neurosci.* 37, 17–27. doi: 10.1503/jpn.100176

Parellada, E., Lomena, F., Font, M., Pareto, D., Gutierrez, F., Simo, M., et al. (2008). Fluordeoxyglucose-PET study in first-episode schizophrenic patients during the hallucinatory state, after remission and during linguistic–auditory activation. *Nuclear Med. Commun.* 29, 894–900. doi: 10.1097/MNM.0b013e328302cd10

Pernet, C., Schyns, P. G., and Demonet, J. F. (2007). Specific, selective or preferential: comments on category specificity in neuroimaging. *Neuroimage* 35, 991–997. doi: 10.1016/j.neuroimage.2007.01.017

Pernet, C. R., McAleer, P., Latinus, M., Gorgolewski, K. J., Charest, I., Bestelmeyer, P. E., et al. (2015). The human voice areas: spatial organization and inter-individual variability in temporal and extra-temporal cortices. *Neuroimage* 119, 164–174. doi: 10.1016/j.neuroimage.2015.06.050

Pinheiro, A. P., Farinha-Fernandes, A., Roberto, M. S., and Kotz, S. A. (2019). Self-voice perception and its relationship with hallucination predisposition. *Cogn. Neuropsychiatry* 24, 237–255. doi: 10.1080/13546805.2019.1621159

Pinheiro, A. P., Rezaii, N., Rauber, A., and Niznikiewicz, M. (2016). Is this my voice or yours? The role of emotion and acoustic quality in self-other voice discrimination in schizophrenia. *Cogn. Neuropsychiatry* 21, 335–353. doi: 10.1080/13546805.2016.1208611

Pinheiro, A. P., Rezaii, N., Rauber, A., Nestor, P. G., Spencer, K. M., and Niznikiewicz, M. (2017). Emotional self–other voice processing in schizophrenia and its relationship with hallucinations: ERP evidence. *Psychophysiology* 54, 1252–1265. doi: 10.1111/psyp.12880

Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., and Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proc. Natl. Acad. Sci. U.S.A.* 103, 7865–7870. doi: 10.1073/pnas.0509989103

Reininghaus, U., Kempton, M. J., Valmaggia, L., Craig, T. K., Garety, P., Onyejiaka, A., et al. (2016). Stress sensitivity, aberrant salience, and threat anticipation in early psychosis: an experience sampling study. *Schizophr. Bull.* 42, 712–722. doi: 10.1093/schbul/sbv190

Rizzolatti, G., and Craighero, L. (2004). The mirror-neuron system. *Annu. Rev. Neurosci.* 27, 169–192. doi: 10.1146/annurev.neuro.27.070203.144230

Rollins, C. P., Garrison, J. R., Simons, J. S., Rowe, J. B., O'Callaghan, C., Murray, G. K., et al. (2019). Meta-analytic evidence for the plurality of mechanisms in transdiagnostic structural MRI studies of hallucination status. *EClinicalMedicine* 8, 57–71. doi: 10.1016/j.eclinm.2019.01.012

Rossell, S. L., and Boundy, C. L. (2005). Are auditory–verbal hallucinations associated with auditory affective processing deficits? *Schizophr. Res.* 78, 95–106. doi: 10.1016/j.schres.2005.06.002

Rousseeuw, P. J., and Hubert, M. (2011). Robust statistics for outlier detection. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 1, 73–79. doi: 10.1002/widm.2

Schirmer, A., and Kotz, S. A. (2006). Beyond the right hemisphere: brain mechanisms mediating vocal emotional processing. *Trends Cogn. Sci.* 10, 24–30. doi: 10.1016/j.tics.2005.11.009

Schmidt, A., Diwadkar, V. A., Smieskova, R., Harrisberger, F., Lang, U. E., McGuire, P., et al. (2015). Approaching a network connectivity-driven classification of the psychosis continuum: a selective review and suggestions for future research. *Front. Hum. Neurosci.* 8:1047. doi: 10.3389/fnhum.2014.01047

Shea, T. L., Sergejew, A. A., Burnham, D., Jones, C., Rossell, S. L., Copolov, D. L., et al. (2007). Emotional prosodic processing in auditory hallucinations. *Schizophr. Res.* 90, 214–220. doi: 10.1016/j.schres.2006.09.021

Simons, C. J., Tracy, D. K., Sanghera, K. K., O'Daly, O., Gilleen, J., Krabbendam, L., et al. (2010). Functional magnetic resonance imaging of inner speech in schizophrenia. *Biol. Psychiatry* 67, 232–237. doi: 10.1016/j.biopsych.2009.09.007

Stephane, M., Thuras, P., Nasrallah, H., and Georgopoulos, A. P. (2003). The internal structure of the phenomenology of auditory verbal hallucinations. *Schizophr. Res.* 61, 185–193. doi: 10.1016/S0920-9964(03)00013-6

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., et al. (2018). The predictive coding account of psychosis. *Biol. Psychiatry* 84, 634–643. doi: 10.1016/j.biopsych.2018.05.015

Swiney, L., and Sousa, P. (2014). A new comparator account of auditory verbal hallucinations: how motor prediction can plausibly contribute to the sense of agency for inner speech. *Front. Hum. Neurosci.* 8:675. doi: 10.3389/fnhum.2014.00675

Thakkar, K. N., Mathalon, D. H., and Ford, J. M. (2021). Reconciling competing mechanisms posited to underlie auditory verbal hallucinations. *Philos. Trans. R. Soc. B* 376:20190702. doi: 10.1098/rstb.2019.0702

Tracy, D. K., and Shergill, S. S. (2006). Imaging auditory hallucinations in schizophrenia. *Acta Neuropsychiatr.* 18, 71–78. doi: 10.1111/j.1601-5215.2006.00129.x

Tseng, H. H., Chen, S. H., Liu, C. M., Howes, O., Huang, Y. L., Hsieh, M. H., et al. (2013). Facial and prosodic emotion recognition deficits associate with specific clusters of psychotic symptoms in schizophrenia. *PLoS One* 8:e66571. doi: 10.1371/journal.pone.0066571

Uddin, L. Q. (2015). Salience processing and insular cortical function and dysfunction. *Nat. Rev. Neurosci.* 16, 55–61. doi: 10.1038/nrn3857

Upthegrove, R., Broome, M. R., Caldwell, K., Ives, J., Oyebode, F., and Wood, S. J. (2016). Understanding auditory verbal hallucinations: a systematic review of current evidence. *Acta Psychiatr. Scand.* 133, 352–367. doi: 10.1111/acps.12531

van Lutterveld, R., Diederen, K. M., Koops, S., Begemann, M. J., and Sommer, I. E. (2013). The influence of stimulus detection on activation patterns during auditory hallucinations. *Schizophr. Res.* 145, 27–32. doi: 10.1016/j.schres.2013.01.004

Van Os, J., Hanssen, M., Bijl, R. V., and Ravelli, A. (2000). Strauss (1969) revisited: a psychosis continuum in the general population? *Schizophr. Res.* 45, 11–20. doi: 10.1016/S0920-9964(99)00224-8

Verdoux, H., and van Os, J. (2002). Psychotic symptoms in non-clinical populations and the continuum of psychosis. *Schizophr. Res.* 54, 59–65. doi: 10.1016/S0920-9964(01)00352-8

von Kriegstein, K., and Giraud, A. L. (2004). Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *Neuroimage* 22, 948–955. doi: 10.1016/j.neuroimage.2004.02.020

von Kriegstein, K., Eger, E., Kleinschmidt, A., and Giraud, A. L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Cogn. Brain Res.* 17, 48–55. doi: 10.1016/S0926-6410(03)00079-X

Warren, J. D., Jennings, A. R., and Griffiths, T. D. (2005). Analysis of the spectral envelope of sounds by the human brain. *Neuroimage* 24, 1052–1057. doi: 10.1016/j.neuroimage.2004.10.031

Warren, J. E., Wise, R. J., and Warren, J. D. (2005). Sounds do-able: auditory–motor transformations and the posterior temporal plane. *Trends Neurosci.* 28, 636–643. doi: 10.1016/j.tins.2005.09.010

Waters, F., and Fernyhough, C. (2017). Hallucinations: a systematic review of points of similarity and difference across diagnostic classes. *Schizophr. Bull.* 43, 32–43. doi: 10.1093/schbul/sbw132

Waters, F., Allen, P., Aleman, A., Fernyhough, C., Woodward, T. S., Badcock, J. C., et al. (2012). Auditory hallucinations in schizophrenia and nonschizophrenia populations: a review and integrated model of cognitive mechanisms. *Schizophr. Bull.* 38, 683–693. doi: 10.1093/schbul/sbs045

Weiss, A. P., and Heckers, S. (1999). Neuroimaging of hallucinations: a review of the literature. *Psychiatry Res. Neuroimaging* 92, 61–74. doi: 10.1016/S0925-4927(99)00041-4

Wilson, S. M., Saygin, A. P., Sereno, M. I., and Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nat. Neurosci.* 7, 701–702. doi: 10.1038/nn1263

Woodruff, P. W., Wright, I. C., Bullmore, E. T., Brammer, M., Howard, R. J., Williams, S. C., et al. (1997). Auditory hallucinations and the temporal cortical response to speech in schizophrenia: a functional magnetic resonance imaging study. *Am. J. Psychiatry* 154, 1676–1682. doi: 10.1176/ajp.154.12.1676

Yang, F., Fang, X., Tang, W., Hui, L., Chen, Y., Zhang, C., et al. (2019). Effects and potential mechanisms of transcranial direct current stimulation (tDCS) on auditory hallucinations: a meta-analysis. *Psychiatry Res.* 273, 343–349. doi: 10.1016/j.psychres.2019.01.059

Zhang, Y., Ding, Y., Huang, J., Zhou, W., Ling, Z., Hong, B., et al. (2021). Hierarchical cortical networks of "voice patches" for processing voices in human brain. *Proc. Natl. Acad. Sci. U.S.A.* 118, e2113887118. doi: 10.1073/pnas.2113887118

Zhuo, C., Jiang, D., Liu, C., Lin, X., Li, J., Chen, G., et al. (2019). Understanding auditory verbal hallucinations in healthy individuals and individuals with psychiatric disorders. *Psychiatry Res.* 274, 213–219. doi: 10.1016/j.psychres.2019.02.040

Zmigrod, L., Garrison, J. R., Carr, J., and Simons, J. S. (2016). The neural mechanisms of hallucinations: a quantitative meta-analysis of neuroimaging studies. *Neurosci. Biobehav. Rev.* 69, 113–123. doi: 10.1016/j.neubiorev.2016.05.037

frontiers | Frontiers in Human Neuroscience

TYPE Hypothesis and Theory
PUBLISHED 29 July 2022
DOI 10.3389/fnhum.2022.851991

Check for updates

OPEN ACCESS

EDITED BY
Douglas M. Shiller,
Université de Montréal, Canada

REVIEWED BY
Marilyn May Vihman,
University of York, United Kingdom
Vikram Ramanarayanan,
University of California, San Francisco, United States

*CORRESPONDENCE
Sam Tilsen
tilsen@cornell.edu

SPECIALTY SECTION
This article was submitted to Speech and Language, a section of the journal Frontiers in Human Neuroscience

RECEIVED 10 January 2022
ACCEPTED 12 July 2022
PUBLISHED 29 July 2022

CITATION
Tilsen S (2022) An informal logic of feedback-based temporal control. *Front. Hum. Neurosci.* 16:851991. doi: 10.3389/fnhum.2022.851991

COPYRIGHT
© 2022 Tilsen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

An informal logic of feedback-based temporal control

Sam Tilsen*

Cornell Phonetics Lab, Department of Linguistics, Cornell University, Ithaca, NY, United States

A conceptual framework and mathematical model of the control of articulatory timing are presented, in which feedback systems play a fundamental role. The model applies both to relatively small timescales, such as within syllables, and to relatively large timescales, such as multi-phrase utterances. A crucial distinction is drawn between internal/predictive feedback and external/sensory feedback. It is argued that speakers modulate attention to feedback to speed up and slow down speech. A number of theoretical implications of the framework are discussed, including consequences for the understanding of syllable structure and prosodic phrase organization.

KEYWORDS

articulation, articulatory timing, speech rate, motor control, feedback, dynamical systems, phonology, prosody

Introduction

Perhaps you have been in a situation in which it was necessary to *shush* someone. For example, imagine you are reading in a library, when a rude person nearby begins talking on their cell phone. You glare at them and say "shhh," transcribed phonetically as [ʃ::]. What determines the duration of this sound? Consider now a different situation: in a coffee shop you are ranting to your friend about the library incident, and your friend tells you to slow down because you are talking too fast. You take a deep breath and proceed more slowly. How do you implement this slowing? The focus of this manuscript is on how variation in the temporal properties of event durations (your "shhh") and variation in event rate (your rapid coffee shop rant) are related. More specifically, what is the mechanistic connection between control of event timing on short timescales and control of speech rate on longer timescales? It is argued that the answer to this question involves a notion of feedback, and that the same feedback mechanisms are involved on both timescales. In other words, control of event timing involves feedback, and control of rate is reducible to control of timing.

Figure 1 provides a graphical depiction of the organization of the manuscript, and lays out the logical structure of the main argument. The motivation for developing a feedback-based model of temporal control is based on three propositions: (i) That current models generally do not provide an empirically adequate account of the role of feedback in the temporal control of articulation (see Section "The need

for a model of feedback-based temporal control"). (ii) That the Task Dynamics (TD)/Articulatory Phonology (AP) framework does not use feedback for temporal control (see Section "Gestural systems and temporal control of gestural activation"). (iii) That empirical phenomena require internal/external feedback control, as well as feedforward control (see Section "Model space and hypotheses"). From these propositions it follows (iv): a model that combines the gestural mechanisms of TD with internal and external feedback-based temporal control is needed. It is important to emphasize that temporal control—control of the timing and relative timing of events—is different from the issue of how movement events are controlled once an intention to achieve articulatory/auditory goals is assumed to be present. The section "The need for a model of feedback-based temporal control" argues that existing models of feedback focus on how tasks/goals are translated to movements but not on how tasks/goals are organized in time. The section "Gestural systems and temporal control of gestural activation" introduces the gestural framework of TD along with the central topic of this manuscript: the question of what causes articulatory events to begin and to end? The section "External feedback vs. internal feedback" defines the notions of internal and external feedback which are employed throughout this manuscript and the sections "Time-representing systems and timing control" and "Deterministic behavior of TiRs and effects of stochastic forces" classify and demonstrate the basic properties of the systems which are used for temporal control in the proposed model. The specific hypotheses of the model and the empirical phenomena which motivate them are detailed in the section "Model space and hypotheses." Further predictions and extensions of the model are discussed in the sections "External influences on parameters," "Parallel domains of competitive selection," and "A model of speech rate control with selectional effects." Finally, some important implications of the model are discussed in the section "General discussion," regarding the control of timing of target achievement (see Section "No direct control of the timing of target achievement"), syllabic and moraic phonological structure (see Section "Reinterpretation of syllabic and moraic structure"), and prosodic phrase structure (see Section "Reinterpretation of prosodic phrase structure and boundaries").

Temporal patterns in speech are challenging to characterize because they exist across a wide range of analysis scales. **Figure 2A** shows rough approximations of timescales associated with various measurements and theoretical vocabularies. Even over the modest range of 20 ms to 5,000 ms (shown in a logarithmic axis), there are many different ways to associate time intervals with theoretical constructs. Furthermore, there are certain terms—"coordination," "boundaries"—which reappear across scales, and problematically necessitate different interpretations.

It is rarely the case that models of small scale phenomena, such as articulatory timing within syllables, are integrated with models of larger scale phenomena, such as boundary-related slowing. One noteworthy exception is the π-*gesture* model (Byrd and Saltzman, 2003), which modulates the rate of a global dynamical clock in the vicinity of phrase boundaries, thereby slowing the timecourse of gestural activation. Another example is the multiscale model of Tilsen (2013), where oscillator-based control of gestural timing is limited to syllable-sized sets of gestures that are competitively selected with a feedback-based mechanism. This early combination of oscillator- and feedback-based control led to the development of Selection-Coordination theory (Tilsen, 2014, 2016), an extension of the AP framework that uses feedback control to account for a variety of cross-linguistic and developmental patterns. A recent proposal in this context is that speech rate is controlled by adjusting the relative contributions of external (sensory) feedback and internal (predictive) feedback (Tilsen, 2018). One of the aims of this manuscript is to elaborate on this idea, advancing that > the generalization that temporal control in speech is largely (but not exclusively) feedback-based. This aim is also the primary novelty of the manuscript: it argues directly that internal and external feedback systems, beyond their commonly held roles in state estimation and error detection/correction, play a fundamental role in the control of timing. In a more general sense, the novelty of the manuscript is its original combination of feedforward control mechanisms described in AP and TD (Saltzman and Munhall, 1989; Saltzman et al., 2008) with feedback control mechanisms used in competitive queuing models (Grossberg, 1987; Bullock and Rhodes, 2002), while explicitly distinguishing internal and external feedback.

A broader aim is to argue for a worldview in which speech patterns are understood to result from interactions of dynamical systems. The "informal logic" developed here advocates for new way of thinking about patterns in speech. It is relevant both for the study of speech motor control, specifically in relation to feedback and control of timing, and for theories of phonological representation, sound patterns, and change. The informal logic challenges the prevailing ontologies of many phonological theories by rejecting the notion that speech is cognitively represented as a structure of hierarchically connected objects, as in **Figure 2B**. It also rejects the notion that such units project "boundaries" onto the temporal dimension of the acoustic signal. Most importantly, the logic holds that speakers never control event durations directly: rather, durational control is accomplished *via* a class of systems which *indirectly* represent time. They do this by integrating the forces they experience from other systems, or from their surroundings.

The systems-oriented approach can provide a more coherent understanding of temporal phenomena across scales. Its logic is qualified as "informal" because, unlike a formal logic, it does not rely heavily on symbolic forms; rather, the schemas presented below are iconic and indexical, designed to help users rapidly interpret complex patterns of system interactions. At the same time, the schemas can be readily mapped to a explicit

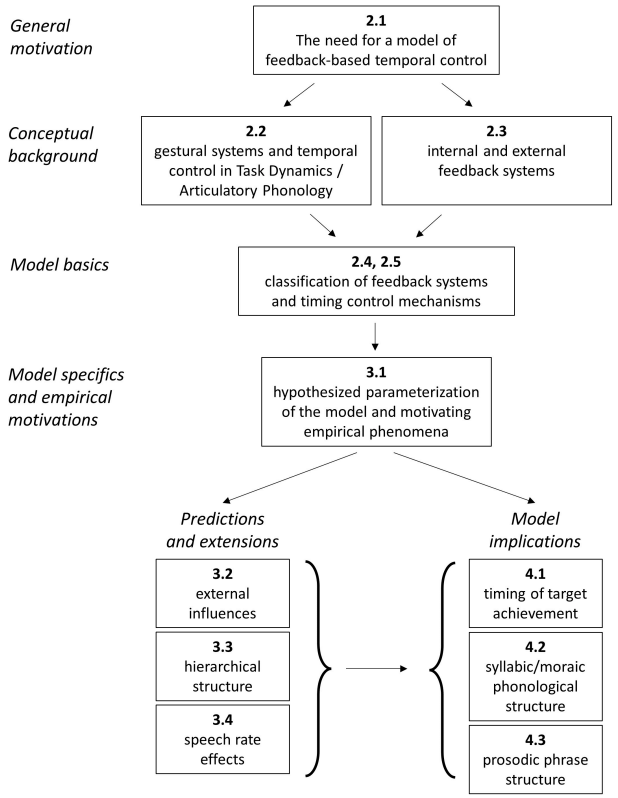


FIGURE 1
Graphical illustration of manuscript organization and propositions comprising the main argument.

mathematical model. All model equations and simulation details are described in **Supplementary Material**, and all code used to conduct simulations and generate figures has been made available in a public repository, here: https://github.com/tilsen/TiR-model.git. The model has been designed to be as simple as possible while being able to generate a variety of temporal patterns. All model equations are presented in the **Supplementary Material** rather than the manuscript, for three reasons. First, the manuscript itself is geared toward a larger audience of readers who are interested in a conceptual understanding of the framework and its applications; hence a graphical rather than symbolic approach to illustrating model structure has been adopted throughout. Second, for the subset of readers who are interested in understanding the mathematical implementation, presenting the equations together in the **Supplementary Material** facilitates this endeavor. Third, in the case of modeling complex cognitive systems, it is important in general not to overemphasize the specific details of mathematical models; following Saltzman and Munhall (1989), I believe that "the primary importance of the work lies not so much in the details of [the] model, but in the problems that can be delineated within its framework" (Saltzman and Munhall, 1989, p. 335). My hope is therefore that the informal logic presented here can be used by researchers to conceptualize empirical phenomena, without requiring them to implement the model I have constructed. And yet, for those who are interested in critiquing or improving the model, or adapting it to interface with other models, I have made substantial efforts to facilitate this. Finally, although the implications of the framework are fairly general, it is nonetheless narrowly focused on describing a logic of *temporal* control. Issues related to "spatial" dimensions of feedback or to feedback modalities are set aside for future extensions.

Background

Below we argue that existing models of speech production do not adequately account for the temporal control of articulatory events, and hence there is a need for a model that focusses on temporal patterns in speech. Subsequent sections introduce some of the key concepts that are incorporated in the model developed here. As a general principle, the components of the model are always viewed as systems and their relations are viewed as interaction forces. Systems are abstract entities which have time-varying internal states. Our analytical task is to formulate change rules to describe how the system states evolve over the course of an utterance, as shown generically in

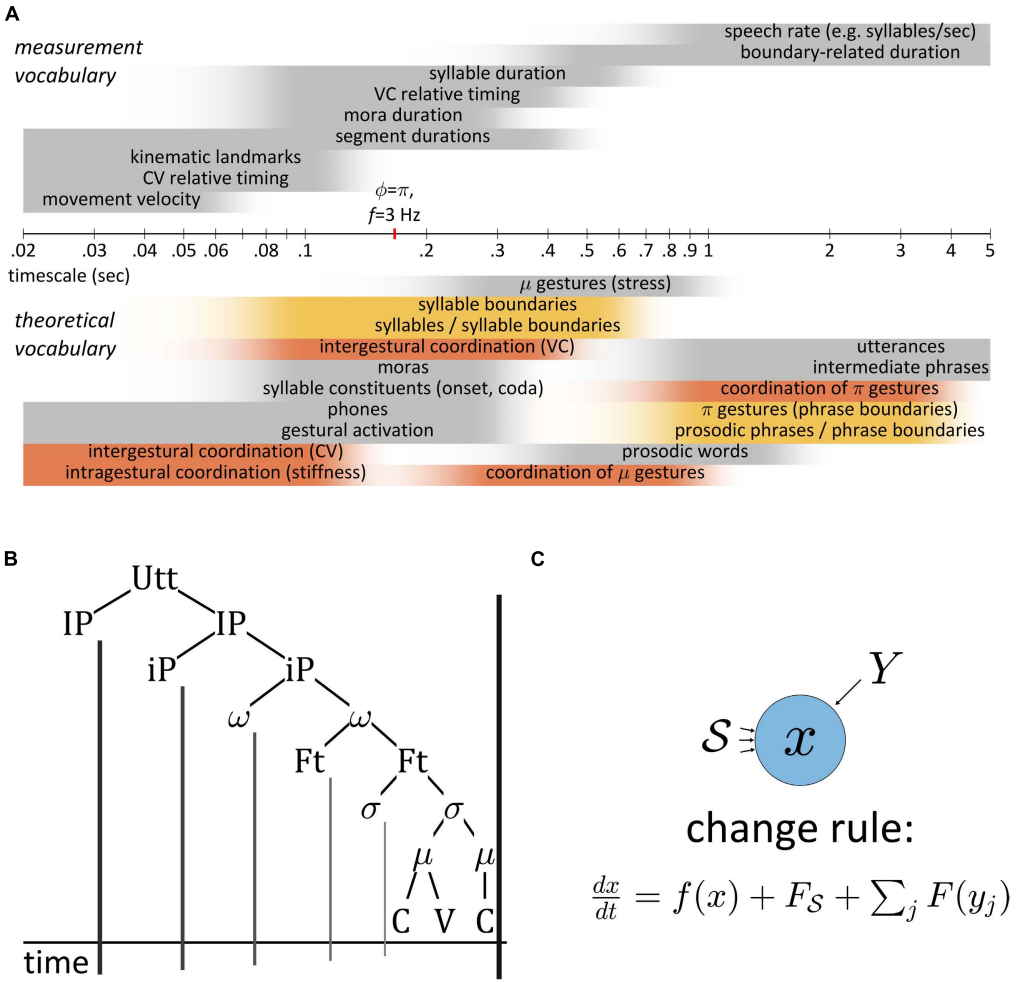


FIGURE 2
(A) Comparison of timescales associated with various measurements and theoretical constructs used to conceptualize temporal patterns. Time axis is logarithmic. Shaded intervals approximately represent ranges of time in which terminology applied. **(B)** Hierarchical conception of prosodic structure and implicit projection of units to boundaries in a temporal coordinate. **(C)** Generic system schema, where change in the state variable *x* is a function of *x* itself and of forces from the surroundings *S* and from other systems *Y*.

Figure 2C. This setup provides a frame in which to analyze and interpret the causes of empirical patterns in speech. Moreover, to draw generalizations about systems and their interactions we must classify them. To accomplish this in the following sections we define terms below such as *internal*, *external*, *feedback*, and *sensory*. These terms are necessarily relative and therefore potentially ambiguous out of context, thus the reader should pay careful attention to these definitions to avoid confusion.

The need for a model of feedback-based temporal control

The motivation for the model developed here is that there currently exists no model of speech motor control that provides an empirically adequate account of articulatory event timing. Importantly, the issue of event timing is different from the issue of how movement is controlled when an intention to generate movement is presupposed. There are several models that provide accounts of how movements are controlled, but only when it is assumed that a speaker has an intention to achieve some goal—these are models which focus on control *from* an intention. As discussed below, most of these models do not address how the intentions themselves are organized in time, i.e., the control *of* (the timing of) intentions. Only one of the models provides explicit mechanisms for governing the temporal organization of intentions, but that model is inadequate from an empirical perspective. By "intention" here I mean the aim of a speaker to achieve a goal-related outcome(s). This abstract term is used in order to generalize over models that are based on different hypothetical entities—often either phones/phonemes or tasks/articulatory gestures.

It is crucial for the reader to understand that control *from* intentions and control *of* intentions are distinct topics: most speech motor control models assume that intentions to conduct movement exist, and ask how those movements are realized and modulated by feedback; in contrast, my interest in this manuscript is what causes the intentions themselves to begin and to end. For example, the questions asked by researchers interested in the control *of* intentions (e.g., Guenther and Hickok, 2016) are "What exactly are the goals, or targets, of the speech production planning process?" and how can the nervous system "generate a complex muscle activation pattern that satisfies the goals of the movement"? These are important questions but they are not the focus of this manuscript, because they are not about the temporal organization of the goals/targets of speech.

Indeed, most speech motor control models do not adequately address the question of temporal control. First, consider the directions into velocities of articulators (DIVA) model (Tourville and Guenther, 2011). In the relatively recent description of this model in Guenther and Hickok (2016), it is stated that "the DIVA model's feedforward commands constitute a form of motor program and, as such, they are closely related to the concepts of a gestural score"; the authors then state that "a gestural score is a stored representation of properly timed vocal tract gestures." It is held that—following (Levelt and Wheeldon, 1994)—frequently used syllables or sequences of syllables are stored as motor programs, and infrequent syllables may be assembled during speech planning from phoneme-sized programs. This characterization of a gestural score as "a stored representation of properly timed vocal tract gestures" is inconsistent both with early formulations of the TD model of speech production, as well as most of the recent theoretical literature on AP and TD (Browman and Goldstein, 1989; Saltzman and Munhall, 1989), which holds that patterns of gestural activation are generated online rather than being stored. This point is discussed more thoroughly in Section "The need for a model of feedback-based temporal control," in the context of a close examination of TD model. Ultimately, the DIVA model alone does not specify what determines the timing of its feedforward commands; rather it presupposes that some timing pattern is already specified.

The gradient ordering (GODIVA) model (Bohland et al., 2010) is an extension of DIVA that incorporates a model of timing, yet this model is empirically inadequate in several ways. GODIVA employs a competitive queuing mechanism to sequentially activate the individual phonemes that are hypothesized to comprise a syllable. Once a syllable is selected, the plan for the first phoneme of that syllable becomes active for a "short duration" (parameter τ of Equation 6 in Bohland et al., 2010), and each subsequent phoneme instantaneously becomes active upon the deactivation of the preceding one. Hence, the model provides a purely sequential account of the temporal organization of intentions (i.e., the goals associated with phonemes). GODIVA is empirically inadequate for several reasons, which are briefly mentioned here and discussed more thoroughly in Section "Model space and hypotheses." First, articulatory movements in adult speech overlap substantially, especially in syllable onsets, where movements associated with consonantal constrictions are largely coextensive in time with vowel-related movements. The GODIVA model does not explain how such extensive temporal overlap could arise from plans which are selected sequentially; in actuality, it predicts the opposite: that consonantal and vocalic movements should occur in a non-overlapping sequence. Second, in complex-onset syllables such as CCV, the order in which the constriction formation movements are initiated empirically is such that the initiation of vocalic movement intervenes between the initiations of the constriction formations: thus GODIVA explicitly imposes a CCVCC sequencing of phones within syllables that does not correspond to the order in which movements are initiated in empirical data (see Section "Empirical motivation for pre-vocalic oscillator-based control" for references). Third, the model does not does not discuss sources of variation in the phoneme duration parameter τ, and therefore it is hard to say what it predicts regarding variability in event durations. Finally, the model does not provide a role for sensory feedback to influence the timing of phone selection; instead, the role of sensory feedback in DIVA/GODIVA is limited to the detection and correction of errors, which can only indirectly influence timing.

The hierarchical state feedback control (HSFC) model of Hickok (2012) argues that both external and internal sensory feedback are used for the detection and correction of errors in speech plans and their outputs. However, the model focuses on the activation of hypothesized syllable and phoneme motor programs; it does not generate articulatory events. Indeed, the words "duration" and "timing" are never used to describe model-generated events in Hickok (2012). As with DIVA/GODIVA, HSFC focuses on the use of feedback for error detection/correction, but not on the temporal organization of the intentions to achieve targets. The equilibrium point model of motor control (Feldman, 1986; Feldman and Levin, 2009) is also not a model of temporal control; it describes how goals (changes in equilibria) are implemented through effector/muscle synergies. This model does not address the issue of when changes in equilibrium points occur. Similarly, the powerful feedback-aware control of tasks in speech (FACTS) model (Parrell et al., 2018, 2019), although avoiding the empirical problems associated with phoneme-sequence conceptions of speech, is a model of how control is achieved given a presupposed temporal pattern of intentions. FACTS does not aim to address how the temporal pattern of intentions is generated in the first place.

Thus, many speech motor control models—DIVA, HSFC, FACTS, equilibrium points—do not directly address the role of feedback in temporal control; instead, they employ feedback

for error detection/correction. The GODIVA model, which contains a mechanism for the sequencing of phones, does not allow feedback any direct role in this sequencing process, and imposes an ordering of phones that is not empirically motivated.

Gestural systems and temporal control of gestural activation

This section describes the understanding of articulatory control adopted here, which originates from TD (Kelso et al., 1986; Saltzman and Munhall, 1989). It is argued that although TD provides a useful framework for thinking about temporal control, the model and its phonological counterpart AP (Browman and Goldstein, 1989) leave important questions regarding articulatory timing unresolved; most importantly, they do not make use of feedback for control of timing. In TD, changes in the physical outputs of speech—vocal tract shape and distributions of acoustic energy—are indirectly caused by systems called *articulatory gestures*. **Figure 3A** schematizes the organization of system interactions in the TD model for production of the word *pop*: gestural systems for bilabial closure (clo), bilabial release (rel), and the vocalic posture of [a] exert driving forces on vocal tract systems of lip aperture (LA) and pharyngeal constriction degree (PHAR), which in turn exert forces on articulator systems for the upper lip (UL), lower lip (LL), jaw, and tongue root (TR). [As an aside, note that the framework attributes no ontological status to phones or phonemes—these are merely "practical tools" (Browman and Goldstein, 1990) or inventions of scientific cultures (Ladefoged, 2001; Port and Leary, 2005).] Gestural system states are defined in normalized activation coordinates which range from zero to one, and gestures are understood to abruptly become active and subsequently deactivate, as shown for the word *pop* in **Figure 3B**—this panel includes the activation intervals of a bilabial closure gesture (LA clo), a bilabial release gesture (LA rel), and a tongue root gesture, which achieves a pharyngeal constriction for the vowel [a]. When their activation is non-zero, gestures exert forces on vocal tract systems, which can lead to movement, as shown in **Figure 3C** for timeseries of lip aperture (LA) and pharyngeal constriction degree (PHAR).

In both a theoretical and technical sense, gestures should be understood as *systems*. They are entities which have internal states and which experience and exert forces. Accordingly, gestures are not movements, nor are they periods of time in which movements occur. To reinforce this point we often refer to them (redundantly) as *gestural systems*. The distinction is important because it is common to refer to movements of vocal organs as "gestures"—but this can cause confusion. Similarly, the periods in which gestural systems obtain states of high activation (shaded intervals in **Figure 3B**) are sometimes called "gestures"—these periods are better described as *gestural activation intervals*. The point here is simply that metonymic extensions of "gesture" to refer to physical movements or activation intervals should not be conflated with the systems themselves. Furthermore, the vocal tract and articulator system states of the TD model are nervous system-internal representations of the physical geometry of the vocal tract/effectors. The actual geometry of the vocal tract is not modeled explicitly in TD and can in principle diverge from these internal representations. Finally, in the TD model, vocal tract system states are defined in position, velocity coordinates, and interactions between gestural systems and vocal tract systems are analogous to mechanical forces. These particular analogies do not apply to forces experienced by gestural systems, nor to other types of systems which we develop below. The systems we construct are better analogized to many-body, open thermodynamic systems: their "activation" states are conceptualized as energies, rather than positions/velocities, and their interactions are analogized to thermodynamic generalized forces. This set of conceptual metaphors is further discussed in the **Supplementary Material**, in the context of the model equations.

The TD framework is particularly valuable because it clarifies the questions that must be addressed in order to understand temporal patterns in speech. There are two questions of paramount importance regarding temporal control: (i) What causes inactive gestural systems to become active? and (ii) What causes active gestural systems to become inactive? These questions correspond to the arrows marking initiations and terminations of the gestural activation in **Figure 3B**.

(i) *What causes gestures to become active?* In answering this question, we temporarily adopt the perspective that the entire set of gestures is a "system." One possible answer then is that there are some *external* systems which exert forces on the gestures. By "external" we mean systems which are "outside" of the set of gestures, and we refer to such systems as *extra-gestural*. Another possibility is that the gestural systems experience forces from each other, in which case the activating forces come from "inside of the system" or are *internal* to the system of gestures, i.e., *inter-gestural*. Note that the first gesture to become active must necessarily be activated by an extra-gestural system, because there is presumably no way for a gestural system to spontaneously "activate itself" or to be activated by inactive gestural systems.

(ii) *What causes gestures to cease to be active?* The extra-gestural and inter-gestural forces described above are both plausible sources of deactivation. A third possibility, unavailable in the case of activating forces, is that deactivation is caused by actions of individual gestural systems on themselves, i.e., *intra-gesturally*. We elaborate below on how this differs from inter-gestural control.

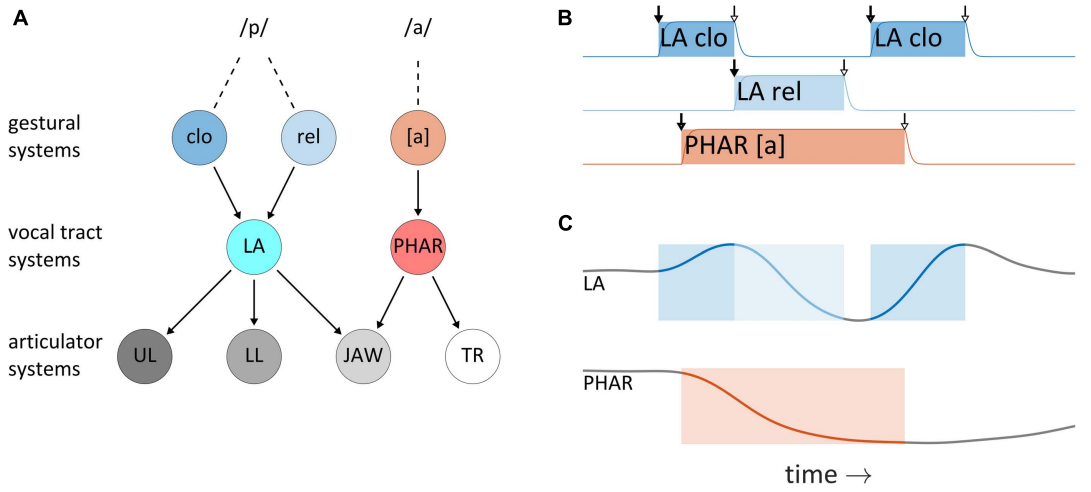


FIGURE 3
System organization and interactions in the Task Dynamics model. **(A)** Organization of system interactions (see text for descriptions of systems). **(B)** Gestural activation intervals for the CVC syllable pop. **(C)** Vocal tract geometry changes resulting from the actions of gestural systems on vocal tract systems. Lip aperture (LA) and pharyngeal constriction (PHAR) timeseries are shown.

The TD model of speech production developed by Saltzman and Munhall (1989) did not resolve which of the various sources of initiating and terminating forces are utilized. Saltzman and Munhall heuristically hand-specified activation intervals to fit empirical data, but they proposed that the model could be extended with the serial network of Jordan (1986) to dynamically control gestural activation. In this serial network, the hidden layers responsible for sequencing might be interpreted as extra-gestural forces. Much attention has been given to the issue of gestural timing in the framework of AP (Browman and Goldstein, 1986, 1988, 1990, 1992). Many early descriptions of timing in AP—in particular references to "phasing"— imply that initiating forces are inter-gestural and that terminating forces are intra-gestural, in line with the explicit interpretations of phasing in Kelso and Tuller (1987). In contrast, later descriptions hypothesize that gestures are activated by a separate system of gestural planning oscillators (Goldstein et al., 2006; Saltzman et al., 2008), which are extra-gestural. These approaches attribute no role to feedback in the initiation or termination of gestures. Thus the current situation is one in which several different possible understandings of feedforward temporal control of gestures have been proposed, none of which specifically implicate feedback.

To summarize, the systems-view of gestural control in the TD framework provides two generic options for what causes gestures to become active or cease to be active—extra-gestural systems or other gestures (inter-gestural control)—along with a third option of intra-gestural control as a form of self-deactivation. There is no theoretical consensus on which of these are actually involved in control of articulatory timing, or in what contexts they may be utilized. Furthermore, feedback has not been incorporated into this framework for the purpose of controlling gestural timing.

External feedback vs. internal feedback

Definitions of external and internal feedback are presented here. The term *feedback* has a variety of different uses. Here *feedback* refers to information which—in either a direct or indirect manner—is produced by some particular system, exists outside of that system, and subsequently plays a role in influencing the state of that same system. Thus feedback is always defined relative to some reference system. In current contexts the reference system is sometimes a particular gestural system, other times the entire set of gestural systems, and most often the central nervous system. Feedback in this sense is a very general notion, and does not presuppose that "sensory" organs such as the cochlea or muscle stretch receptors are involved.

Note also that the "information" referred to in the above definition of feedback can be plausibly given a technical interpretation (Shannon, 1948), but the actual quantification of said information faces many obstacles. Strictly speaking, information is produced when an observer's uncertainty in the state of a system is reduced. Quantification of information production requires knowledge of the probability distribution over states of an observed system, along with definition of the observed and observing systems. For example, a vocal tract system "observes" the forces it experiences from a gestural system, but to quantify the information produced by this observation we need a probability distribution over all possible gestural system forces. However, to simply determine whether information meets the definition of feedback, we need only to

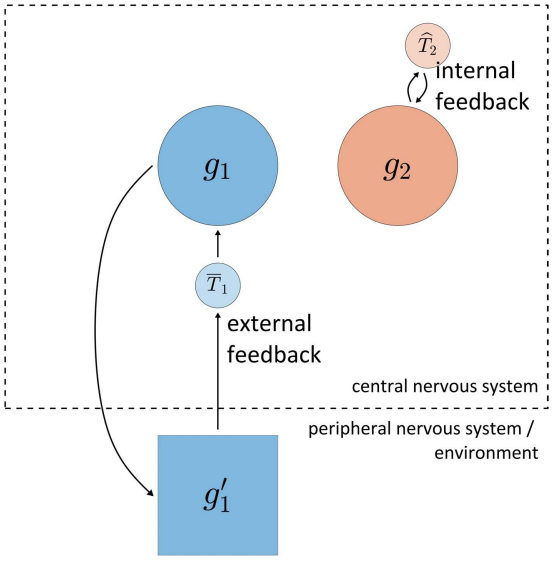


FIGURE 4
Schematic illustration of distinction between internal and external feedback. The dashed line represents the boundary of the central nervous system. Systems g_1 and g_2 are gestural systems, g'_1 is system which represents information associated with g_1 outside of the central nervous system, and T_1 and T_2 are hypothetical systems which use feedback to act on g_1/g_2.

identify the chain of interactions associated with information production. If that chain forms a loop back to the particular system of interest, then it meets the definition of feedback.

For a logic of feedback-based temporal control of speech it is crucial to distinguish between *external feedback* and *internal feedback*, as illustrated in Figure 4. The reference system is the central nervous system (CNS, consisting of cortex, brainstem, and spinal cord). External feedback involves information that (i) is originally generated within the CNS, (ii) causes information to be produced outside of the CNS, and (iii) in turn causes information to be produced within the CNS; correlations must obtain between the information in these three stages. For example, activation of the gestural system g_1 causes the production of various forms of information in the environment (movement of articulators, generation of acoustic energy), which is in turn transduced in the peripheral nervous system (depolarization of hair cells in the cochlea and sensory muscle fibers) and subsequently produces information in cortical systems. For current purposes we draw no distinctions between various sensory modalities, which are lumped together as system g'_1 in Figure 4. The information associated with g'_1 can ultimately influence the state of g_1, and hence meets our definition of feedback. Notice that Figure 4 includes a system labeled $\overline{T}_1$, which uses the external feedback from g'_1 to act on g_1.

In contrast to external feedback, internal feedback is information which never exists outside of the CNS. For example, in Figure 4 the gestural system g_2 generates information that system $\hat{T}_2$ uses to act on g_2. Thus the contrast between external and internal feedback is based on whether the relevant information at some point in time exists "outside of"/"external to" the central nervous system. External feedback may be also described as "sensory" feedback, but with a caveat: one could very well also describe internal feedback as "sensory," in that any experience of force—regardless of its origins—can reasonably be considered a form of *sensation*. The point is simply that the word "sensory" is ambiguous regarding what is being sensed, and so the qualifiers *internal* and *external* are preferred, with the CNS being the implied reference system. Internal feedback can also be described as "predictive," but we should be cautious because this term strongly evokes an agentive interpretation of systems.

The distinction between external and internal feedback is only partly orthogonal to the distinction between extra-gestural, inter-gestural, and intra-gestural control. The full system of gestures is by definition within the CNS; hence feedback associated with inter-gestural and intra-gestural control is by definition internal feedback. In contrast, extra-gestural control may involve either external feedback (e.g., auditory or proprioceptive information) or internal feedback from CNS-internal systems. This can be confusing because "extra"-gestural control does not entail external feedback—hence the necessity to keep tabs on the system boundaries to which our vocabulary implicitly refers. When describing feedback, the reference system is the CNS. When describing control of gestural activation, the reference system is either the full system of gestures (for extra-gestural control) or individual gestural systems (for inter- vs. intra-gestural control).

The Task Dynamic model incorporates no feedback of any form for gestural systems. Nonetheless, Saltzman and Munhall cited the necessity of eventually incorporating sensory feedback, stating: "without feedback connections that directly or indirectly link the articulators to the intergestural level, a mechanical perturbation to a limb or speech articulator could not alter the timing structure of a given movement sequence" (Grossberg, 1987, p. 360). Note that here Saltzman and Munhall expressed a concern with the *temporal* effects of perturbation rather than *spatial* effects—in this manuscript, we are similarly focused on timing but recognize that a complete picture should incorporate a fully embodied and sensorially differentiated model of the articulatory and acoustic dimensions of feedback.

Time-representing systems and timing control

To augment our classification of the ways in which gestural systems may be activated or deactivated, we need to think about how time may be "measured," "estimated," or "represented" by the nervous system. Researchers have adopted various ways of talking about different types of systems that serve

this function (Kelso and Tuller, 1987; Schöner, 2002)—timers, clocks, timekeepers, virtual cycles, etc., with the discussion of Schöner (2002) being particularly informative. For current purposes, we describe such systems as "time-representers" (TiRs) and develop a multidimensional classification. Despite this name, we emphasize that temporal representations are *always indirect*: the states of TiR systems are never defined in units of time.

Before classifying TiRs, we make a couple points regarding their interactions with gestures. First, each gestural system is associated with a gating system, labeled "G" in **Figure 5A**. The gating system states are treated as binary: gates are either open or closed. When a gestural gate is open, the activation state of the associated gestural system transitions rapidly toward its normalized maximum activation of 1. Conversely, when the gate is closed, the gestural system transitions rapidly toward its minimum value. For current purposes, transitions in gestural activation states occur in a single time step, as in Saltzman and Munhall (1989). Nothing hinges on this simplified implementation and the model can be readily extended to allow for activation ramping or non-linearities to better fit empirical tract variable velocity profiles (Sorensen and Gafos, 2016).

Second, TiRs act on gestural gating systems, not directly on gestures, and thus function to activate/deactivate gestural systems indirectly. One reason for including gating systems as intermediaries between TiRs and gestures is that they allow for the dynamics of gestural systems to be dissociated from the forces that control gestural activation. The actions of TiRs are modeled as brief, pulse-like forces, and always depend on TiR-internal states: each TiR has threshold parameters (τ) which specify the internal states (in units of activation) at which the TiR acts on gating systems. The action threshold parameters are labeled on the arrows of **Figure 5A**. To reduce visual clutter in model schemas, gating systems are omitted from subsequent figures.

One main dimension of TiR classification involves whether a TiR is autonomous or non-autonomous. An *autonomous* TiR does not depend on either gestural or sensory system input to maintain an indirect representation of time. **Figure 5B** shows two examples of autonomous TiRs. The first is ϵ', which activates gestures g_1 and g_2. The second is ϵ_1, which deactivates g_1. Note that autonomous TiRs *do* require an external input to begin representing time—they need to be "turned on"/de-gated—but subsequently their state evolution is determined by a growth rate parameter. This parameter may vary in response to changes in a hypothesized "surroundings" or contextual factors.

In contrast to autonomous TiRs, the states of *non-autonomous* TiRs depend on input from a gestural or sensory system. Non-autonomous TiRs integrate the forces that they experience from a given system. An example is $\widehat{\mathbf{T}}_2$ in **Figure 5B**, which receives input from g_2 and deactivates g_2 upon reaching a threshold state of activation, here $\tau = 0.25$. Non-autonomous TiRs are associated with integration rate parameters α, which determine how much the forces they experience contribute to changes in their internal states.

The key difference between autonomous TiRs and non-autonomous ones is that the states of the autonomous TiRs evolve independently from the states of gestures or sensory systems. In the example of **Figure 5B** the states of autonomous TiRs ϵ' and ϵ_1 are assumed to be 0 at the beginning of the simulation and increase linearly in a way that represents elapsed time. In this example (but not in general), the growth rates of autonomous TiR states were set to $1/\Delta t$ (where Δt is the simulation time step); consequently, their activation states exactly correspond to elapsed time. This is convenient for specifying threshold parameters that determine when TiRs act on other systems. Similarly, the integration rate parameters of non-autonomous TiRs were parameterized to represent the time elapsed from the onset of gestural activation. In general, the correspondence between TiR activation values and elapsed time is neither required nor desirable, and we will see how changes in TiR growth rates/integration rates are useful for modeling various empirical phenomena.

Another dimension of TiR classification involves the sources of input that non-autonomous TiRs make use of to represent time. Non-autonomous TiRs can be described as *external* or *internal*, according to whether they integrate external or internal feedback. This distinction is illustrated in **Figure 6A**, where the non-autonomous TiR $\widehat{\mathbf{T}}_1$ can be described as internal because it integrates feedback directly from gesture g_1. In contrast, the non-autonomous TiR $\overline{\mathbf{T}}_2$ is external because it integrates feedback from sensory systems which encode the actions of g_2 outside of the CNS.

Non-autonomous, internal TiRs are further distinguished according to whether they are inter-gestural or intra-gestural (internal to a gesture). Intra-gestural internal TiRs can only act on the particular gestural system that they are associated with, and can integrate forces only from that gesture. Inter-gestural TiRs can act on and experience forces from any gestural system. For example, in **Figure 6B**, the deactivation of g_1 is controlled by an intra-gestural TiR $\widetilde{\mathbf{T}}_1$, but the inter-gestural TiRs $\widehat{\mathbf{T}}_1$ and $\widehat{\mathbf{T}}_2$ activate and deactivate g_2, respectively. The distinction is useful if we wish to impose the condition that a TiR is isolated from all systems other than a particular gesture.

The distinction between inter-gestural and intra-gestural TiRs can be viewed in relation to different aspects of the virtual cycles that Kelso and Tuller (1987) proposed to govern gestural timing. Tuller and Kelso held that each gesture could be associated with a virtual cycle, which might be described as a "single-shot" oscillation. Different phases of the cycle were hypothesized to correspond to events such as gesture initiation, achievement of maximum velocity, target achievement, and gesture termination. It was suggested in Browman and Goldstein (1995) that when a virtual cycle phase of $3\pi/2$ rad (270°) is reached, a gesture is deactivated. In this regard intra-gestural TiRs can implement the functions

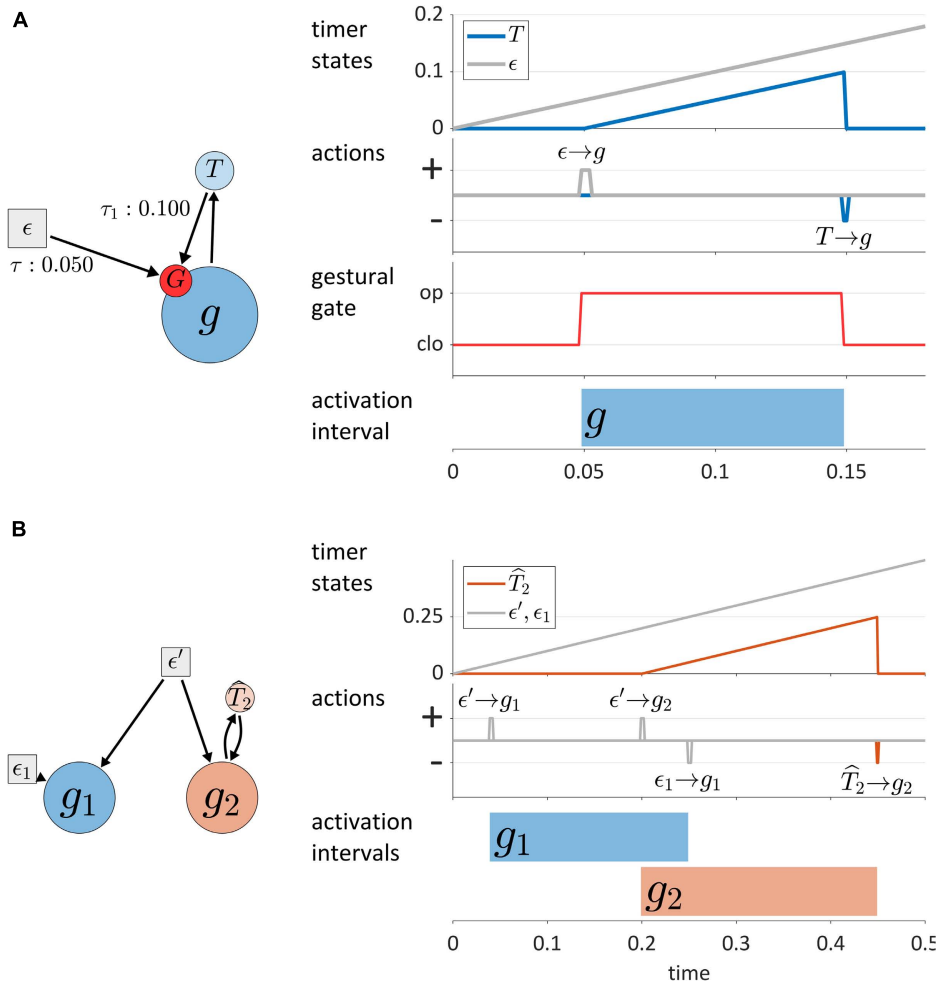


FIGURE 5
(A) Model of interactions between gestures and TiRs, with depiction of the gestural gating system G that TiRs act upon. Panels on the right show timer states, timer actions on gestures, gestural gating system states, and gestural activation interval. **(B)** Distinction between autonomous TiRs (ϵ', ϵ_1) and non-autonomous TiRs ($\widehat{T}_2$).

of virtual cycles: their activation states can be converted to a normalized coordinate that ranges from 0 to 2π, and their growth rates can be adjusted to match the natural frequency of an undamped harmonic oscillator. However, Kelso and Tuller (1987) also proposed that intergestural timing might involve specification of the initiation of the virtual cycle of one gesture relative to the virtual cycle of another. Only inter-gestural TiRs can serve this function, because unlike intra-gestural TiRs, they can act on gestural systems that they are not directly associated with. For all of the purposes that follow in this manuscript, intra-gestural TiRs are unnecessary and we make use of inter-gestural TiRs instead.

Autonomous TiRs can differ in whether their state evolution is aperiodic or periodic. Periodic (or technically, quasi-periodic) TiRs are used in the coupled oscillators model (Saltzman et al., 2008), where each gesture is associated with an oscillatory system called a *gestural planning oscillator*. The planning oscillators are autonomous TiRs because they do not integrate gestural or sensory system states, as can be seen in Figure 7. They are often assumed to have identical frequencies and to be strongly phase-coupled, such that the instantaneous frequencies of the oscillators are accelerated or decelerated as a function of their phase differences. When a given planning oscillator reaches a particular phase, it "triggers" the activation of the corresponding gestural system. The "triggering" in our framework means that the TiR acts upon a gestural system, in the same way that other TiRs act upon gestural systems. The schema in Figure 7 illustrates a system of three periodic TiRs in which θ_1 and θ_3 are repulsively phase coupled to one another while being attractively phase coupled to θ_2.

The phase coupling configuration in Figure 7 generates a pattern of relative phase that—*via* phase-dependent actions on gestural systems—leads to a symmetric displacement of initiations of gestures g_1 and g_3 relative to initiation of g_2.

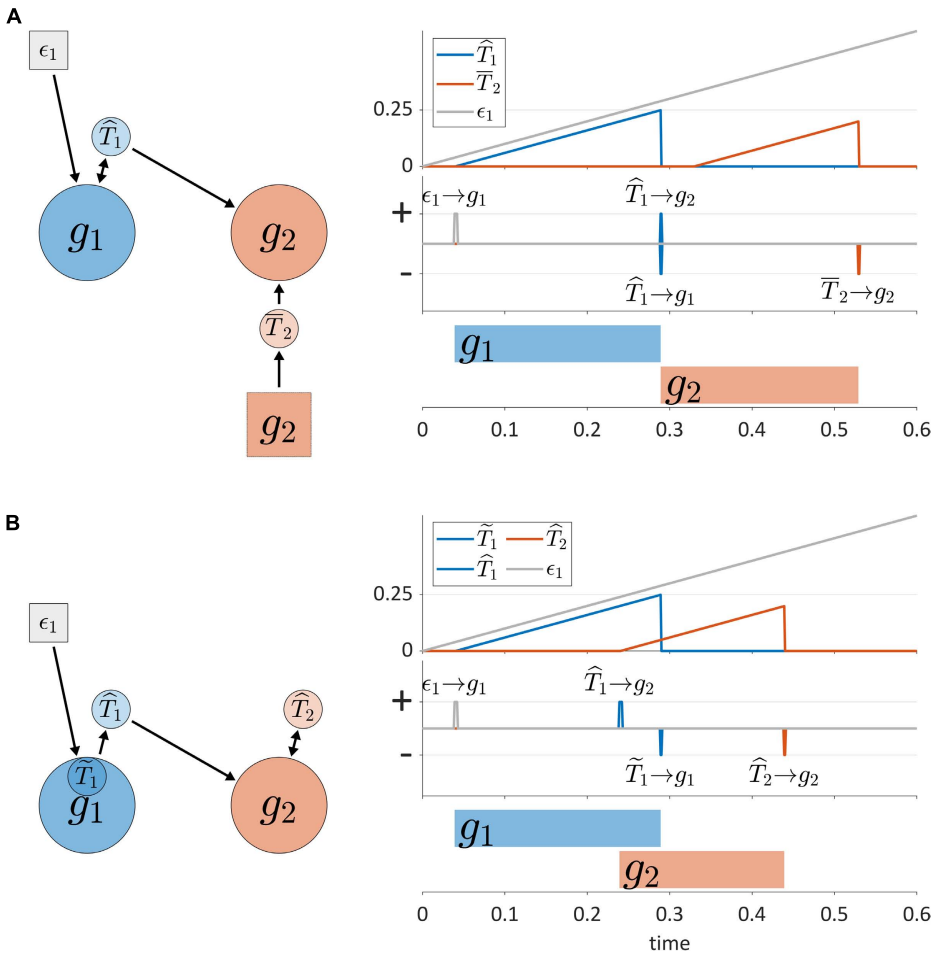


FIGURE 6
(A) External vs. internal sources of feedback for non-autonomous TiRs. Panels on the right show timer states, timer actions, and gestural activation intervals. **(B)** Example of inter-gestural vs. isolated/intra-gestural TiRs.

Statistical tendencies toward symmetric displacement patterns of this sort are commonly observed in two phonological environments: in simple CV syllables, the initiations of constriction formation and release are displaced in opposite directions in time from the initiation of the vocalic gesture (Tilsen, 2017); in complex onset CCV syllables, the initiations of the first and second constriction are equally displaced in opposite directions from initiation of the vocalic gesture (Browman and Goldstein, 1988; Marin and Pouplier, 2010; Tilsen et al., 2012).

The coupled oscillators model has not been used to govern gestural deactivation. Furthermore, a gating mechanism is needed to prevent oscillators from re-triggering gestural systems in subsequent cycles or to prevent them from triggering gestures prematurely. To address this, in the current implementation each oscillator is described by three state variables: a phase angle, a radial amplitude, and the derivative of the radial amplitude. Furthermore, each oscillator is associated with a gating system that controls oscillator amplitude dynamics. As shown in **Figure 7**, intergestural TiRs close these oscillator gates. Moreover, a condition is imposed such that oscillators can only trigger gestural activation when their amplitudes are above a threshold value. The "oscillations" panel of **Figure 7** shows a representation of oscillator states that combines phase and amplitude dimensions (the product of the amplitude and the cosine of phase). Further details are provided in the **Supplementary Material**.

An important hypothesis is that oscillator frequencies are constrained in a way that aperiodic TiR growth rates are not. We refer to this as the *frequency constraint hypothesis*. The rationale is that the oscillator states are believed to represent periodicity in a short-time integration of neuronal population spike-rates; this periodicity is likely to be band-limited due to intrinsic time-constants of the relevant neural circuits and neurophysiology. A reasonable candidate band is theta, which ranges from about 3–8 Hz (Buzsáki and Draguhn, 2004; Buzsaki, 2006), or

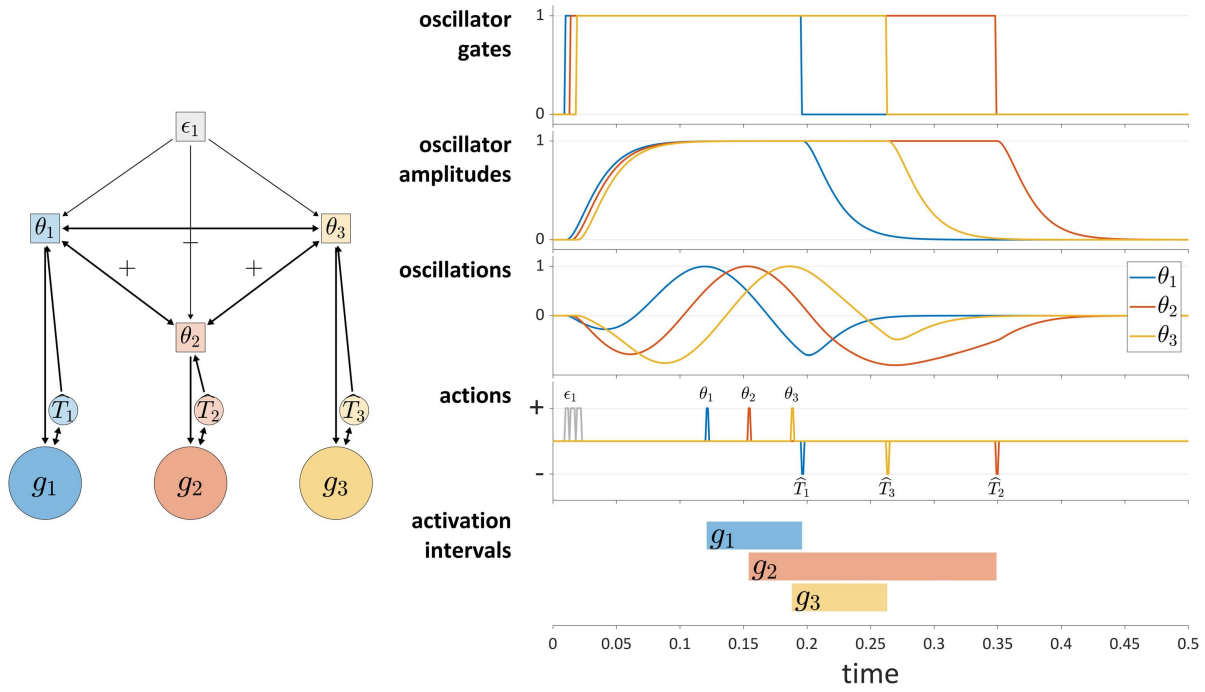


FIGURE 7
The coupled oscillators model in the TiR framework. Periodic TiRs θ1, θ2, and θ3 are phase coupled as indicated by +/− symbols. The oscillator gates, radial amplitudes, and oscillations (amplitude × cosine of phase) are shown. Due to the pattern of phase coupling imposed here, initiation of gestural systems g_1 and g_3 are symmetrically displaced from initiation of g_2.

periods of about 330 to 125 ms. On the basis of these limits, certain empirical predictions regarding temporal patterns can be derived, which we examine in detail below.

Stepping back for a moment, we emphasize that all TiRs can be understood to "represent" time, but this representation is *not* in units of time. The representation results either (i) from the integration of gestural/sensory system forces (non-autonomous TiRs), (ii) from a constant growth rate/frequency (autonomous TiRs) understood to be integration of surroundings forces, or (iii) from a combination of surroundings forces and forces from other TiRs (as in the case of coupled oscillators). Thus the systems we hypothesize represent time indirectly and imperfectly, in units of experienced force.

The utility of TiRs lies partly in their ability to indirectly represent time and partly in their ability to act on gestures or other systems. **Table 1** below summarizes the types of TiRs discussed above. All TiRs are associated with a parameter vector τ that specifies the activation states at which the TiR acts upon other systems, along with a parameter vector χ whose sign determines whether actions open or close gestural gating systems. Autonomous TiRs are associated with a parameter ω which is either a growth rate (aperiodic TiRs) or angular frequency (periodic TiRs). The latter are also associated with a phase-coupling matrix. Non-autonomous TiRs are associated with a vector α of integration factors, which determines how input forces contribute to the growth of activation. Additional simulation parameters and details are described in **Supplementary Material**.

The motivations for including the different types of TiRs defined above relate to the goal of generating various empirical phenomena, which are described more specifically in Section "A hybrid model of gestural timing and speech rate control." Broadly speaking, inter-gestural and extra-gestural non-autonomous TiRs are intended to provide mechanisms for control that involve internal and external feedback, respectively (see Section "Gestural systems and temporal control of gestural activation"). Autonomous periodic TiRs (coupled oscillators) provide precise control over the relative timing of movements, allowing the model to generate symmetric displacement patterns. Autonomous aperiodic TiRs allow the model to initiate and terminate a sequence of actions; as we develop in Section "A hybrid model of gestural timing and speech rate control," these can be used to implement competitive selection, which is a sequencing mechanism.

Deterministic behavior of time-representers and effects of stochastic forces

In order to better understand the behavior of TiRs, it is important to examine the covariance patterns of timing

TABLE 1 Summary of TiRs.

Symbols	Autonomous/non-autonomous	Feedback source	Sub-classes	Periodic/aperiodic	Parameters
ε	Autonomous			Aperiodic	ω, χ/τ
θ	Autonomous			Periodic	ω, χ/τ, Φ
$\overline{T}$	Non-autonomous	CNS-external	Extra-gestural		α, χ/τ
$\hat{T}$	Non-autonomous	CNS-internal	Inter-gestural		α, χ/τ
$\tilde{T}$	Non-autonomous	G-internal	Inter-gestural		α, χ/τ

intervals that are generated by them. The analysis of covariance in temporal intervals is a basic tool for drawing inferences about the organization of temporal control in general (Wing and Kristofferson, 1973; Vorberg and Wing, 1996), and for articulatory timing in particular (Shaw et al., 2009, 2011; Tilsen, 2017). In order for interesting covariance patterns to arise, sources of stochastic variation must be present in the system. This section first establishes the deterministic, non-stochastic properties of temporal intervals in the current framework, and then examines how those temporal intervals covary in the presence of stochastic forces.

Under certain conditions, the time δ when a TiR acts on some other system (δ is relative to when TiR activation began to grow) is fully determined by its parameters. In the case of autonomous, aperiodic TiRs, the growth rate ω and action threshold τ determine δ. In two-dimensional ω/τ parameter space, constant δ are straight lines of positive slope, since increases of ω (which shorten δ) can be offset by increases of τ (which lengthen δ). Thus either changes in TiR rate ω or in its action threshold τ, or in some combination of the two, can generate the same change in action timing. This holds for τ and the integration rate α of non-autonomous TiRs as well, as long as the input force to the TiR is constant. For coupled oscillator TiRs, δ depends in complicated ways on the initial phases of the systems, the oscillator frequencies, and the strengths of phase coupling forces (putting aside oscillator amplitude dynamics).

For even a simple system of three gestures, there is a rich set of possible ways in which temporal control can be organized. How can the organization of control be inferred from empirical observations? What we call "noise" may be quite useful in this regard. An essential characteristic of natural speech is that it is unavoidably stochastic, and as a consequence, no two utterances are identical. We interpret stochastic forces here as variation across utterances in the influence of the surroundings on time-representing systems. Moreover, in modeling noise we distinguish between *global noise*—stochastic variation that affects all TiRs equally—and *local noise*—stochastic variation that differentially affects TiRs. This distinction is important because the relative amplitudes of local and global noise can influence timing patterns.

The analysis of stochastic variation below focuses on correlations of successive time intervals between gestural initiations in three-gesture systems. These intervals are referred to as Δ12 and Δ23. We examine correlations (henceforth "Δ-correlations") rather than interval durations, because correlations more directly reflect interactions between systems. Five different local and global noise levels were crossed, from 0 to a maximum level (see **Supplementary Material**: Simulations for further detail). **Figures 8A–F** show the structures of each model tested, and corresponding panels in **Figures 8A'–F'** show how Δ-correlation varies as a function of global and local noise levels. Each line corresponds to a fixed level of global noise, and horizontal values of points represent different local noise levels.

The "shared trigger" model (**Figure 8A**) shows that if both non-initial gestures are activated by feedback from the initial one, Δ-correlation is trivially equal to 1, regardless of noise. The reason for this is simply that the same TiR (here $\hat{1}$) activates g_2 and g_3. Note that this trivial correlation occurs for external feedback control as well (not shown). The coupled oscillators model (**Figure 8B**) is unique among the systems examined in that it always produces non-trivial positive correlations. The reason for this has to do with phase coupling. Even when oscillator frequencies are heterogenous due to local noise, phase-coupling forces stabilize the oscillators at a common frequency. As long as phase-coupling forces are strong, local noise has relatively small effects on the phase evolution of oscillators. Global frequency noise always leads to positive correlations because it results in simulation-to-simulation variation in frequency that equally influences Δ12 and Δ23, causing them to covary positively. However, a more complex analysis of correlation structure in the coupled oscillators model in Tilsen (2017) has shown that when coupling strengths are also subject to noise, the model can generate negative correlations.

The external and internal feedback "chain models" (**Figures 8C,D**) exhibit nearly identical, complex patterns of correlation that depend on the relative levels of global and local noise. The patterns are nearly identical because the two models are topologically similar—they are causal chains—differing only in regard to the temporal delay associated with external sensory feedback. When there is no local noise, these chain models exhibit Δ-correlations of 1, since the global noise has identical effects on Δ12 and Δ23. Conversely, when there is no global noise, Δ-correlation is 0, since local noise has independent effects on Δ12 and Δ23. In-between those extremes, the correlation depends on the relative levels of local

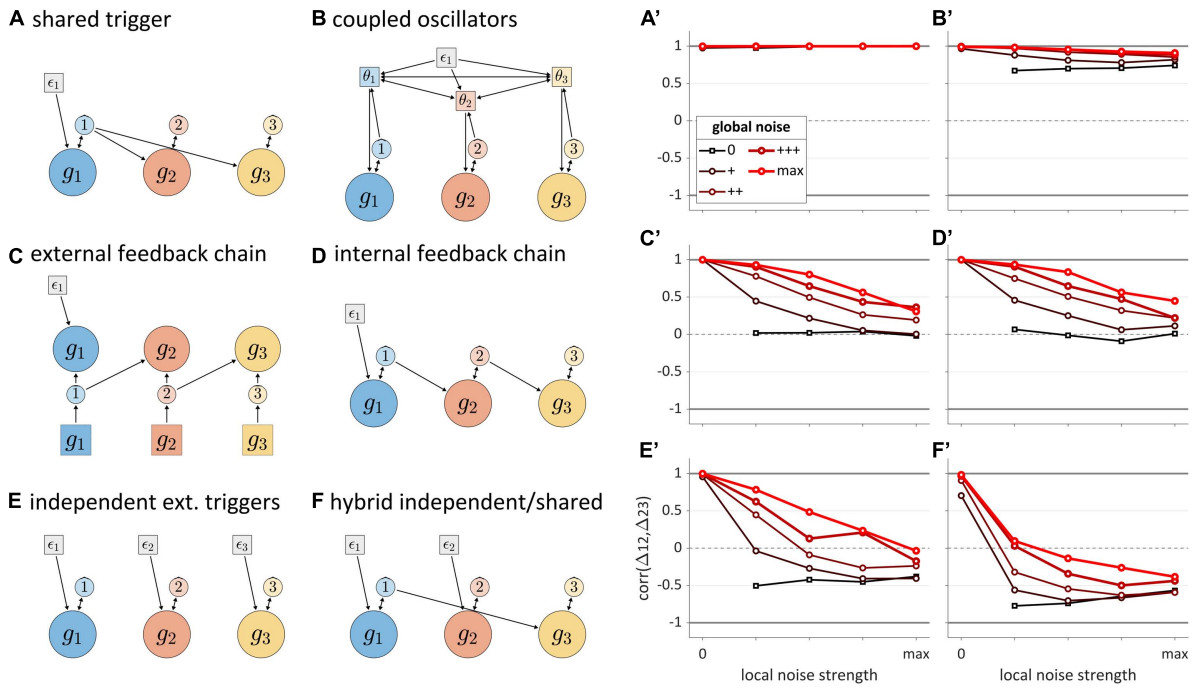


FIGURE 8
Noise-related correlation patterns for a variety of three-gesture systems. Panels **(A–F)** show model schemas and corresponding panels **(A'–F')** show correlations of intervals between initiation of gestural systems. Local noise levels increase along the horizontal axes, while global noise levels are indicated by the lines in each panel. Cases where both global and local noise are zero are excluded.

and global noise: increasing local relative to global noise leads to decorrelation of the intervals.

Unlike the other models, the independent extra-gestural triggers model (Figure 8E) and hybrid model (Figure 8F) can generate substantial negative correlations. In particular, negative correlations arise when g_2 is influenced by local noise. This occurs because whenever the TiR which activates g_2 does so relatively early or late, $\Delta 12$ and $\Delta 23$ will be influenced in opposite ways. Note that the negative correlations are stronger when the activation of g_1 and g_3 are caused by the same TiR, as is the case for the hybrid model (Figure 8F). At the same time, global noise induces positive Δ-correlation, counteracting the negative correlating effect of local noise. When we examine speech rate variation below, we will see that the opposing effects of global and local noise are not specific to "noise" *per se*: any source of variation which has similar effects on all TiRs tends to generate positive interval correlations, while the absence of such variation can lead to zero or negative correlation.

A hybrid model of gestural timing and speech rate control

Equipped with a new logic of temporal control, we now develop a hybrid model of gestural timing which is designed to accommodate a wide range of empirical phenomena. The primary requirement of the model is that for each gesture which is hypothesized to drive articulatory movement in an utterance, the model must generate commands to activate and deactivate that gesture.

Model space and hypotheses

For even a single CVC syllable, the set of all logically possible models is very large. Nonetheless, there are a number of empirical and conceptual arguments that we make to greatly restrict this space. Below we consider various ways in which gestural activation might be controlled for a CVC syllable uttered in isolation. Note that we adopt the modern "split-gesture" analysis in which constriction formation and constriction release are driven by separate gestural systems; this analysis has been discussed and empirically motivated in Nam (2007) and Tilsen, 2011, 2017. With that in mind we use the following gestural labeling conventions: C/c and R/r correspond to constriction formation and release gestures, respectively; upper case labels C/R correspond to pre-vocalic gestures (or, gestures associated with syllable onsets); lower case labels c/r correspond to post-vocalic gestures (or, gestures associated with syllable codas); and gestures/gesture pairs are subscripted according to the order in which they are initiated.

The schemas in **Figures 9A–C** show "extreme" models that—though logically possible—are conceptually and empirically problematic. **Figure 9A** shows a "maximally sensory" model, where all gestural activation/deactivation is controlled by external feedback systems. This model is problematic because the time delay between efferent motor signals and afferent feedback is too long to be useful for some relative timing patterns, such as the relative timing of consonantal constriction and release in normal speech. **Figure 9B** shows a "maximally internal" model, where all gestural activation and deactivation is induced by inter-gestural TiRs (keeping in mind that initiation of activation of the first gesture in an utterance is always external). The maximally internal model is problematic because it has no way of allowing for external/sensory feedback to influence timing.

Schema (**Figure 9C**) shows an "oscillator triggered" model, where all gestures are activated by coupled oscillators. Under standard assumptions, this model is problematic because it cannot generate some empirically observed combinations of pre-vocalic and post-vocalic consonantal timing, as discussed in Tilsen (2018). For example, in a CVC syllable, the temporal intervals between the initiation of the vocalic gesture and the initiations of onset and coda consonantal gestures cannot be produced by a system of oscillators that govern all three of these events, given certain constraints on oscillator frequency, triggering, and coupling. The "standard" assumptions are: (i) that all oscillators have (approximately) the same frequency; (ii) that all oscillators trigger gestural initiation at the same phase of their cycle; and (iii) that only in-phase and anti-phase coupling are allowed. With these constraints, the model cannot generate empirically common combinations of pre-vocalic and post-vocalic temporal intervals, where prevocalic CV intervals are generally in the range of 50–100 ms (Tilsen, 2017) and post-vocalic VC intervals—periods of time from V initiation to post-vocalic C initiation—are in the range of 150–400 ms. Moreover, relaxing any of the three assumptions may be undesirable. Allowing oscillators to have substantially different frequencies can lead to instability and chaotic dynamics, unless coupling forces are made very strong. Allowing oscillators to trigger gestures at arbitrary phases is inconsistent with the neurophysiological interpretation: presumably one particular phase of the cycle represents maximal population spike rate and should be associated with the strongest triggering force. Allowing for arbitrary relative phase coupling targets, such as a relative phase equilibrium of $3\pi/2$, may not be well-motivated from a behavioral or neurophysiological perspective.

Although the relatively extreme/monolithic models of **Figures 9A–C** are individually problematic, the mechanisms that they employ are practically indispensable for a comprehensive understanding of timing control. The hybrid control model (**Figure 9D**) is hypothesized to represent temporal control in typical adult speech. The model is described as "hybrid" because it uses coordinative/oscillator-based control for pre-vocalic timing, while allowing for internal or external feedback control for vocalic and post-vocalic timing. The model can be viewed as the combination of the following two more specific hypotheses:

> *Pre-vocalic coordinative control hypothesis.* Control of the initiation of pre-vocalic consonantal constriction formation (C), release (R), and vocalic (V) gestures is governed by a system of coupled oscillators.

> *Vocalic/post-vocalic feedback control hypothesis.* The deactivation of vowel gestures and the activation/deactivation of post-vocalic constriction (c) and release (r) gestures is governed by either internal or external feedback.

Below, we explain how each component of the model is motivated by a specific set of empirical phenomena.

Empirical motivation for pre-vocalic oscillator-based control

A major rationale for oscillator-triggered control is the phenomenon of symmetric displacement patterns (Tilsen, 2017, 2018). Such patterns were first described as the "c-center effect" in syllables with complex onsets (Browman and Goldstein, 1988). For a syllable with the form C1C2V, studies from a variety of languages have observed that the movements associated with the formation of the C1 constriction precede the movement associated with the vocalic posture, while the movements associated with the C2 constriction follow the movements associated with the vocalic posture; the C1 and C2 movement initiations tend to be approximately equally displaced in opposite directions in time from the initiation of the vocalic movement (Sproat and Fujimura, 1993; Byrd, 1995, 1996; Honorof and Browman, 1995; Kuhnert et al., 2006; Goldstein et al., 2007; Marin and Pouplier, 2010; Hermes et al., 2011, 2013; Tilsen et al., 2012). The pattern is remarkable because the order in which articulatory movements are initiated in such forms deviates from the order of segments in linear symbolic representations. The understanding of the c-center effect was significantly generalized by Nam (2007) and Tilsen (2017), where it was shown that a similar pattern of temporal displacement applies to the formation and release of the consonantal constriction in simple CV syllables: the constriction formation and release are displaced in opposite directions in time from the initiation of the vocalic movement. The only mechanism that has been proposed to explain symmetric displacement patterns is one in which the initiations of the gestures are governed by a system of coupled oscillators. With a combination of repulsive phase coupling between the oscillators that trigger consonantal gestures and attractive phase coupling between consonantal and vocalic oscillators, such a system

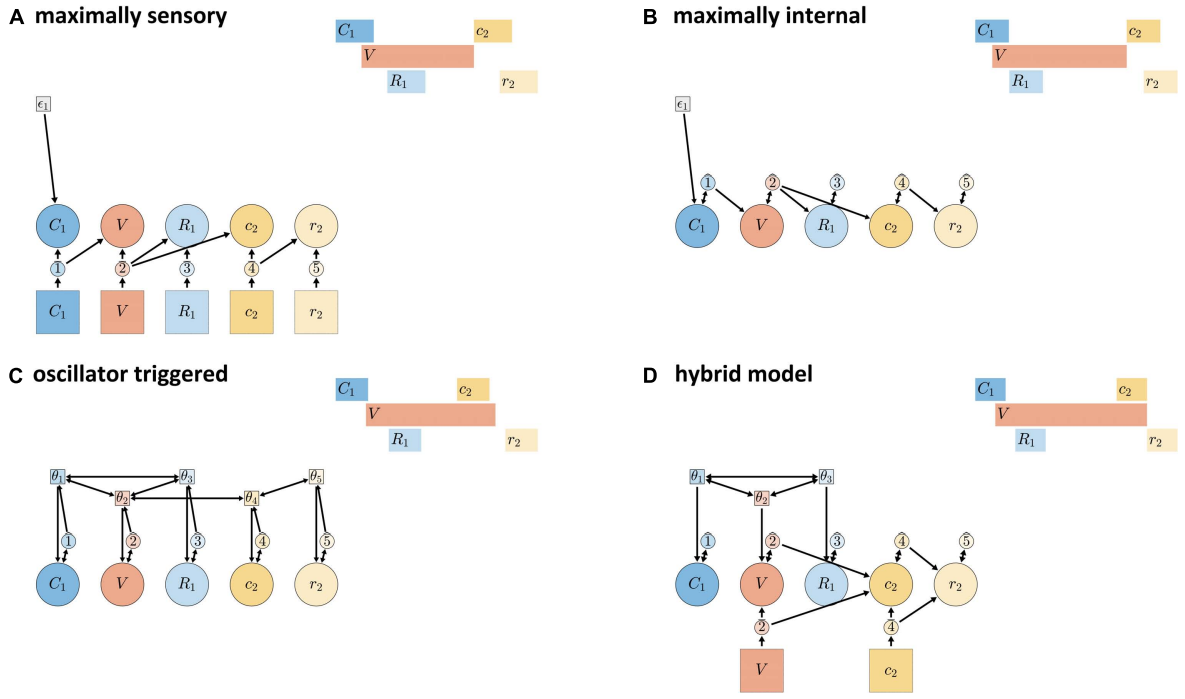


FIGURE 9
Candidate models of CVC syllables. **(A)** Maximally sensory model where all activation and deactivation is controlled by external sensory feedback. **(B)** Maximally internal model where all control is governed by internal feedback. **(C)** Fully oscillator-triggered model where all gestures are initiated by oscillators. **(D)** Hybrid model in which pre-vocalic gestural activation is oscillator-governed while post-vocalic activation is governed by either internal or external feedback.

naturally evolves toward a steady-state in which consonantal oscillator phases are displaced in opposite directions from the vocalic oscillator phase. Although the existence of symmetric displacement timing patterns does not prove that oscillators govern gestural timing, it is important to recognize that there exist no alternative models of these pervasive patterns.

A more indirect motivation for oscillator-triggered control comes from the observation that in the babbling stage of speech development, children employ an oscillatory cycle of jaw opening and closing to bootstrap the acquisition of CV syllables (MacNeilage et al., 1997; MacNeilage and Davis, 2000; Oller, 2000; Iverson et al., 2007). Furthermore, several studies have reported a coincidence of rhythmic activities in speech and non-speech domains (Thelen, 1979; Eilers et al., 1993; Iverson et al., 2007). It was argued in Tilsen (2014) that the oscillatory character of babble and its relation to oscillatory behaviors in non-speech actions suggest that oscillatory systems control the initiation of articulatory movements in CV forms.

Empirical motivation for vocalic/post-vocalic external feedback control

The primary motivation for incorporating external feedback control systems in the model is the common observation that word durations are lengthened in the presence of feedback perturbations (Houde and Jordan, 1998; Larson et al., 2001; Purcell and Munhall, 2006; Villacorta et al., 2007; Tourville et al., 2008; Cai et al., 2011). Such durational changes occur when auditory feedback is naturally or artificially degraded, and this occurs even in laboratory studies in which speakers are not accommodating listeners; the effect is known to be at least partly involuntary (Garnier et al., 2010; Zollinger and Brumm, 2011; Luo et al., 2018). It follows that there must be some temporal control mechanism that is responsible for increases in word duration in the absence of sensory feedback. The external feedback systems hypothesized to control the timing of vocalic/post-vocalic gestures are a minimal expansion of the model and are necessary for modeling the temporal effects of feedback perturbations.

Furthermore, recent evidence indicates that the temporal effects of auditory feedback perturbations are specific to vocalic/post-vocalic timing. The study in Oschkinat and Hoole (2020) found that post-vocalic intervals respond to temporal perturbations of feedback and that pre-vocalic intervals do not; specifically, subtle temporal delays of feedback imposed during a complex onset did not induce compensatory timing adjustments, while the same perturbations applied during a complex coda did. Another recent study (Karlin et al., 2021) found that temporal perturbations induced compensatory adjustments of vowel duration but not of onset consonant duration. Although the hybrid character of the model is a

complication compared to purely feedforward or feedback control structure, it seems necessary to account for the dissociation in feedback sensitivity that was observed by these studies.

Moreover, there are a host of more indirect reasons for dissociating pre-vocalic and vocalic/post-vocalic control mechanisms. These are discussed in depth in Tilsen (2016) but are briefly re-iterated here. First, the coarticulatory patterns exhibited by young children differ substantially between pre-vocalic and post-vocalic contexts: children show hyper-coarticulatory patterns between CV but hypo-coarticulatory patterns between VC (Kent, 1983; Hawkins, 1984; Repp, 1986; Goodell and Studdert-Kennedy, 1993; Sussman et al., 1999; Katz and Bharadwaj, 2001). Second, the patterns of sequencing errors exhibited by children in the early word stage are highly asymmetric for onsets and codas [see section 3.2 of Tilsen (2016) for a comprehensive analysis]. Third, a unified understanding of several forms of typological variation in syllable structure is made possible by hypothesizing pre-/post-vocalic asymmetries in the use of feedback for temporal control (Tilsen, 2016).

Empirical motivation for internal feedback control

The primary motivation for including internal feedback control systems in addition to external ones is the observation that temporal control is possible under circumstances in which external feedback is not available, for example during loud cocktail parties, for speakers with complete hearing loss, or during subvocal rehearsal (internal speech) with no articulatory movement. Thus in order for a model of temporal control to be empirically adequate, it is necessary to include internal feedback systems. There is a wide range of argumentation and evidence for the use of internal feedback control of movement, both generally (Miall and Wolpert, 1996; Kawato and Wolpert, 1998; Kawato, 1999; Thoroughman and Shadmehr, 1999; Shadmehr and Krakauer, 2008) and specifically in speech motor control (Guenther and Perkell, 2004; Max et al., 2004; Hickok, 2012; Guenther and Hickok, 2016; Parrell et al., 2018). Moreover, internal feedback systems are incorporated in a variety of speech production models (Tourville and Guenther, 2011; Hickok, 2012; Guenther and Hickok, 2016). However, most of the studies providing evidence for internal feedback control focus on the control of movement *via* predictive state estimation and error correction. These functions are instances of control *from* intentions, rather than control *of* (the timing of) intentions.

The inclusion of internal TiRs for control of timing in the hypothesized model follows from the reasoning that, in the absence of external feedback, some mechanism is needed to govern timing. For the reasons discussed above, this mechanism cannot be oscillator-based control. Because internal feedback systems are already motivated by their role in predictive state estimation and error correction, they are a natural candidate for a parsimonious model of timing control.

External influences on parameters

Here and following sections, some specific predictions of the hypotheses are examined. A key point about the model is that parameters of TiRs are context-dependent: they vary in ways that are conditioned on factors associated with TiR system surroundings, so-called "external factors." Here we demonstrate two ways in which external factors may influence timing. An innovation of the model is the idea that these factors can have differential influences on external vs. internal TiR parameters.

Figures 10A–C demonstrate the effects of variation in a hypothetical contextual factor of *self-attention*, or "attention to one's own speech," which is represented by a variable, λ. The self-attention variable λ ranges from 0 to 1, where 0 represents minimal attention to one's own speech, and 1 represents maximal attention. The figure summarizes simulations of the system shown in **Figure 10A**, where activation of a post-vocalic constriction gesture c_1 is potentially caused by an internal or external TiR representing feedback from the vocalic gesture V_1. This is the hypothesized organization of post-vocalic control in the hybrid model. By hypothesis, the force integration rates of internal and external TiRs are differentially modulated by self-attention λ, such that $\alpha = \alpha'/(1 + \beta\lambda)$, where $\beta_{internal} < \beta_{external}$. This reflects the intuition that when one attends to feedback more closely, feedback-accumulation (i.e., force-integration) rates of TiR systems are diminished, so that TiRs take longer to act on gestures. This diminishing effect applies more strongly to internal feedback than external feedback. As a consequence, there is a value of λ such that as λ is increased, initiation of g_2 switches from being governed by the internal TiR to the external one. In the example, the transition occurs around $\lambda = 0.425$, where a change is visible in the slope relating the control parameter λ and the interval δ (the time between initiation of V_1 and c_1). Gestural activation intervals associated with three values of λ are shown in **Figure 10C**.

Figure 10B shows that when TiR parameters are differentially modulated by an external influence, transitions between internal and external feedback control can occur. In the above example, the external influence was posited to represent "self-attention" and its state was encoded in the variable λ. This variable was then hypothesized to differentially adjust external vs. internal non-autonomous TiR growth rates. Another way in which the same effect can be derived is by allowing the external variable λ to differentially adjust TiR action-thresholds.

Another parameter that can respond to external factors is the frequency of the coupled oscillators which are hypothesized to govern prevocalic gestural initiation, as in **Figure 10D**. Suppose that the external factor here is a mechanism that controls oscillator frequency *via* an external variable called "pace." As with self-attention, the external variable of pace ranges from 0 to 1, with 0 corresponding to minimal pace and 1 corresponding to maximal pace. However, because of the

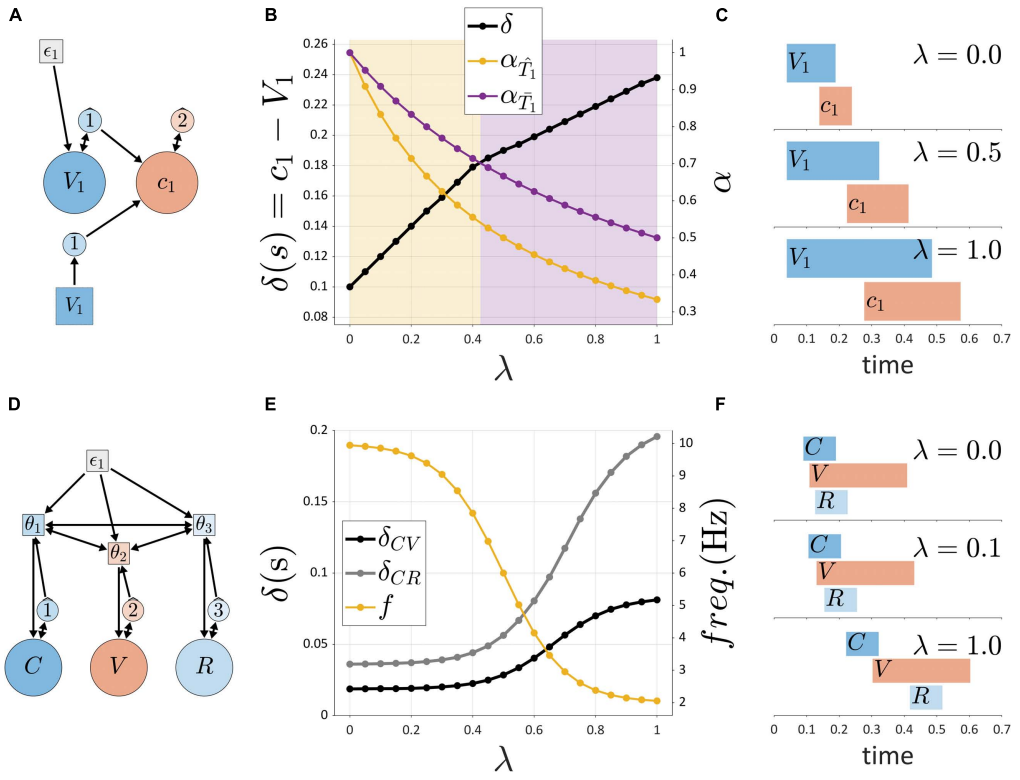


FIGURE 10
Simulations of external influences on parameters. **(A)** Schema for post-vocalic control with both internal and external TiRs. **(B)** Dual axis plot showing how δ (left side) and integration rates α (right side) change with self-attention parameter λ. **(C)** Gestural activation intervals for several values of λ. **(D)** Model schema of pre-vocalic coordinative control. **(E)** Dual axis plot showing effect of rate parameter λ on δ-values (left side) and frequencies (right side). **(F)** Gestural activation intervals for several values of λ.

frequency constraint hypothesis, we cannot simply allow the oscillator frequencies to respond linearly to changes in pace. Instead, we impose soft upper and lower frequency bounds by attenuating the effect of the pace parameter λ on frequency *f*. This is accomplished by making the effective frequency a non-linear function of λ, as shown in **Figure 10E** (right side) and **Figure 10F**. The consequence of this limitation on *f* is that intervals which are governed by coordinative control are predicted to exhibit non-linear responses to variation in the external factor: here we can see that the δ_{CV} and δ_{CR} plateau at extreme values of λ.

In the section "A model of speech rate control with selectional effects," we combine the above effects of self-attention and pace into a general model of the control of speech rate. But first we introduce another important mechanism, which allows the model to organize the subsystems of larger utterances.

Parallel domains of competitive selection

Competitive selection (or competitive queuing) is a dynamical mechanism that, given some number of actions, iteratively selects one action while preventing the others from being selected. The concept of competitive selection of actions originates from Grossberg (1987), and many variations of the idea have been explored subsequently, both within and outside of speech (Bullock and Rhodes, 2002; Bullock, 2004; Bohland et al., 2010; Bhutani et al., 2013; Tilsen, 2013; Glasspool, 2014; Kristan, 2014). One of the key ideas behind the mechanism is that a serial order of actions is encoded in an initial activation gradient, such that prior to the performance of an action sequence, the first action in the sequence will have the highest relative activation, the second action will have the next highest activation, and so on. The growth of activation is a "competition" of systems to be selected, and selection is achieved by reaching an activation threshold. Moreover, action selection is mutually exclusive, such that only one action can be selected at a time.

Figure 11 shows how these ideas are understood in the current model. The "actions" which are competitively selected in this example are three CV syllables, and the selection of these actions is governed by systems that we refer to as μ-*systems*. As shown in the model schema, each μ-system de-gates a system of coupled oscillators, which in turn activate gestures. Each of the μ-systems is associated with a μ-gating

system that—when open—allows the corresponding μ-system activation to grow. Notice that at time 0 (before the production of the sequence), the pattern of relative activation of μ-systems corresponds to the order in which they are selected. When μ-system gates are open, μ-system activations grow until one of the systems reaches the selection threshold. At this point, all μ-gating systems are closed, which halts growth of μ-system activation. The selected μ-system is eventually suppressed (its activation is reset to 0) by feedback—specifically by the inter-gestural TiR associated with the last gesture of the syllable, in this case the vowel gesture. This causes all μ-systems to be de-gated, allowing their activations to grow until the next most highly active μ-system reaches the selection threshold. This three-step process—(i) de-gating and competition, (ii) selection and gating of competitors, and (iii) feedback-induced suppression of the selected system—iterates until all of the μ-systems have been selected and suppressed. See **Supplementary Material**: Model details for further information regarding the implementation.

A more abstract depiction of a competitive selection trajectory is included in the activation potentials of **Figure 11**. The potentials without arrows are relatively long epochs of time in which μ-systems exhibit an approximately steady-state pattern of activation. The potentials with arrows correspond to abrupt intervening transitions in which the relative activation of systems is re-organized by the competitive selection/suppression mechanism. Along these lines, the dynamics of competitive selection have been conceptualized in terms of operations on discrete states in Tilsen, 2019a,c.

There are two important questions to consider regarding the application of a competitive selection mechanism to speech. First, exactly what is responsible for suppressing the currently selected μ-system? In the example above, which involves only CV-sized sets of gestures, it was the internal TiR associated with the last gesture of each set. Yet a more general principle is desirable. Second, what generalizations can we make about the gestural composition of μ-systems? In other words, how is control of gestural selection organized, such that some gestures are selected together (*co-selected*) and coordinatively controlled, while others are competitively selected *via* feedback mechanisms? This question has been discussed extensively in the context of the Selection-coordination theory of speech production (Tilsen, 2014, 2016), where it is hypothesized that the organization of control follows a typical developmental progression. In this progression, the use of external sensory feedback for suppression/de-gating is replaced with the use of internal feedback, a process called *internalization of control*.

The are two important points to make about internalization. First, internalization of control is partly optional, resulting in various patterns of cross-linguistic and inter-speaker variation which are detailed in Tilsen (2016) and which we briefly discuss in the section "No direct control of the timing of target achievement." Second, internalization is flexible within and across utterances, such that various contextual factors (e.g., self-attention) can influence whether external or internal feedback TiRs are responsible for suppressing selected μ-systems.

Furthermore, a recently developed theory of syntactic organization in speech (Tilsen, 2019c) argues that there are two interacting domains of competitive selection. This is known as the *parallel domains hypothesis*. One of these domains involves "gestural-motoric" organization of the sort illustrated above, where gestures are organized into competitively selected sets (μ-systems). The other involves "conceptual-syntactic" organization in which concept systems are organized into competitively selected sets. The hypotheses advanced in Tilsen (2019c) hold that sets of co-selected conceptual systems correspond loosely to the prosodic unit called the *phonological word* (a.k.a. p-wrd, or ω), which has the property that there is a single accentual gesture associated with set of co-selected conceptual systems. Moreover, under normal circumstances speakers do not interrupt (for example by pausing) the gestural competitive selection processes which are induced by selection of a phonological word.

These parallel domains of conceptual-syntactic and gestural-motoric competitive selection are illustrated **Figure 12** for an utterance which would typically be analyzed as four prosodic words, such as [*a dog*] [*and a cat*] [*chased*] [*the monkey*]. Note that to conserve visual space release gestures have been excluded. The top panel shows the sequence of epochs in competitive selection of concept systems $\mathcal{C}$. Each of these could in general be composed of a number of co-selected subsystems (not shown). For each epoch of concept system selection, there is a corresponding series of one or more epochs of competitive selection of gestural systems. The model accomplishes this by allowing the concept systems to de-gate the corresponding sets of μ-systems. Within each of these sets of μ-systems, the appropriate initial activation gradient is imposed. Further detail on the implementation is provided in the **Supplementary Material**.

Although there is no *a priori* constraint on the number of domains of competitive selection that might be modeled, the parallel domains hypothesis that we adopt makes the strong claim that only two levels are needed—one for conceptual-syntactic organization and one for gestural-motoric organization. We examine some of the important consequences of these ideas in Section "Reinterpretation of prosodic phrase structure and boundaries," regarding phrasal organization. One aspect of prosodic organization which we do not elaborate on specifically in this manuscript involves the metrical (stress-related) organization of gestures, but see Tilsen (2019b) for the idea that the property of "stress" relates to which sets of co-selected gestures (μ-systems) may include accentual gestures, which in turn are responsible for transient increases in self-attention.

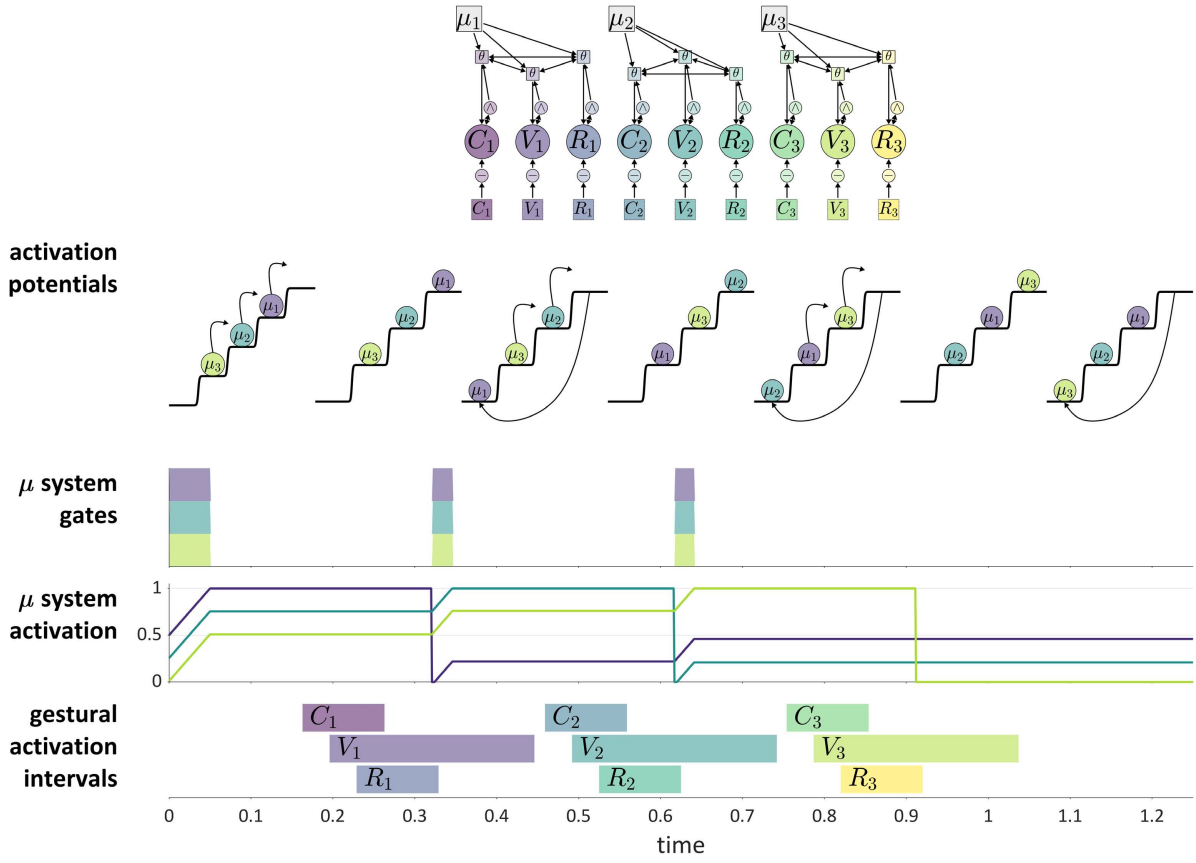


FIGURE 11
Illustration of competitive selection for a sequence of three CV syllables. **(Top)** Model schema. Activation potentials with arrows show transitions between states, and potentials without arrows shown quasi-steady states. μ-gating system states are shown (shaded intervals are open states). **(Bottom)** Gestural activation intervals.

A model of speech rate control with selectional effects

When given verbal instructions to "talk fast" or "talk slow," speakers are able to produce speech that listeners can readily judge to be relatively fast or slow. To quantify this sort of variation in tempo, speech rate is often measured as a count of events per unit time, e.g., syllables per second or phones per second. There are several important points to consider about these event-rate quantities, which call into question whether speakers control tempo as a rate of events, *per se*. First, in order to be practically useful, an event rate must be measured over a period of time in which multiple events occur. Hence event rates are unlikely to be controlled instantaneously, since such measures cannot be robustly defined in a moment-to-moment fashion. Second, there is no consensus on which events are the appropriate ones to count—phones, syllables, words, or something else? In the current framework, many commonly used units do not even have an ontological status. In order for an event rate to be controlled, it stands to reason that the relevant events should have some degree of cognitive reality. Third, even if we ignore the above problems, there is no evidence to my knowledge that speakers directly control rate quantities such as syllables/second or phones/second. Hence there is reason to doubt that the quantity which speakers attempt to control should be conceptualized as a rate of events. If speakers do not in fact control speech rate as an event rate *per se*, then what are speakers controlling in order to speak fast or slow?

The *attentional modulation hypothesis* (Tilsen, 2018) holds that speakers control rate by modulating their attention to feedback of their own speech (*self-attention*), and specifically do so in a way that, as self-attention increases, prioritizes external/sensory feedback over internal feedback. Furthermore, along with modulating self-attention, speakers may adjust pacing, that is, the frequencies of gestural planning oscillators. The separate effects of varying these external factors were already demonstrated in Section "External influences on parameters."

In addition, a mechanism is needed to account for the phenomenon of boundary-related lengthening. Many empirical

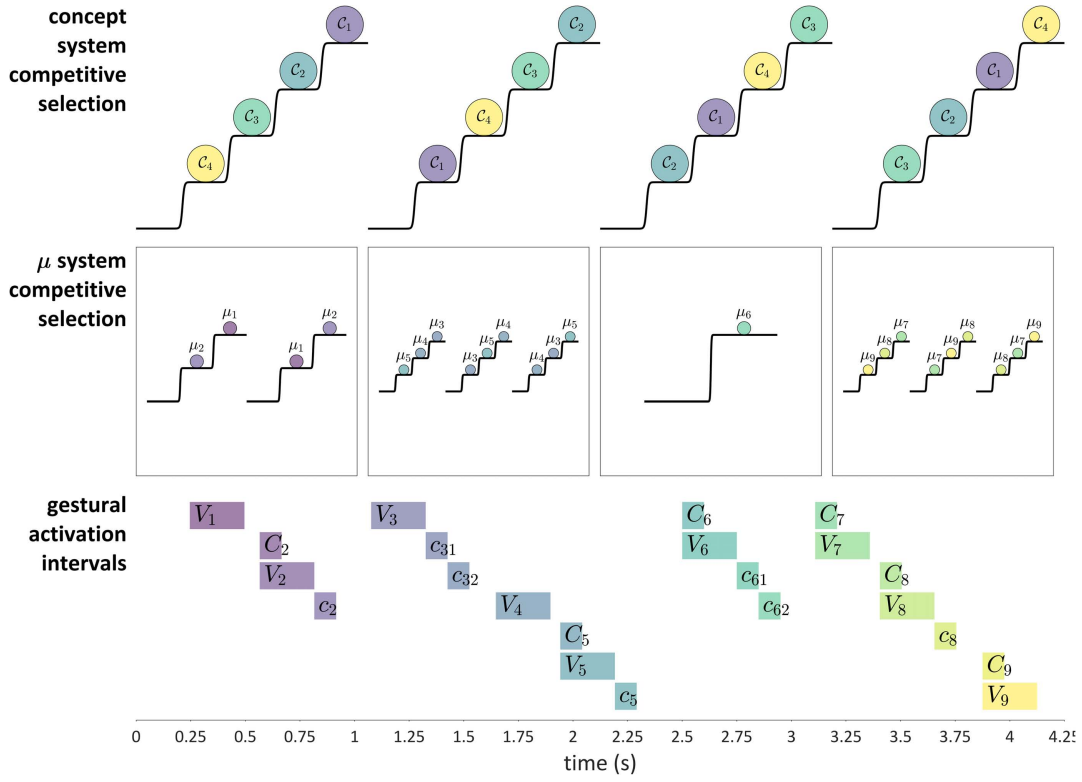


FIGURE 12
Illustration of parallel domains of competitive selection for an utterance with the structure. **(Top)** Concept systems C are competitive selected. **(Middle)** Selection a concept system de-gates corresponding μ-systems which themselves are competitively selected. **(Bottom)** Gestural activation intervals generated by the model.

studies have shown that speech slows down as speakers approach the ends of phrases, with greater slowing and increased likelihood of pausing statistically associated with "higher-level" phrase boundaries (Byrd and Saltzman, 1998, 2003; Byrd, 2000; Byrd et al., 2006; Turk and Shattuck-Hufnagel, 2007, 2020b; Krivokapić, 2014). One approach to understanding the mechanism responsible for such effects is the π-gesture model of Byrd and Saltzman (2003), in which it was hypothesized that boundary-related lengthening is caused by a special type of clock modulating system, a "π-gesture." This clock-modulating system, when active, slows down the rate of a hypothesized nervous system-internal global clock, relative to real time. Gestural activation dynamics evolve in the internal clock coordinate, and so gestural activation intervals are extended in time when a π-gesture is active. Furthermore, it was suggested in Byrd and Saltzman (2003) that the degree of activation of a π-gesture varies in relation to the strengths of prosodic boundaries, such that stronger/higher-level boundaries are associated with greater π-gesture activation and hence more slowing.

How can the phenomenon of boundary-related lengthening be conceptualized in the current framework, where there is no global internal clock for gestural systems? A fairly straightforward solution is to recognize that in effect, each gestural system has its own "local clocks," in the form of the internal and external feedback TiRs, whose integration rates are modulated by self-attention. In that light, it is sensible to adapt the π-gesture mechanism by positing that self-attention effects on TiR parameters tend to be greater not only in the final set of gestures selected in each prosodic word (i.e., final μ-system), but also in the final set of co-selected conceptual systems (i.e., the final μ-system). As for why it is the final set of selected systems that induces these effects, we reason that speakers may attend to sensory feedback to a greater degree when there are fewer systems that remain to be selected. At the end of an utterance, there are no more systems that remain to be selected, and thus self-attention is greatest. We refer to this idea as the *selectional anticipation hypothesis*, because anticipation of upcoming selection events is proposed to distract a speaker from attention to feedback of their own speech. Although this hypothesis is admittedly a bit ad hoc, and alternative explanations should be considered, we show below that the implementation of this idea is sufficient to generate the lengthening that occurs at the ends of phrases.

Putting the above ideas together, **Figure 13** shows how interval durations change as a function of attentional

modulation. The utterance here is a competitively selected sequence of three syllables with forms CVC, CV, CVC, as shown in **Figure 13A**. Note that the organization of each syllable conforms to the hybrid control model, entailing that pre-vocalic timing is coordinative and vocalic/post-vocalic timing is feedback-based. As in the section "External influences on parameters," the integration rates of external (sensory) and internal TiRs, along with oscillator frequencies, are made to vary in response to changes in a control parameter λ; these relations are shown in **Figure 13B**. In addition, the integration rate parameters associated with the final set of gestures are even more strongly modulated by λ (dotted lines of **Figure 13B**), to implement the selectional anticipation hypothesis; the consequences of this are evident in the contrast between word 1 and word 3 durations in **Figure 13D**. The initiation times of gestures for each of the 11 values of λ that were simulated are shown vertically in **Figure 13C**.

By simulating variation in speech rate, we are able to generate some of the most essential predictions of the hybrid control model, introduced in Section "Model space and hypotheses." Recall that this model combined two hypotheses: prevocalic coordinative control and post-vocalic feedback-control. These hypotheses are associated with the following three predictions:

(i) *Prevocalic attenuation*. The prevocalic coordinative control hypothesis holds that initiation of the prevocalic constriction and release gestures, along with initiation of the vocalic gesture, is controlled by a system of coupled oscillators. Moreover, the frequency constraint hypothesis was shown in the section "External influences on parameters" to predict that intervals between these initiations attenuate as rate is increased or decreased. This effect can be seen in **Figure 13E** for the C_3-R_3 interval, which is the interval between constriction formation and release. In other words, the prediction is that prevocalic timing is only so compressible/expandible, no matter how quickly or slowly a speaker might choose to speak.

(ii) *Postvocalic expandability*. Conversely, the post-vocalic feedback-control hypothesis holds that there is a transition from internally to externally governed control, and that there should be no limits on the extent to which increasing self-attention can increase the corresponding interval durations. This prediction is shown in **Figure 13E** for the R_3-c_3 interval (which loosely corresponds to acoustic vowel duration) and the c_3-r_3 interval (related to constriction duration). These intervals continue to increase as attention to feedback is increased.

(iii) *Sensitivity to feedback perturbation*. Finally, a third prediction of the model is that, when external feedback governs post-vocalic control (as is predicted for slow rates), perturbations of sensory feedback will influence post-vocalic control but not prevocalic control.

How do these predictions fare in light of current evidence? The ideal tests of predictions (i) and (ii) require measurements of temporal intervals produced over a wide range of variation in global speech rate. Unfortunately, most studies of the effects of speech rate do not sufficiently probe extremal rates, since many studies use categorical adverbial instructions (e.g., *speak fast* vs. *speak normally* vs. *speak slowly*). One exception is a recent study using an elicitation paradigm in which the motion rate of a visual stimulus iconically cued variation in speech rate (Tilsen and Hermes, 2020). Utterance targets were words with either intervocalic singleton or geminate bilabial nasals (/ima/ and /imma/). The study observed that the timing of constriction formation and release of singleton /m/ exhibited a non-linear plateau at slow rates, similar to the prediction for the c_3-r_3 interval in **Figure 13E**. This is expected given the assumption that the formation and release gestures are organized in the onset of the second syllable of the target words. In contrast, the durations of constriction formation-to-release intervals of geminate /mm/ did not attenuate: they continued to increase as rate slowed. This is expected if the initiation of the geminate bilabial closure is associated with the first syllable and its release with the second. Although the dissociation of effects of rate on singletons vs. geminates is not the most direct test of the hybrid model hypothesis, it shows that more direct tests are warranted.

Regarding prediction (iii), a recent study has indeed found evidence that post-vocalic intervals respond to temporal perturbations of feedback and that pre-vocalic intervals do not (Oschkinat and Hoole, 2020). This study found that subtle temporal delays of feedback imposed during a complex onset did not induce compensatory timing adjustments, while the same perturbations applied during a complex coda did. This dissociation in feedback sensitivity is a basic prediction of the hybrid model. Another recent study (Karlin et al., 2021) has found that temporal perturbations induced compensatory adjustments of vowel duration but not of onset consonant duration (codas were not examined). There may be other reasons why temporal feedback perturbations have differential effects on prevocalic and vocalic/post-vocalic intervals, and certainly there is much more to explore with this promising experimental paradigm. Nonetheless, effects that have been observed so far are remarkably consistent with the predictions of the hybrid control model.

General discussion

The informal logic developed here has many consequences for phonological theories. Below we discuss an important point about control of target timing along with two of the most important consequences of the model. First, the framework does not allow for direct control over the timing of articulatory target achievement, and we will argue that this is both conceptually desirable and empirically consistent. Second, structural entities

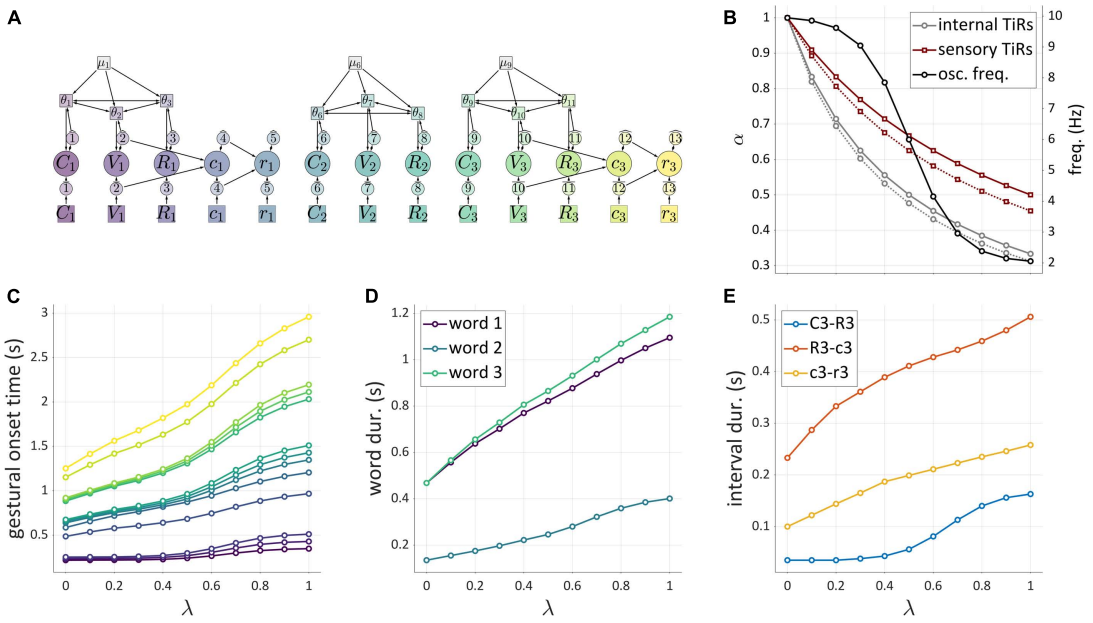


FIGURE 13
Simulation of variation in speech rate, as controlled by correlated chagnes in self-attention and pacing, both indexed by λ. **(A)** Model schema showing three syllables with the forms CVC, CV, and CVC. **(B)** Relations between λ and feedback TiR integration rates (α) and oscillator frequencies. **(C)** Times of gestural initiation for each value of λ simulated. **(D,E)** Word durations and interval durations of the third word.

such as syllables and moras can be re-interpreted in relation to differences in the organization of control. Third, there is no need to posit the existence of different types of phrases, nor a hierarchical organization of phrases: the appearance of prosodic "structure" above the phonological word can reinterpreted more simply as variation in self-attention conditioned on selection of prosodic words.

No direct control of the timing of target achievement

Some researchers in the AP/TD framework have explicitly hypothesized that control of timing of target achievement is a basic function available in speech (Gafos, 2002), or have implicitly assumed such control to be available (Shaw et al., 2011). More generally, outside of the AP/TD framework, it has been argued that speakers prioritize control of the timing of articulatory and acoustic target events over control of the initiation of the very same actions that are responsible for achieving those targets (Turk and Shattuck-Hufnagel, 2014, 2020a,b). "Target achievement" is defined here as an event in which the state of the vocal tract reaches a putative target state that is associated with a gestural system.

Direct control of the timing of gestural target achievement is prohibited by our logic because TiRs control when gestural systems become active and cease to be active, and neither of these events fully determines the time at which targets are achieved. The TiR framework of course allows for *indirect* control of target achievement timing, *via* the trivial fact that target achievement depends in part on when a gesture is activated. Yet other factors, which are outside the scope of the TiR model, play a role as well. In standard TD (Saltzman and Munhall, 1989) these factors include the strengths of the forces that gestural systems exert on a tract variable systems—both driving forces and dissipative damping forces—as well as how these forces are blended when multiple gestural systems are active. Or, in an alternative model of how gestures influence tract variable control systems (Tilsen, 2019a), the relevant factors are the strengths, timecourses, and distributions of inhibitory and excitatory forces that gestural systems exert on spatial fields that encode targets. In either case, target achievement cannot be understood to be controlled directly by TiRs.

A major conceptual issue with direct control of the timing of target achievement is that it requires an unrealistically omniscient system that has accurate knowledge of the future. In order to control exactly when a target is achieved, a control system must initiate a movement at precisely the right time, which in turn requires that the system is able to anticipate the combined influences on the vocal tract state of all currently active subsystems and all subsystems which might become active in the near future. This all-knowing planner must accomplish these calculations before the critical time at which the movement must be initiated. While such calculations are not in principle impossible, they do require a system which has access to an implausibly high degree of information from many subsystems.

A primary empirical argument for direct control of target achievement is premised on the claim that there is less variability associated with timing of target achievement than variability associated with timing of movement onsets. This is argued in Turk and Shattuck-Hufnagel, 2014 to suggest that timing of target achievement is not only independently controlled, but also prioritized over timing of movement initiation. The difference in variability upon which the argument is premised has been observed in non-speech studies in which an actor must hit or catch a moving object. Yet these sorts of non-speech examples do not necessarily translate to speech, because in articulation there are no uncontrolled moving objects that the effectors must collide with at the right place in space and time—speech is simply not like catching a ball. Indeed, only one study of speech appears to have concluded that there is less variability in target vs. initiation timing (Perkell and Matthies, 1992), and this interpretation of the data is highly questionable due to differences in how the two events were measured.

Empirically observed phonetic and phonological patterns indeed provide the strongest argument *against* direct control of target achievement timing. Phonetic reduction of targets, which can arise from insufficient allotment of time for a target to be achieved, is rampant in speech. The "perfect memory" example of Browman and Goldstein (1990) shows how at fast speech rates the word-final [t] can be not only acoustically absent but also quite reduced kinematically when the preceding and following velar and bilabial closures overlap. If speakers prioritized the timing of the [t] target relative to either the preceding or following targets, this sort of reduction presumably would happen far less often. The prevalence of historical sound changes that appear to involve deletion of constriction targets argues against the notion that speakers are all that concerned with achieving targets. Certainly, the consequences of failing to achieve a target are usually not so severe: in order to recognize the intentions of speakers, listeners can use semantic/contextual information and acoustic cues that are not directly related to target achievement. Rather than being a priority, our informal logic views target achievement as an indirect and often not-so-necessary consequence of activating gestural systems.

Reinterpretation of syllabic and moraic structure

Many phonological theories make use of certain entities—syllables (σ) and moras (μ)—as explanatory structures for phonological patterns. These entities are viewed structurally as groupings of segments, with moras being subconstituents of syllables, as was shown in **Figure 2B**. Selection-coordination theory (Tilsen, 2014, 2016) has argued that these entities, rather than being parts of a structure, should be thought of as different classes of phonological patterns that are learned in different stages of a particular developmental sequence, over which the organization of control changes. This idea is referred to as the *holographic hypothesis*, because it holds that what appears to be a multi-level structure of syllables and moras is in fact a projection over developmental time of two different forms of organization which do not exist simultaneously. This is loosely analogous to a hologram, which encodes a three-dimensional image in two dimensions.

The holographic hypothesis is exemplified in **Figure 14** (top) for a CVC syllable. Early in development, the post-vocalic constriction gesture is controlled entirely by sensory feedback (i.e., extra-gestural TiRs), and so phonological patterns learned at this time are associated with a moraic structure, reflecting a stronger differentiation in control of pre-vocalic and post-vocalic articulation. Subsequently, speakers learn to activate and deactivate the post-vocalic constriction/release with internal TiRs, which is an instance of internalization. This leads to initiation of the post-vocalic constriction before termination of the vocalic gesture, and hence an increase in articulatory overlap/coarticulation. Phonological patterns learned in conjunction with this internalized organization of control are associated with syllables, rather than moras. Similar reasoning applies to other syllable shapes such as {C}{CV}→{CCV} and {CV}{V}→{CVV}, where developmental transitions in the internalization of control can account for cross-linguistic phonetic and phonological variation (Tilsen, 2016).

Exactly what causes internalization and governs its progression are open questions that presumably relate to information transmission. More internalization is associated with a greater rate of information production in speech, or in other words, increased efficiency of communication. Conversely, too much internalization can result in degrees of articulatory overlap which sacrifice perceptual recoverability (Liberman et al., 1967; Fowler and Rosenblum, 1991; Chitoran and Goldstein, 2006; Gick et al., 2006), reflecting constraints on channel capacity. It is far from clear how these opposing considerations—information rate vs. channel capacity—might be mechanistically manifested in a model of utterance-timescale processes. Informational aspects of speech, which by definition require analysis of the space of possible state trajectories of gestural systems, necessarily involve attention to patterns on lifespan timescales and speech-community spatial scales. Thus the challenge lies in understanding how these relatively large timescale informational forces translate to changes in utterance-scale control.

Reinterpretation of prosodic phrase structure and boundaries

There are many prosodic theories in which prosodic words (ω) are understood to be hierarchically structured into various types of phrases. A "phrase" in this context simply refers

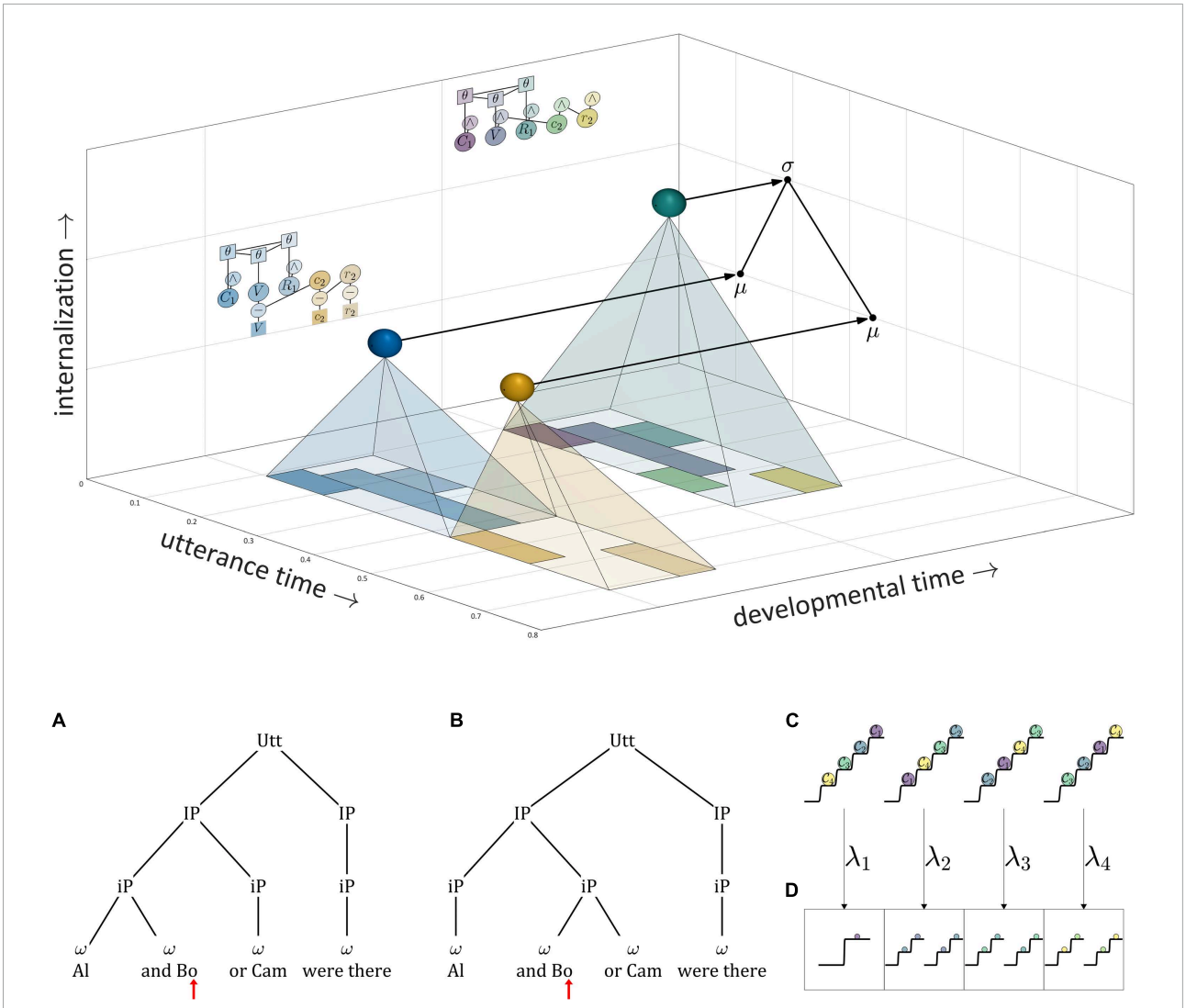


FIGURE 14
(Top) Visualization of the holographic hypothesis, for a CVC form. In an early stage of development, control over the post-vocalic constriction is based entirely on sensory feedback. Phonological patterns learned in this stage of development are described with moraic structure. In a later stage of development, control has been internalized, and phonological patterns learned in this stage are described with syllabic structure. **(Bottom)** Hierarchical prosodic structure reinterpreted as variation in attentional modulation of control parameters. **(A** vs. **B)** Alternative hierarchical prosodic structures purported to encode a difference in conceptual grouping. Red arrows indicate timepoint discussed in the text. **(C,D)** In different epochs of concept system selection, self-attention (λ) may differ, resulting in differences in temporal control.

to a grouping of prosodic words. Different types of phrases have been proposed, with two of the most popular being the "intonational phrase" (IP) and "intermediate phrase" (iP) from Beckman and Pierrehumbert (1986); these were shown in **Figure 2B**. Many theories additionally posit that these types of phrases can be recursively hierarchically organized (Ladd, 1986; Féry, 2010; Ito and Mester, 2013), such that a given type of phrase can contain instances of itself. In general, the motivations for positing phrase structures of this sort are diverse and too complex to address in detail here, but most of them relate either (i) to the likelihood that certain phonological patterns will occur in some portion of an utterance, or (ii) to statistical patterns in measures of pitch or duration observed in longer utterances.

To provide an example, consider the question: *Who was in the library?*, answered with the utterance *Al and Bo or Cam were there*. This response has two probable interpretations, and in many theories these would be disambiguated by the prosodic structures shown in **Figure 14** (bottom: **A vs. B**):

The motivation for positing the structural distinction between **Figure 14A** and **Figure 14B** is that it can account for certain empirical patterns related to conceptual grouping. Consider specifically the period of time in the vicinity of the red arrows, near the end of the production of *Bo*, which is

often conceptualized as a phrase "boundary." Here utterance **Figure 14A**, compared to **Figure 14B**, will tend to exhibit a larger fall of pitch, greater boundary-related lengthening, and a greater likelihood of a pause. The pitch of the following word may also start at a higher value. Hierarchical structural analyses hold that these differences occur because there is a "higher-level boundary" at this location in **Figure 14A** than in **Figure 14B**, that is, an intermediate phrase boundary vs. a prosodic word boundary.

The logic of multilevel competitive selection makes hierarchical or recursive phrasal structure unnecessary. If anything, our framework corresponds to a flat, anarchical organization of prosodic words—though more appropriately it rejects the notion that prosodic words are parts of structures in the first place, and "boundaries" are seen as wholly metaphoric. How can regularities in intonational patterns such as in **Figure 14A** vs. **Figure 14B** be understood, without the notions of phrase hierarchies and boundaries?

Recall that each prosodic word is one set of co-selected concept systems, which are associated with some number of sets of co-selected gestural systems (**Figure 11**). Furthermore, recall that boundary-related lengthening was interpreted as a decrease in integration rates of feedback TiRs, and this parameter modulation is proposed to be greater for the last set of systems in a competitively selected set (the selectional anticipation hypothesis), as shown for the word durations in **Figure 13D**. This reasoning leads to an alternative understanding of why there exists phonetic and phonological variation that correlates with prosodic organization: rather than being due to "structural" differences, the variation arises from differences in how TiR parameters are modulated for each prosodic word, as suggested by the arrows in **Figures 14C,D**. Rather than constructing a structure of prosodic words for each utterance, speakers simply learn to adjust self-attention in a way that can reflect conceptual relations between systems of concepts. Presumably many forms of discourse-related and paralinguistic information can be signaled in this way, including focus phenomena such as emphatic and contrastive focus. In other words, to emphasize information for listeners, speakers simply emphasize that information for themselves.

Conclusion

To conclude, we return to the initial questions of this manuscript: (i) what determines the duration of that *shush* that you gave to the loud person in the library, and (ii) how do you slow down the rant to your friend in the coffee shop? According to the feedback-based logic of temporal control, your *shush* duration is most likely determined by a sensory feedback-based control system (an external, non-autonomous TiR), and depending upon various factors (how angry you are, how far away the loud student is), you will diminish the integration rate of the TiR and/or increase its threshold to extend the duration of the sound. Later on in the coffee shop, you slow down your rant in effect by doing the same thing: increasing self-attention.

One possible criticism of the framework presented here is that it is too complex. While it is fair to assert that the model proposed here is complex compared to other models, this manuscript has shown that in all cases the complexity is warranted, in order to for the model to be empirically adequate. Simpler models are simply not able to generate the full range of temporal patterns which occur in speech. Given that empirically observed temporal patterns in speech are complicated, it is not surprising that the mechanisms used to generate speech must reflect that complexity.

There are several important conceptual and theoretical implications of our informal logic. First, all control of timing must be understood in terms of systems and their interactions, and this understanding involves the formulation of change rules to describe how system states evolve in time. Second, the systems which control timing do not "represent" time in any direct sense; the states of systems are defined in units of activation, and activation is never a direct reflection of elapsed time. Instead, it is more appropriate to say that timing is controlled *via* the integration of force, in combination with learned yet adjustable thresholds that determine when systems act. Third, the timing of target achievement is not a controlled event. Finally, much of the theoretical vocabulary that spans the range of timescales portrayed in **Figure 2** is contestable, and new interpretations of empirical patterns can be derived from our logic. This applies to units such as syllables and moras, and also to hierarchical and recursive organizations of phrases. Ultimately the logic is useful because it facilitates a unified understanding of temporal patterns in speech, from the short timescale of articulatory timing to the large timescale of variation in speech rate.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/tilsen/TiR-model.git.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Acknowledgments

The author would like to thank members of the Cornell Phonetics Lab for discussion of the ideas in this manuscript.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnhum.2022.851991/full#supplementary-material

References

Beckman, M. E., and Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology* 3, 255–309. doi: 10.1017/S095267570000066X

Bhutani, N., Sureshbabu, R., Farooqui, A. A., Behari, M., Goyal, V., and Murthy, A. (2013). Queuing of concurrent movement plans by basal ganglia. *J. Neurosci.* 33, 9985–9997. doi: 10.1523/JNEUROSCI.4934-12.2013

Bohland, J. W., Bullock, D., and Guenther, F. H. (2010). Neural representations and mechanisms for the performance of simple speech sequences. *J. Cogn. Neurosci.* 22, 1504–1529. doi: 10.1162/jocn.2009.21306

Browman, C., and Goldstein, L. (1988). Some notes on syllable structure in articulatory phonology. *Phonetica* 45, 140–155. doi: 10.1159/000261823

Browman, C., and Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology* 6, 201–251. doi: 10.1017/S0952675700001019

Browman, C., and Goldstein, L. (1990). "Tiers in articulatory phonology, with some implications for casual speech," in *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, eds J. Beckman and M. Kingston (New York, NY: Cambridge University Press), 341–376. doi: 10.1017/CBO9780511627736.019

Browman, C., and Goldstein, L. (1992). Articulatory phonology: an overview. *Phonetica.* 49, 155–180. doi: 10.1159/000261913

Browman, C., and Goldstein, L. (1995). "Gestural syllable position effects in American English," in *Producing Speech: Contemporary Issues*, eds F. Bell-Berti and L. Raphael (Woodbury, NY: American Institute of Physics), 19–33.

Browman, C. P., and Goldstein, L. M. (1986). Towards an Articulatory Phonology. *Phonol. Yearb.* 3, 219–252. doi: 10.1017/S0952675700000658

Bullock, D. (2004). Adaptive neural models of queuing and timing in fluent action. *Trends Cogn. Sci.* 8, 426–433. doi: 10.1016/j.tics.2004.07.003

Bullock, D., and Rhodes, B. (2002). Competitive queuing for planning and serial performance. *CAS CNS Tech. Rep. Ser.* 3, 1–9.

Buzsaki, G. (2006). *Rhythms of the Brain*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780195301069.001.0001

Buzsáki, G., and Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science* 304, 1926–1929. doi: 10.1126/science.1099745

Byrd, D. (1995). C-centers revisited. *Phonetica* 52, 285–306. doi: 10.1159/000262183

Byrd, D. (1996). Influences on articulatory timing in consonant sequences. *J. Phon.* 24, 209–244. doi: 10.1006/jpho.1996.0012

Byrd, D. (2000). Articulatory vowel lengthening and coordination at phrasal junctures. *Phonetica* 57, 3–16. doi: 10.1159/000028456

Byrd, D., Krivokapić, J., and Lee, S. (2006). How far, how long: on the temporal scope of prosodic boundary effects. *J. Acoust. Soc. Am.* 120, 1589–1599. doi: 10.1121/1.2217135

Byrd, D., and Saltzman, E. (1998). Intragestural dynamics of multiple prosodic boundaries. *J.Phon.* 26, 173–199. doi: 10.1006/jpho.1998.0071

Byrd, D., and Saltzman, E. (2003). The elastic phrase: modeling the dynamics of boundary-adjacent lengthening. *J. Phon.* 31, 149–180. doi: 10.1016/S0095-4470(02)00085-2

Cai, S., Ghosh, S. S., Guenther, F. H., and Perkell, J. S. (2011). Focal manipulations of formant trajectories reveal a role of auditory feedback in the online control of both within-syllable and between-syllable speech timing. *J. Neurosci.* 31, 16483–16490. doi: 10.1523/JNEUROSCI.3653-11.2011

Chitoran, I., and Goldstein, L. (2006). "Testing the phonological status of perceptual recoverability: articulatory evidence from Georgian," in *Proceedings of the 10th Conference on Laboratory Phonology, June 29–July 1*, Paris, 69–70.

Eilers, R. E., Oller, D. K., Levine, S., Basinger, D., Lynch, M. P., and Urbano, R. (1993). The role of prematurity and socioeconomic status in the onset of canonical babbling in infants. *Infant Behav. Dev.* 16, 297–315. doi: 10.1016/0163-6383(93)80037-9

Feldman, A. G. (1986). Once more on the equilibrium-point hypothesis (λ model) for motor control. *J. Mot. Behav.* 18, 17–54. doi: 10.1080/00222895.1986.10735369

Feldman, A. G., and Levin, M. F. (2009). The equilibrium-point hypothesis–past, present and future. *Prog. Mot. Control* 629, 699–726. doi: 10.1007/978-0-387-77064-2_38

Féry, C. (2010). Recursion in prosodic structure. *Phonol. Stud.* 13, 51–60.

Fowler, C. A., and Rosenblum, L. D. (1991). "The perception of phonetic gestures," in *Modularity and the Motor Theory of Speech Perception*, eds I. G. Mattingly and M. Studdert-Kennedy (Hillsdale, NJ: Erlbaum), 33–59.

Gafos, A. I. (2002). A grammar of gestural coordination. *Nat. Lang. Linguist. Theory* 20, 269–337. doi: 10.1023/A:1014942312445

Garnier, M., Henrich, N., and Dubois, D. (2010). Influence of sound immersion and communicative interaction on the Lombard effect. *J. Speech Lang. Hear. Res.* 53, 588–608. doi: 10.1044/1092-4388(2009/08-0138)

Gick, B., Campbell, F., Oh, S., and Tamburri-Watt, L. (2006). Toward universals in the gestural organization of syllables: a cross-linguistic study of liquids. *J. Phon.* 34, 49–72. doi: 10.1016/j.wocn.2005.03.005

Glasspool, D. W. (2014). *Competitive Queuing and the Articulatory Loop*. London: Psychology Press.

Goldstein, L., Byrd, D., and Saltzman, E. (2006). "The role of vocal tract gestural action units in understanding the evolution of phonology," in *Action to Language via the Mirror Neuron System*, ed. M. Arbib (Cambridge: Cambridge University Press), 215–249. doi: 10.1017/CBO9780511541599.008

Goldstein, L., Chitoran, I., and Selkirk, E. (2007). "Syllable structure as coupled oscillator modes: evidence from Georgian vs. Tashlhiyt Berber," in *Proceedings of the XVIth International Congress of Phonetic Sciences*, Saarbrücken, 241–244.

Goodell, E. W., and Studdert-Kennedy, M. (1993). Acoustic evidence for the development of gestural coordination in the speech of 2-year-olds: a longitudinal study. *J. Speech Lang. Hear. Res.* 36, 707–727. doi: 10.1044/jshr.3604.707

Grossberg, S. (1987). The adaptive self-organization of serial order in behavior: speech, language, and motor control. *Adv. Psychol.* 43, 313–400. doi: 10.1016/S0166-4115(08)61766-5

Guenther, F. H., and Hickok, G. (2016). "Neural models of motor speech control," in *Neurobiology of Language*, eds G. Hickok and S. L. Small (Amsterdam: Elsevier), 725–740. doi: 10.1016/B978-0-12-407794-2.00058-4

Guenther, F. H., and Perkell, J. S. (2004). "A neural model of speech production and its application to studies of the role of auditory feedback in speech," in *Speech Motor Control in Normal and Disordered Speech*, eds B. Maassen, R. Kent, H. F. M. Peters, P. Van Lieshout, and W. Hulstijn (Oxford: Oxford University Press), 29–49.

Hawkins, S. (1984). "On the development of motor control in speech: evidence from studies of temporal coordination," in *Speech and Language: Advances in Basic Research and Practice*, Vol. 11, ed. N. J. Lass (New York, NY: Academic Press), 317–374. doi: 10.1016/B978-0-12-608611-9.50012-7

Hermes, A., Mücke, D., and Grice, M. (2013). Gestural coordination of Italian word-initial clusters: the case of 'impure s'. *Phonology* 30, 1–25. doi: 10.1017/S095267571300002X

Hermes, A., Ridouane, R., Mücke, D., and Grice, M. (2011). "Kinematics of syllable structure in Tashlhiyt Berber: the case of vocalic and consonantal nuclei," in *Proceedings of the 9th International Seminar on Speech production*, Montreal, QC, 401–408.

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nat. Rev. Neurosci.* 13, 135–145. doi: 10.1038/nrn3158

Honorof, D. N., and Browman, C. (1995). "The center or edge: How are consonant clusters organized with respect to the vowel," in *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Stockholm, 552–555.

Houde, J. F., and Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science* 279, 1213–1216. doi: 10.1126/science.279.5354.1213

Ito, J., and Mester, A. (2013). Prosodic subcategories in Japanese. *Lingua* 124, 20–40. doi: 10.1016/j.lingua.2012.08.016

Iverson, J. M., Hall, A. J., Nickel, L., and Wozniak, R. H. (2007). The relationship between reduplicated babble onset and laterality biases in infant rhythmic arm movements. *Brain Lang.* 101, 198–207. doi: 10.1016/j.bandl.2006.11.004

Jordan, M. I. (1986). *Serial Order: a Parallel Distributed Processing Approach. Technical Report, June 1985-March 1986.* San Diego, CA: Inst. for Cognitive Science.

Karlin, R., Naber, C., and Parrell, B. (2021). Auditory feedback is used for adaptation and compensation in speech timing. *J. Speech Lang. Hear. Res.* 64, 3361–3381. doi: 10.1044/2021_JSLHR-21-00021

Katz, W. F., and Bharadwaj, S. (2001). Coarticulation in fricative-vowel syllables produced by children and adults: a preliminary report. *Clin. Linguist. Phon.* 15, 139–143. doi: 10.3109/02699200109167646

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Curr. Opin. Neurobiol.* 9, 718–727. doi: 10.1016/S0959-4388(99)00028-8

Kawato, M., and Wolpert, D. (1998). Internal models for motor control. *Sens. Guid. Mov.* 218, 291–307. doi: 10.1002/9780470515563.ch16

Kelso, J. A., Saltzman, E. L., and Tuller, B. (1986). The dynamical perspective on speech production: data and theory. *J. Phon.* 14, 29–59. doi: 10.1016/S0095-4470(19)30608-4

Kelso, J. A. S., and Tuller, B. (1987). "Intrinsic time in speech production: theory, methodology, and preliminary observations," in *Motor Sensory and Motor Processes in Language*, Vol. 203, eds E. Keller and M. Gopnik (Hillsdale, NJ: Erlbaum), 222.

Kent, R. (1983). "The segmental organization of speech," in *The Production of Speech*, ed. P. F. MacNeilage (New York, NY: Springer), 57–89. doi: 10.1007/978-1-4613-8202-7_4

Kristan, W. B. (2014). Behavioral sequencing: competitive queuing in the fly CNS. *Curr. Biol.* 24, R743–R746. doi: 10.1016/j.cub.2014.06.071

Krivokapić, J. (2014). Gestural coordination at prosodic boundaries and its role for prosodic structure and speech planning processes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 369:20130397. doi: 10.1098/rstb.2013.0397

Kuhnert, B., Hoole, P., and Mooshammer, C. (2006). "Gestural overlap and C-center in selected French consonant clusters," in *Proceedings of the 7th International Seminar on Speech Production*, eds H. C. Yehia, D. Demolin, and R. Laboissière (Pampulha: CEFALA), 327–334.

Ladd, D. R. (1986). Intonational phrasing: the case for recursive prosodic structure. *Phonology* 3, 311–340. doi: 10.1017/S0952675700000671

Ladefoged, P. (2001). *Vowels and Consonants: An Introduction to the Sounds of the World*. Malden, MA: Blackwell Publications.

Larson, C. R., Burnett, T. A., Bauer, J. J., Kiran, S., and Hain, T. C. (2001). Comparison of voice F0 responses to pitch-shift onset and offset conditions. *J. Acoust. Soc. Am.* 110, 2845–2848. doi: 10.1121/1.1417527

Levelt, W., and Wheeldon, L. (1994). Do speakers have access to a mental syllabary? *Cognition* 50, 239–269. doi: 10.1016/0010-0277(94)90030-2

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychol. Rev.* 74, 431–461. doi: 10.1037/h0020279

Luo, J., Hage, S. R., and Moss, C. F. (2018). The Lombard effect: from acoustics to neural mechanisms. *Trends Neurosci.* 41, 938–949. doi: 10.1016/j.tins.2018.07.011

MacNeilage, P. F., and Davis, B. L. (2000). Deriving speech from nonspeech: a view from ontogeny. *Phonetica* 57, 284–296. doi: 10.1159/000028481

MacNeilage, P. F., Davis, B. L., and Matyear, C. L. (1997). Babbling and first words: phonetic similarities and differences. *Speech Commun.* 22, 269–277. doi: 10.1016/S0167-6393(97)00022-8

Marin, S., and Pouplier, M. (2010). Temporal organization of complex onsets and codas in American English: testing the predictions of a gestural coupling model. *Mot. Control* 14, 380–407. doi: 10.1123/mcj.14.3.380

Max, L., Guenther, F. H., Gracco, V. L., Ghosh, S. S., and Wallace, M. E. (2004). Unstable or insufficiently activated internal models and feedback-biased motor control as sources of dysfluency: a theoretical model of stuttering. *Contemp. Issues Commun. Sci. Disord.* 31, 105–122. doi: 10.1044/cicsd_31_S_105

Miall, R. C., and Wolpert, D. M. (1996). Forward models for physiological motor control. *Neural Netw.* 9, 1265–1279. doi: 10.1016/S0893-6080(96)00035-4

Nam, H. (2007). "Syllable-level intergestural timing model: split-gesture dynamics focusing on positional asymmetry and moraic structure," in *Laboratory Phonology*, eds J. Cole and J. I. Hualde (Berlin, NY: Walter de Gruyter), 483–506.

Oller, D. K. (2000). *The Emergence of the Speech Capacity [Internet]*. Mahwah, NJ: Lawrence Erlbaum. doi: 10.4324/9781410602565

Oschkinat, M., and Hoole, P. (2020). Compensation to real-time temporal auditory feedback perturbation depends on syllable position. *J. Acoust. Soc. Am.* 148, 1478–1495. doi: 10.1121/10.0001765

Parrell, B., Ramanarayanan, V., Nagarajan, S. S., and Houde, J. F. (2018). "FACTS: a hierarchical task-based control model of speech incorporating sensory feedback," in *Proceedings of Interspeech*, 1497–1501, Hyderabad. doi: 10.21437/Interspeech.2018-2087

Parrell, B., Ramanarayanan, V., Nagarajan, S., and Houde, J. (2019). The FACTS model of speech motor control: fusing state estimation and task-based control. *bioRxiv* [Preprint]. doi: 10.1101/543728

Perkell, J. S., and Matthies, M. L. (1992). Temporal measures of anticipatory labial coarticulation for the vowel/u: within-and cross-subject variability. *J. Acoust. Soc. Am.* 91, 2911–2925. doi: 10.1121/1.403778

Port, R. F., and Leary, A. P. (2005). Against formal phonology. *Language* 81, 927–964. doi: 10.1353/lan.2005.0195

Purcell, D. W., and Munhall, K. G. (2006). Compensation following real-time manipulation of formants in isolated vowels. *J. Acoust. Soc. Am.* 119, 2288–2297. doi: 10.1121/1.2173514

Repp, B. H. (1986). Some observations on the development of anticipatory coarticulation. *J. Acoust. Soc. Am.* 79, 1616–1619. doi: 10.1121/1.393298

Saltzman, E., and Munhall, K. (1989). A dynamical approach to gestural patterning in speech production. *Ecol. Psychol.* 1, 333–382. doi: 10.1207/s15326969eco0104_2

Saltzman, E., Nam, H., Krivokapic, J., and Goldstein, L. (2008). "A task-dynamic toolkit for modeling the effects of prosodic structure on articulation," in *Proceedings of the 4th International Conference on Speech Prosody*, Campinas, 175–184.

Schöner, G. (2002). Timing, clocks, and dynamical systems. *Brain Cogn.* 48, 31–51. doi: 10.1006/brcg.2001.1302

Shadmehr, R., and Krakauer, J. W. (2008). A computational neuroanatomy for motor control. *Exp. Brain Res.* 185, 359–381. doi: 10.1007/s00221-008-1280-5

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423,623–656. doi: 10.1002/j.1538-7305.1948.tb00917.x

Shaw, J., Gafos, A. I., Hoole, P., and Zeroual, C. (2009). Syllabification in Moroccan Arabic: evidence from patterns of temporal stability in articulation. *Phonology* 26, 187–215. doi: 10.1017/S0952675709001754

Shaw, J., Gafos, A. I., Hoole, P., and Zeroual, C. (2011). Dynamic invariance in the phonetic expression of syllable structure: a case study of Moroccan Arabic consonant clusters. *Phonology* 28, 455–490. doi: 10.1017/S0952675711000224

Sorensen, T., and Gafos, A. (2016). The gesture as an autonomous nonlinear dynamical system. *Ecol. Psychol.* 28, 188–215. doi: 10.1080/10407413.2016.1230368

Sproat, R., and Fujimura, O. (1993). Allophonic variation in English/l/and its implications for phonetic implementation. *J. Phon.* 21, 291–311. doi: 10.1016/S0095-4470(19)31340-3

Sussman, H. M., Duder, C., Dalston, E., and Cacciatore, A. (1999). An acoustic analysis of the development of CV coarticulation: a case study. *J. Speech Lang. Hear. Res.* 42, 1080–1096. doi: 10.1044/jslhr.4205.1080

Thelen, E. (1979). Rhythmical stereotypies in normal human infants. *Anim. Behav.* 27, 699–715. doi: 10.1016/0003-3472(79)90006-X

Thoroughman, K. A., and Shadmehr, R. (1999). Electromyographic correlates of learning an internal model of reaching movements. *J. Neurosci.* 19, 8573–8588. doi: 10.1523/JNEUROSCI.19-19-08573.1999

Tilsen, S. (2011). Effects of syllable stress on articulatory planning observed in a stop-signal experiment. *J. Phon.* 39, 642–659. doi: 10.1016/j.wocn.2011.04.002

Tilsen, S. (2014). *Selection-Coordination Theory. Cornell Working Papers in Phonetics and Phonology, 2014*, 24–72. City: Ithaca, NY.

Tilsen, S. (2016). Selection and coordination: the articulatory basis for the emergence of phonological structure. *J. Phon.* 55, 53–77. doi: 10.1016/j.wocn.2015.11.005

Tilsen, S. (2017). Exertive modulation of speech and articulatory phasing. *J. Phon.* 64, 34–50. doi: 10.1016/j.wocn.2017.03.001

Tilsen, S. (2018). *Three Mechanisms for Modeling Articulation: Selection, Coordination, and Intention. (Cornell Working Papers in Phonetics and Phonology 2018)*. city: Ithaca, NY.

Tilsen, S. (2019a). Motoric mechanisms for the emergence of non-local phonological patterns. *Front. Psychol.* 10:2143. doi: 10.3389/fpsyg.2019.02143

Tilsen, S. (2019b). Space and time in models of speech rhythm. *Ann. N. Y. Acad. Sci.* 1453, 47–66. doi: 10.1111/nyas.14102

Tilsen, S. (2019c). *Syntax with Oscillators and Energy Levels (Studies in Laboratory Phonology)*. Berlin: Language Science Press.

Tilsen, S., and Hermes, A. (2020). "Nonlinear effects of speech rate on articulatory timing in singletons and geminates," in *Proceedings of the 12th International Seminar on Speech Production*, New Haven, CT.

Tilsen, S., Zec, D., Bjorndahl, C., Butler, B., L'Esperance, M. J., Fisher, A., et al. (2012). A cross-linguistic investigation of articulatory coordination in word-initial consonant clusters. *Cornell Work. Pap. Phon. Phonol.* 2012, 51–81.

Tilsen, S. A. (2013). Dynamical model of hierarchical selection and coordination in speech planning. *PLoS One* 8:e62800. doi: 10.1371/journal.pone.0062800

Tourville, J. A., and Guenther, F. H. (2011). The DIVA model: a neural theory of speech acquisition and production. *Lang. Cogn. Process.* 26, 952–981. doi: 10.1080/01690960903498424

Tourville, J. A., Reilly, K. J., and Guenther, F. H. (2008). Neural mechanisms underlying auditory feedback control of speech. *Neuroimage* 39, 1429–1443. doi: 10.1016/j.neuroimage.2007.09.054

Turk, A., and Shattuck-Hufnagel, S. (2014). Timing in talking: what is it used for, and how is it controlled? *Philos. Trans. R. Soc. B Biol. Sci.* 369:20130395. doi: 10.1098/rstb.2013.0395

Turk, A., and Shattuck-Hufnagel, S. (2020a). *Speech Timing: Implications for Theories of Phonology, Speech Production, and Speech Motor Control*, Vol. 5. New York, NY: Oxford University Press. doi: 10.1093/oso/9780198795421.001.0001

Turk, A., and Shattuck-Hufnagel, S. (2020b). Timing evidence for symbolic phonological representations and phonology-extrinsic timing in speech production. *Front. Psychol.* 10:2952. doi: 10.3389/fpsyg.2019.02952

Turk, A. E., and Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. *J. Phon.* 35, 445–472. doi: 10.1016/j.wocn.2006.12.001

Villacorta, V. M., Perkell, J. S., and Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *J. Acoust. Soc. Am.* 122, 2306–2319. doi: 10.1121/1.2773966

Vorberg, D., and Wing, A. (1996). Modeling variability and dependence in timing. *Handb. Percept. Action* 2, 181–262. doi: 10.1016/S1874-5822(06)80007-1

Wing, A. M., and Kristofferson, A. B. (1973). Response delays and the timing of discrete motor responses. *Percept. Psychophys.* 14, 5–12. doi: 10.3758/BF03198607

Zollinger, S. A., and Brumm, H. (2011). The Lombard effect. *Curr. Biol.* 21, R614–R615. doi: 10.1016/j.cub.2011.06.003

frontiers | Frontiers in Human Neuroscience

TYPE Original Research
PUBLISHED 25 August 2022
DOI 10.3389/fnhum.2022.905365

Check for updates

OPEN ACCESS

EDITED BY
Lucie Menard,
Université du Québec à Montréal, Canada

REVIEWED BY
Pascal Perrier,
UMR 5216 Grenoble Images Parole Signal Automatique (GIPSA-Lab), France
Takayuki Ito,
UMR 5216 Grenoble Images Parole Signal Automatique (GIPSA-Lab), France

*CORRESPONDENCE
Daniel R. Nault
daniel.nault@queensu.ca

SPECIALTY SECTION
This article was submitted to Speech and Language, a section of the journal Frontiers in Human Neuroscience

RECEIVED 27 March 2022
ACCEPTED 04 August 2022
PUBLISHED 25 August 2022

CITATION
Nault DR, Mitsuya T, Purcell DW and Munhall KG (2022) Perturbing the consistency of auditory feedback in speech.
Front. Hum. Neurosci. 16:905365.
doi: 10.3389/fnhum.2022.905365

COPYRIGHT
© 2022 Nault, Mitsuya, Purcell and Munhall. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Perturbing the consistency of auditory feedback in speech

Daniel R. Nault[1]*, Takashi Mitsuya[2,3], David W. Purcell[2,3] and Kevin G. Munhall[1]

[1]Department of Psychology, Queen's University, Kingston, ON, Canada, [2]School of Communication Sciences and Disorders, Western University, London, ON, Canada, [3]National Centre for Audiology, Western University, London, ON, Canada

Sensory information, including auditory feedback, is used by talkers to maintain fluent speech articulation. Current models of speech motor control posit that speakers continually adjust their motor commands based on discrepancies between the sensory predictions made by a forward model and the sensory consequences of their speech movements. Here, in two within-subject design experiments, we used a real-time formant manipulation system to explore how reliant speech articulation is on the accuracy or predictability of auditory feedback information. This involved introducing random formant perturbations during vowel production that varied systematically in their spatial location in formant space (Experiment 1) and temporal consistency (Experiment 2). Our results indicate that, on average, speakers' responses to auditory feedback manipulations varied based on the relevance and degree of the error that was introduced in the various feedback conditions. In Experiment 1, speakers' average production was not reliably influenced by random perturbations that were introduced every utterance to the first (F1) and second (F2) formants in various locations of formant space that had an overall average of 0 Hz. However, when perturbations were applied that had a mean of +100 Hz in F1 and −125 Hz in F2, speakers demonstrated reliable compensatory responses that reflected the average magnitude of the applied perturbations. In Experiment 2, speakers did not significantly compensate for perturbations of varying magnitudes that were held constant for one and three trials at a time. Speakers' average productions did, however, significantly deviate from a control condition when perturbations were held constant for six trials. Within the context of these conditions, our findings provide evidence that the control of speech movements is, at least in part, dependent upon the reliability and stability of the sensory information that it receives over time.

KEYWORDS

speech motor control, speech production, auditory feedback, perturbation, consistency, variability, compensation

Introduction

Painted on the window of a café in the Norrmalm district of Stockholm is information to help customers find their way in. Within an arrow pointing to the left is the text, "Entrance 8,47 M". What makes this signage funny is its precision. Knowing the door's location to the hundredth of a meter when you are steps away from entering is excessive and it makes passersby smile when they see it. People have an intuitive feel for what information they need and how precise it should be.

Current models of the control of actions include sensory information that is used to coordinate the movements accurately or is needed to maintain the stability of the motor system [see Parrell et al. (2019a) for a review of recent speech models]. Such models include closed-loop processing of sensory information to guide immediate motor responses and predictive algorithms where sensory information is used to tune representations of the effectors and their activities. In both types of sensorimotor control, the required precision of the sensory information and reliability of that information is a part of the control system.

The present paper addresses this issue of the precision of perceptual information for action in a specific context—spoken language. All the papers in this special issue present studies of how the auditory feedback for speech is processed and how it influences the accuracy of talking. The technique that is employed in these papers is the real-time modification of the sounds that talkers produce so that they hear themselves say sounds slightly differently than they actually spoke. Studies have shown that introducing errors in the timing (Mitsuya et al., 2014), amplitude (Heinks-Maldonado and Houde, 2005), pitch (Kawahara, 1995), and spectral details (Houde and Jordan, 1998) of the auditory feedback cause talkers to modify their speech in compensation. The question we are asking here is: How "off" can the feedback be?

The best answer to that question is: it depends. It depends on the vocal parameter. Timing, amplitude, and frequency parameters may be related in spoken language, but they are the purview of different articulatory subsystems, and they convey different communicative information in speech. They are measured in different physical qualities with different units. Thus, there is no simple one-to-one correspondence between their signal ranges or their variabilities.

Here we report studies of variability in speech produced in a very restricted context. Specifically, we present a series of studies of vowel formant feedback produced in repetitive citation format. This choice is determined by factors both pragmatic and strategic. Practically, the custom real-time processor that we use (Purcell and Munhall, 2006) is designed for cued production of a stimulus set where real-time formant tracking is optimized for a particular vowel. Repetitive productions of the same syllable are ideal for this paradigm.

Our strategic reason for using repetitive syllable production is that we aim to understand the operating principles of the most basic speech utterances spoken at a normal rate. By using feedback perturbations on a syllabic unit, we are trying to carry out system identification for speech motor behavior. With controlled conditions, and the subject performing the same task (e.g., moving to the same target), the character of the dynamic system that controls articulation can be uncovered[1]. This is an admittedly reductionist approach, but we believe it serves as important baseline behavior of the much more complex system.

Our focus here will be on trial-to-trial variability within and between subjects. Variability is one of the hallmarks of speech and motor systems generally, and it can be the result of 'noise' at many levels in the nervous system (Faisal et al., 2008: cellular, synaptic, sensory, motor, etc.). Such noise can be seen as a challenge for control but is also thought to be beneficial in some circumstances (e.g., in learning and skill acquisition: Dhawale et al., 2017; Sternad, 2018). Here we treat it as a biomarker of the state of the system (Riley and Turvey, 2002) as we assess changes in the predictability of auditory feedback in speech.

Vowel production in both acoustic and articulatory terms shows considerable variability (e.g., Whalen et al., 2018) but variability that is consistent across vowels and correlated for acoustics and articulation. While this variability can change over the course of a day, it is relatively stable across days (Heald and Nusbaum, 2015). Because of these attributes, changes in variability are frequently used as an index of developmental stage (Sosa, 2015) and clinical status (Miller, 1992). We will use this parameter as an index of how the speech system responds to changes in the predictability of auditory feedback.

Studying the predictability of auditory feedback has several important advantages. Experimentally, it is something that can be manipulated in the real-time feedback paradigm. Critically, it is also at the heart of most current computational models of speech, including DIVA (Villacorta et al., 2007), GEPPETO (Patri et al., 2018, 2019), and FACTS (Parrell et al., 2019b). Forward models are proposed to predict the sensory consequences of speaking and adjust future motor commands to the computed discrepancies between model and sensory feedback. Sensorimotor speech control is thought to be inherently predictive.

The present studies

Below in **Figure 1** is a modified version of the production half of Denes and Pinson's (1973) speech chain. The figure portrays a closed loop between intention and the feedback that talkers hear of their own speech. The red arrow indicates our experimental intervention. Our proposal is that, if subjects are

1 One complication in our approach is that speech targets are an unknown quantity. Unlike eye-hand coordination, where targets can be experimentally defined and error be measured from a physical target location, speech targets can only be experimentally defined by a linguistic category. Subjects are instructed to say a word or syllable and they select their target. The target and target width can only be inferred from repetitive utterances produced under the same conditions.

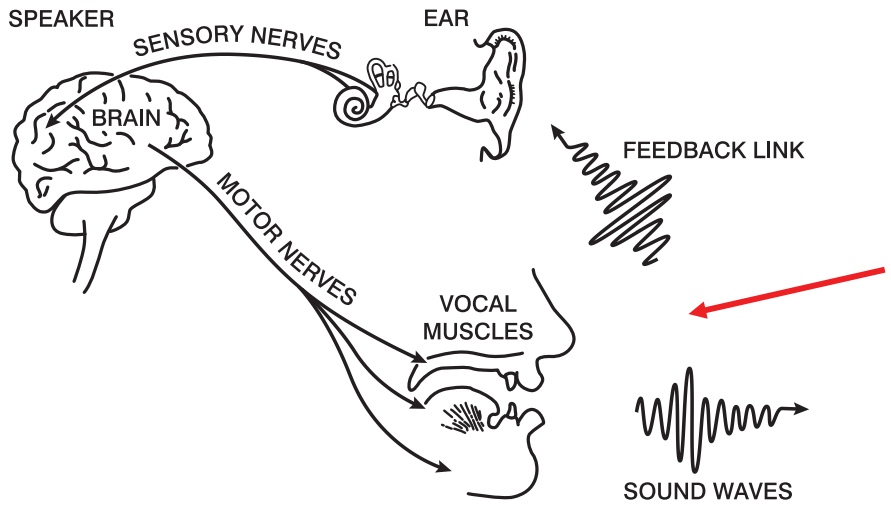


FIGURE 1
The Speech Chain (Denes and Pinson, 1973).

producing naturally-paced syllables[2], we can manipulate the regularity of the auditory feedback and determine the assiduity of the forward model.

Prior to considering our manipulations, it is useful to reflect on what is known of the boundary conditions of the auditory feedback system. For both temporal and spectral perturbations, there are demonstrated ranges over which subjects respond to change. In formant perturbation studies that increase or ramp the changes by small amounts on successive utterances, subjects do not produce compensations on average until the perturbation is beyond a threshold (Purcell and Munhall, 2006). It is as if there is a tolerance for variation in production and small errors do not require correction. At the other end of the perturbation range, subjects' compensations increase linearly with steps in the ramp until the perturbations become too large (MacDonald et al., 2010). Compensations in both the first and second formant reach an asymptote and, as perturbations in the experiment continue to increase with each utterance, the compensation starts to decrease. Finally, the auditory feedback system operates optimally with simultaneous feedback and less so with delays (Mitsuya et al., 2017). Mitsuya et al. (2017) showed that with delays decreasing from 100 ms, the compensation grew linearly to simultaneity. This presents a picture of a formant control system that has inherent variability and that operates within a bounded set of conditions. It does not correct changes smaller than a range of about plus or minus 50 Hz. The control system does not make changes specifically tied to large perturbations of more than 250 Hz and compensates most strongly when there are no delays in auditory feedback. Sudden step changes in formant frequency within this span of conditions are compensated over a series of utterances (approximately 10 trials) rather than on the next trial.

Here, we aim to explore, within the scope of these conditions, how reliant speech articulation is on a predictable auditory feedback environment over a sequence of utterances. In the study of visuomotor and force field paradigms for limb movement, manipulations of feedback predictability have advanced Bayesian perspectives on motor adaptation and sensorimotor control (see Krakauer et al., 2019 for a review). The extension of this approach to speech production has been limited. Daliri and Dittman (2019) have addressed these issues in a series of papers. Their work suggests that task relevance and the magnitude of the error influence the magnitude of the observed compensation. Here, we extend this work by applying manipulations to the probability of perturbation and the consistency or range of the errors that speakers hear.

The data presented in this paper stem from two separate experiments, each involving multiple conditions. Experiment 1 was conducted at the University of Western Ontario, while Experiment 2 was conducted at Queen's University. The raw data are publicly available on OSF here: osf.io/n4pgf.

Experiment 1

In Experiment 1, we directly manipulated the predictability and, therefore, the variability of the auditory feedback of

2 There is a strategy in several laboratories to instruct subjects to prolong their syllables to study rapid, closed-loop control of speech. While this strategy has been used to increase experimental efficiency and to enable the study of both rapid closed-loop and feed-forward control in the same trials, we are concerned that the prolonged utterances are a different phenomenon than naturally-paced productions. Indeed, it has been suggested that online responses to feedback perturbations and the between-trial effects that we are studying are controlled by different neural mechanisms (Raharjo et al., 2021).

speakers' formant frequencies during vowel production. In three experimental conditions, we constrained the auditory feedback speakers received into specific regions of the F1/F2 vowel space. Our aim was to examine the influence of systematic variability in auditory feedback on speakers' moment-to-moment and average speech production patterns.

We selected three different types of feedback variability that varied in the range of the feedback error introduced, and in the degree of independence of the perturbations to F1 and F2:

1. Randomly and independently perturbing F1 and F2 on each trial over the frequency range that would change the syllable "head" to either "hid" or "had" (F1 ±200 Hz; F2 ±250 Hz) but with an overall mean perturbation of 0 Hz in both formants.
2. The same random perturbations over the same frequency range but only for F1. No perturbation was applied to F2. As in the first condition, the overall mean perturbation was 0 Hz.
3. A more phonetic perturbation that randomly varied the feedback for F1 and F2 on each trial in a coupled manner as if the feedback was being shifted between "head" and "had." This varied the vowel quality within a small region of the vowel space and smaller region of the acoustic space (F1 +200 Hz; F2 −250 Hz). We used this condition to also test whether introducing a bias to the randomization would influence the behavior of the speech motor system. In this condition, the mean perturbation across trials was F1 = 100 Hz and F2 = −125 Hz.

These feedback perturbations are only a subset of the ways that unpredictability could alter feedback processing in fluent speech. However, they sample distinct modes of noise in speech feedback and will serve to test in a broad way the dependence on similar noise levels in F1 and F2. They also provide an initial test of the effects of the range of perturbation variability.

Materials and methods

Participants

Eighteen female speakers fluent in Canadian English ranging in age from 21 to 30 years of age (M_{age} = 24.06, SD_{age} = 2.26) participated in the study. Eight speakers reported being fluent in at least one other language in addition to English. To reduce variability in formant values due to sex differences, only female participants were recruited. All participants had normal audiometric hearing thresholds between 500 and 4,000 Hz (≤20 dB hearing level) and reported having no speech or language impairments. All participants provided written, informed consent prior to participating and all experimental procedures were approved by the Health Sciences Research Ethics Board at Western University.

Equipment

The equipment used for Experiment 1 was the same as previously reported in Mitsuya et al. (2017). Participants sat in front of a computer monitor in a sound-attenuated booth (Eckel Industries of Canada, model C2) and wore headphones (Sennheiser HD 265). Their speech was recorded using a portable headset microphone (Shure WH20). The microphone signal was amplified (Tucker-Davis Technologies MA3 microphone amplifier), low-pass filtered with a cut-off frequency of 4,500 Hz (Frequency Devices type 901) and digitized at a sampling rate of 10 kHz. The signal was then filtered in real-time to produce formant feedback perturbations (National Instruments PXI-8106 embedded controller). The processed speech signal was presented back to participants with Sennheiser HD 265 headphones at approximately 80 dBA sound pressure level (SPL) with speech shaped noise (Madsen Itera) of 50 dBA SPL.

Acoustic processing

Voicing was detected using a statistical amplitude threshold, and formant manipulations were introduced in real time using an infinite impulse response filter (see Purcell and Munhall, 2006). An iterative Burg algorithm (Orfanidis, 1988) was implemented to estimate formant changes every 900 μs. Formant estimates were then used to calculate filter coefficients. A pair of spectral zeros were used to deemphasize energy present in the existing formant frequency, and a pair of spectral poles were used to emphasize energy present in the new desired formant.

Prior to data collection, talkers were cued to randomly produce six tokens of each English vowel in the /hVd/ context ("heed," "hid," "hayed," "head," "had," "hawed," "hoed," "who'd," "hood," and "heard"). This was carried out to estimate a parameter that determined the number of coefficients used in the real-time filtering of the vowels in the experiment. Participants were presented with a visual prompt of each word that remained on a computer screen for 2.5 s (with an inter-stimulus interval of approximately 1.5 s).

Formants were analyzed offline in the same manner as previously reported in Munhall et al. (2009). For each utterance, vowel boundaries of the vowel segment were estimated using an automated process based on the harmonicity of the power spectrum. Vowel boundaries were then inspected by hand and corrected, if necessary. Trials were occasionally removed from the dataset when participants made an error (i.e., pronounced the wrong word, failed to produce the correct vowel, coughed or lip smacked during production). The same algorithm that was used for real-time formant

10.3389/fnhum.2022.905365

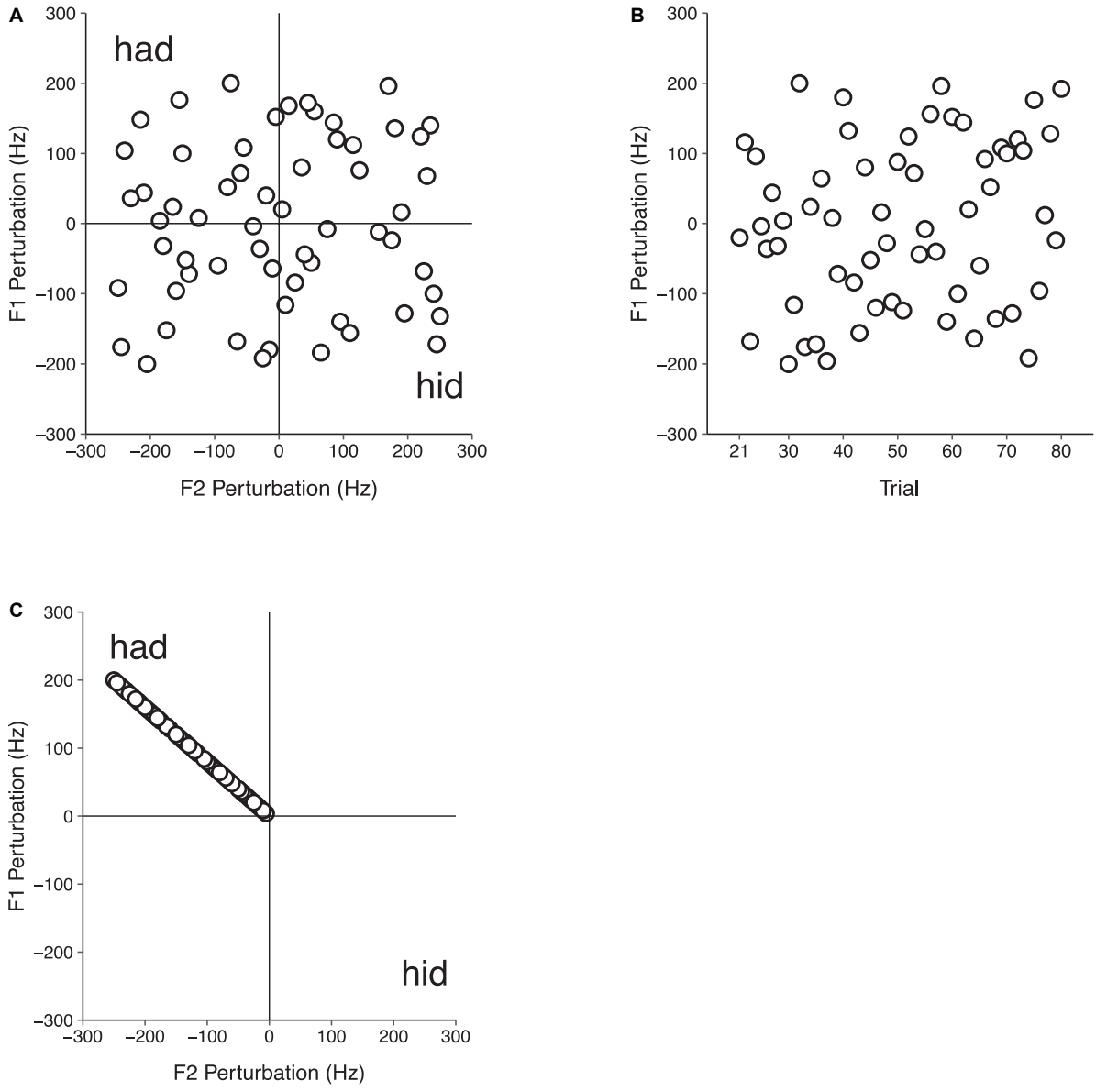


FIGURE 2
Auditory feedback perturbation values in the Perturbation phase (trials 21–80) of Experiment 1 in the F1/F2 Random Perturbation Condition **(A)**, F1-Only Random Perturbation Condition **(B)**, and F1/F2 Coupled Random Perturbation Condition **(C)**. In the F1/F2 Random Perturbation Condition, perturbation magnitudes were not related. Half of the perturbations were positive, and half were negative. The overall average perturbation value in F1 and F2 was 0 Hz. In the F1-Only Random Perturbation Condition, only F1 was perturbed. An equal number of random positive and negative F1 perturbations were applied, and the overall average perturbation value in F1 was 0 Hz. In the F1/F2 Coupled Random Perturbation Condition, speakers received feedback that was biased toward the vowel /æ/ in "had" in F1/F2 space. Random perturbation magnitudes in F1 and F2 were related, with F1 and F2 perturbations being applied in multiples of 4 and −5 Hz, respectively.

tracking was also used offline to estimate the first three formant frequencies (F1, F2, and F3) for each utterance. Formants were estimated from the middle 40–80% of each vowel's duration. On the occasion when a formant was incorrectly categorized as another (e.g., F1 was categorized as F2), it was manually corrected by inspecting the utterance with all the "steady state" F1, F2, and F3 estimates marked for that participant.

Design and procedure

Prior to the experiment, participants filled out a questionnaire to indicate their native language and current language(s) spoken, and to screen for any known vision, hearing, speech, and language impairments. Each participant also completed a hearing screening test at octave frequencies of 500, 1,000, 2,000, and 4,000 Hz prior to beginning the speech experiment.

Participants sat in a sound-attenuated booth in front of a computer monitor and said the monosyllabic word "head" 140 times in each of three conditions: (1) F1/F2 Random Perturbation Condition, (2) F1-Only Random Perturbation Condition, and (3) F1/F2 Coupled Random Perturbation Condition. In each condition, three successive experimental phases that were not indicated to participants were tested. In the Baseline phase (trials 1–20), participants spoke while receiving natural, unaltered auditory feedback. In the Perturbation phase (trials 21–80), participants' auditory feedback was manipulated. In the F1/F2 Random Perturbation Condition, this involved randomly perturbing F1 and F2 in multiples of 4 and 5 Hz, respectively, on each trial. The magnitude of the perturbations in F1 and F2 were not related. However, the directions of the perturbations in F1 and F2 were equally distributed. A quarter (i.e., 15) of the perturbations in F1/F2 were both positive (F1 + F2+), both negative (F1 − F2−), and one positive and one negative (F1 + F2−; F1 − F2 +). The overall average perturbation magnitude during the Perturbation phase of the F1/F2 Random Perturbation Condition was 0 Hz in F1 and F2 (see **Figure 2A**). In the F1-Only Random Perturbation Condition, perturbations were applied in the same way, but only in F1. As in the F1/F2 Random Perturbation Condition, an equal number of positive and negative F1 perturbations were applied during the Perturbation phase (see **Figure 2B**). In the F1/F2 Coupled Random Perturbation Condition, speakers were presented with perturbations that biased the auditory feedback they received from the vowel /ɛ/ in "head" toward the vowel /æ/ in "had" in F1/F2 space (see **Figure 2C**). This was achieved by randomly applying positive F1 perturbations in multiples of 4 Hz ranging from +4 to +200 Hz. The average F1 perturbation value was +100 Hz. Perturbation values in F2 were negative and were determined by dividing the value of the F1 perturbation by four and multiplying by negative five. All subjects received the same randomization of perturbations in each condition. The final Return phase (trials 81–140) was the same in all three conditions; participants' natural unaltered auditory feedback was restored.

The order of each condition was counterbalanced across participants. Before the experiment began, the experimenter instructed participants to speak in their normal conversational voice, and to keep the loudness and pitch of their voice as stable as possible throughout the experiment. To ensure participants returned to baseline speech production after each condition, the experimenter entered the sound booth, and engaged in a short conversation with the participant for a few minutes.

Data analysis

The procedure for data analysis involved first eliminating trials 1–5 from the dataset to minimize the impact of subjects' familiarization with the speech task and with speaking while receiving feedback through headphones. Each speaker's utterances were then normalized for each condition by subtracting that speaker's mean Baseline formant frequencies from each of their utterances. This procedure facilitated our ability to compare formant frequencies across speakers. Speakers' normalized F1 and F2 values were used as the dependent variable in all reported analyses. Descriptive statistics of raw formant values are provided in the **Supplementary material**.

In both experiments, linear mixed-effects modeling (LMM) was used to examine the influence of condition and phase on speakers' normalized speech production. Modeling was carried out using the lme4 package (v1.1-27; Bates et al., 2015) in R (R Core Team, 2020). Analyzing our data in this way allowed for the simple handling of missing data. It also allowed us to maximize our control over unexplained variance in formant frequencies among individual speakers by including a random-effects term. For each experiment, two linear mixed-effects models were constructed—one for F1 and one for F2. As per the guidelines set forth by Barr et al. (2013), the random effects structure for each model was kept as maximal as possible based on our experimental design and the satisfaction of model convergence criteria. In each model, this involved including a random intercept for speakers causing non-independence in the data and, if possible, a random slope for each within-unit predictor if there were no convergence errors. If convergence criteria were not satisfied, the random effects structure was simplified by removing the random slope that explained the smallest amount of variance. This process was continued until the random effects model converged (Barr et al., 2013). The random effects structure for each model was determined prior to adding any fixed effects.

In each LMM analysis, we refer to the model with the best fit to the data as the Best Fit Model. In all cases, Best Fit Models were determined using a "backward-fitting" model selection approach (Bates et al., 2015). This involved first testing a model with the maximal random effects structure that satisfied convergence criteria and all fixed effects of interest (i.e., condition, phase, and their interaction term). Fixed effects were then removed one at a time and alternative models were compared for goodness of fit to the data using likelihood ratio tests (LRTs). Two-tailed p-values and confidence intervals were estimated using a Wald t-distribution with Satterthwaite approximation. The Best Fit Model for each analysis always significantly outperformed all other testable models and satisfied convergence criteria. In cases where significant fixed effects were observed, the emmeans package (v.1.7.0; Lenth, 2019) was used to conduct pairwise comparisons with the Bonferroni correction. In secondary analyses, within-subjects ANOVAs (one for F1, one for F2) were used to examine whether average within-speaker variability (i.e., standard deviation) differed by condition and phase.

We also investigated the possibility of oscillations in compensation throughout the Perturbation phase of each condition. This was achieved by computing an amplitude

spectrum for each subject in each condition using the normalized F1 values from the Perturbation phase as time series. The spectra were calculated in MATLAB (2020b) using a discrete Fourier transform with a Hanning window and a sampling rate of one sample per trial. The resulting amplitude spectra had normalized units of frequency (normalized by the sampling rate and reported as cycles per trial) and were averaged across subjects for each condition. If there was a prominent oscillation of F1 values across trials in the Perturbation phase of any condition, it would be expected to appear as a peak in the frequency spectrum. By averaging only the amplitude spectra, between-subject variability in the temporal position of cycles of a potential oscillation across trials in the Perturbation phase will not diminish detection of the oscillation in the average spectrum.

Results

The primary dataset for Experiment 1 involved a total of 7,290 utterances (18 speakers * 3 conditions * 135 trials = 7,290). Thirty F1 values and 43 F2 values were omitted from the dataset due to issues with formant tracking. The reported results involve normalized formant frequencies. We begin by visually presenting the average normalized results for F1 and F2 in each condition. We then report the results from the Best-Fit Models used to predict normalized speech production in F1 and F2, followed by analyses of average within-speaker variability.

The average normalized results for F1 and F2 across all three phases of each condition in Experiment 1 are shown in **Figure 3**. The general pattern apparent in **Figure 3** is that the random perturbations with a mean of zero relative formant frequency in the F1/F2 Random Perturbation Condition and F1-Only Random Perturbation Condition had minimal effects on average formant production. In contrast, when the random perturbations had a mean of F1 = 100 Hz and F2 = −125 Hz in the F1/F2 Coupled Random Perturbation Condition, the average compensations resembled those produced in experiments with a step perturbation (e.g., Munhall et al., 2009; MacDonald et al., 2011).

In the LMM analysis of speakers' normalized F1 speech production values, the Best-Fit Model produced a significantly better fit to the data than a null model that only included the random effects, $\chi^2(8) = 505.67$, $p < 0.001$. It also significantly outperformed alternative models that only included the fixed effect of Phase [$\chi^2(6) = 301.12$, $p < 0.001$] or Condition, $\chi^2(6) = 492.82$, $p < 0.001$. The Best Fit Model was a significantly better fit to the data than another alternative model that did not include the interaction between Condition and Phase, $\chi^2(4) = 288.24$, $p < 0.001$.

Results from the Best-Fit Model revealed a significant Phase effect. Pairwise comparisons using the Bonferroni correction revealed that speakers' normalized F1 values were significantly more negative during the Perturbation ($M = -14.02$, $SE = 2.95$) and Return ($M = -7.64$, $SE = 2.95$) phases than they were during the Baseline phase ($M = -0.04$, $SE = 3.06$), all ps < 0.001. The main effect of Condition was not significant. However, there was a significant interaction between Condition and Phase. Adjusting for multiple comparisons, pairwise tests showed that there were significant mean differences between the F1 values produced by speakers during the Perturbation phase of the F1/F2 Coupled Random Perturbation Condition ($M = -32.47$, $SE = 3.64$), and the Perturbation phases of the F1/F2 Random Perturbation Condition ($M = -4.80$, $SE = 4.71$) and F1-Only Random Perturbation Condition ($M = -4.79$, $SE = 3.78$), both ps < 0.001. In the F1-Only Random Perturbation Condition, speakers' F1 values were significantly more negative during the Return phase ($M = -8.88$, $SE = 3.78$) than during the Baseline phase ($M = -0.009$, $SE = 4.03$), $p < 0.001$. In the F1/F2 Coupled Random Perturbation Condition, speakers' F1 values were also significantly more negative during the Return phase ($M = -13.43$, $SE = 3.64$) than during the Baseline phase ($M = 0.03$, $SE = 3.89$), $p < 0.001$. Thus, on average, speakers did not reliably compensate for random F1 perturbations that had a relative overall average of 0 Hz. However, when random F1 perturbations had an average that deviated from zero, speakers demonstrated significant compensatory behavior. In two conditions, speakers' average F1 production also remained significantly negative as compared to the Baseline phase following the restoration of their natural auditory feedback during the Return phase. A full list of pairwise comparisons and their significance values are provided in the **Supplementary material**. Best-Fit Model coefficients are shown in **Table 1**.

The Best-Fit Model predicting speakers' normalized F2 production was a significantly better fit to the data than a null model that only had the random effects, $\chi^2(8) = 373.2$, $p < 0.001$. An alternative model that did not have the Condition effect failed to converge. The Best-Fit Model significantly outperformed alternative models that did not have the Phase effect [$\chi^2(6) = 354.59$, $p < 0.001$], or the interaction between Condition and Phase [$\chi^2(4) = 208.91$, $p < 0.001$], both ps < 0.001.

Results from the Best-Fit Model in F2 revealed that the main effects of Condition and Phase were not significant. However, there was a significant interaction between Condition and Phase. Pairwise comparisons using the Bonferroni correction revealed that, on average, speakers' F2 production was significantly more positive during the Perturbation phase of the F1/F2 Coupled Random Perturbation Condition ($M = 38.26$, $SE = 6.99$) than during the Perturbation phases of the F1/F2 Random Perturbation Condition ($M = 3.07$, $SE = 7.51$) and the F1-Only Random Perturbation Condition ($M = -14.86$, $SE = 5.05$), both $ps < 0.001$. In the F1-Only Random Perturbation Condition, there were significant mean differences between speakers' F2 values produced during the Baseline phase ($M = -0.20$,

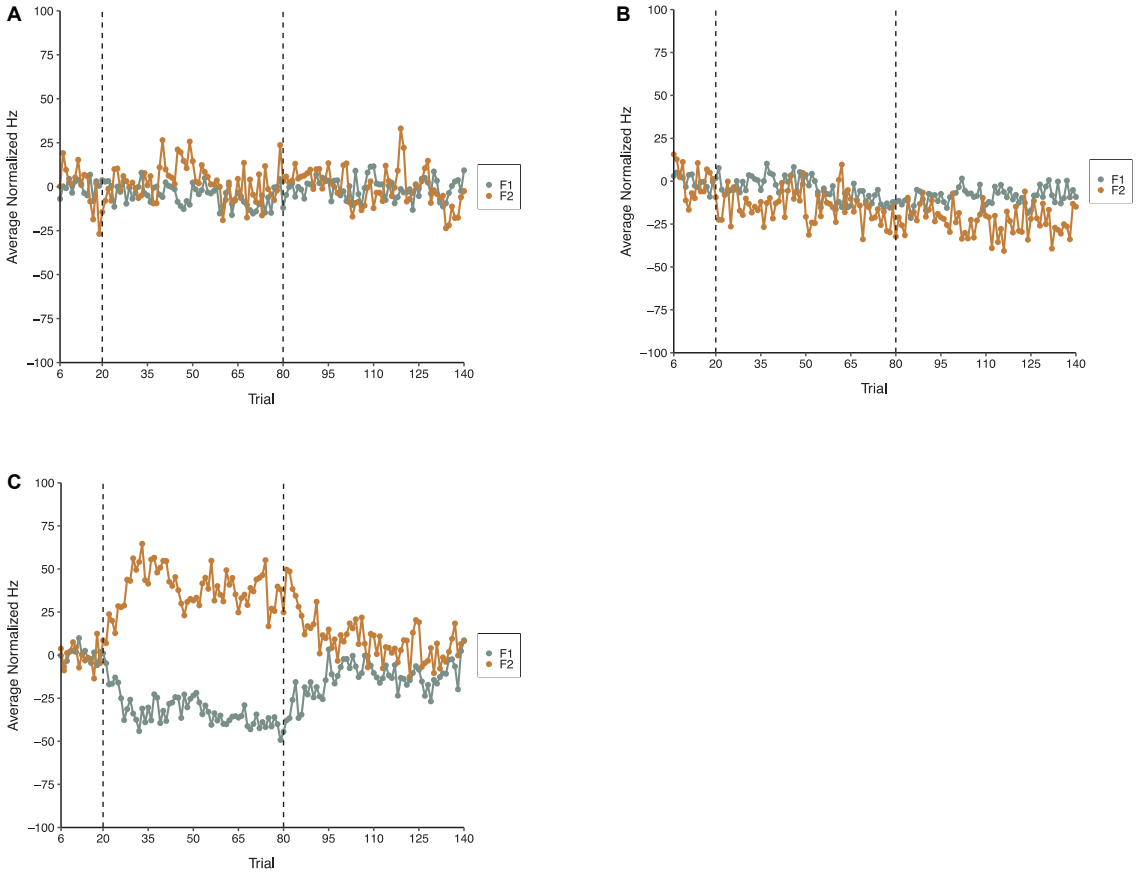


FIGURE 3
Average normalized F1 (gray) and F2 (gold) speech production values from 18 speakers in the F1/F2 Random Perturbation Condition **(A)**, F1-Only Random Perturbation Condition **(B)**, and F1/F2 Coupled Random Perturbation Condition **(C)** of Experiment 1. From left to right, the dotted lines denote boundaries between the Baseline, Perturbation, and Return phases, respectively.

TABLE 1 Coefficients from the Best-Fit Model used to predict speakers' normalized F1 values during Experiment 1.

Fixed effects	Estimate (*SE*)	95% CI	*t*-value	*P*-value	Random effects	*SD*
Intercept (F1/F2 condition baseline)	−0.15 (4.92)	[−10.39, 10.08]	−0.03	0.975	*Speaker*	
F1-only condition	0.14 (5.16)	[−10.49, 10.78]	0.03	0.978	Intercept (F1/F2 random condition)	19.71
Linear condition	0.18 (4.80)	[−9.68, 10.05]	0.04	0.970	F1-only condition	19.66
Perturbation phase	**−4.65 (1.80)**	**[−8.18,−1.11]**	**−2.58**	**0.010**	Linear condition	17.94
Return phase	−0.46 (1.80)	[−4.00, 3.07]	−0.26	0.797	*Residual*	26.27
F1-only*perturbation	−0.14 (2.54)	[−5.13, 4.85]	−0.05	0.956		
Linear*perturbation	**−27.85 (2.54)**	**[−32.84,−22.87]**	**−10.95**	**<0.001**		
F1-only*return	**−8.41 (2.54)**	**[−13.39,−3.42]**	**−3.30**	**<0.001**		
Linear*return	**−13.00 (2.54)**	**[−17.98,−8.01]**	**−5.11**	**<0.001**		

Significant effects are bolded. 95% confidence intervals and *p*-values were computed using a Wald *t*-distribution with a Satterthwaite approximation. Number of observations = 7,260; Number of speakers = 18.

SE = 5.56) and Return phase (*M* = −22.93, *SE* = 5.05), and between the Perturbation phase (*M* = −14.86, *SE* = 5.05) and the Return phase, both *p*s < 0.001. In the F1/F2 Coupled Perturbation Condition, speakers' average F2 production also significantly differed in the Baseline phase (*M* = 0.34, *SE* = 7.36) as compared to the Return phase (*M* = 9.93, *SE* = 6.99; *p* = 0.047), and in the Perturbation phase (*M* = 38.26, *SE* = 6.99) as compared to the Return phase, *p* < 0.001. Hence, as in the F1 model, speakers' compensatory behavior in F2 was most pronounced during the F1/F2 Coupled Random Perturbation

Condition, where average relative perturbation magnitudes deviated from zero. A full list of pairwise comparisons and their significance values are provided in the **Supplementary material**. Best-Fit Model coefficients for F2 are shown in **Table 2**.

Two repeated-measures ANOVAs (one for F1, one for F2) were carried out to examine the influence of Condition (F1/F2 Random Perturbation, F1-Only Random Perturbation, F1/F2 Coupled Random Perturbation) and Phase (Baseline, Perturbation, and Return) on average within-subject speech production variability (i.e., standard deviation; SD). In the F1 model, the Phase effect violated the sphericity assumption, Mauchly's Test of Sphericity, $p = 0.002$. The Greenhouse–Geisser correction was thus used to make decisions about the statistical significance of this effect. The main effect of Condition was not significant at the 0.05 level, $F(2,34) = 3.14$, $p = 0.056$, $\eta_p = 0.156$. However, there was a significant Phase effect, $F(1.30,22.09) = 5.20$, $p = 0.025$, $\eta_p = 0.234$. Follow-up comparisons revealed that, on average, speakers were significantly less variable in F1 during the Baseline phase ($M = 20.72$; $SE = 0.984$) than they were during the Perturbation ($M = 24.05$, $SE = 1.21$) and Return ($M = 24.47$, $SE = 1.27$) phases, both $ps < 0.029$. The difference in within-speaker F1 variability in the Perturbation and Return phases was not significant, $p = 0.552$. The interaction between Condition and Phase was also not significant, $F(4,68) = 12.52$, $\eta_p = 0.038$, $p = 0.609$. Average within-subject variability in Experiment 1 is shown in **Figure 4**.

There were no significant effects in the F2 model. Within-speaker standard deviation in F2 did not significantly differ by Condition [$F(2,34) = 0.819$, $p = 0.450$, $\eta_p = 0.046$] or Phase, $F(2,34) = 52.92$, $p = 0.323$, $\eta_p = 0.064$. The interaction between Condition and Phase was also not significant, $F(2.5,42.51) = 0.608$, $p = 0.585$, $\eta_p = 0.035$.

The average amplitude spectrums computed to examine oscillations in speakers' F1 compensatory behavior throughout the Perturbation phase of each condition in Experiment 1 are presented in **Figure 5**. The frequency zero represents the DC-offset and reflects the mean change in normalized F1 values in the Perturbation phase relative to the Baseline phase. As can be seen, the mean amplitude at zero cycles/trial for the F1/F2 Coupled Random Perturbation Condition is numerically larger than the other two conditions. This is consistent with the LMM results above. At higher frequencies, all three conditions display low amplitudes and are intermingled, indicating that there were no prominent oscillations of F1 within the Perturbation phase of any condition.

Discussion

The massive unpredictability of the F1/F2 Random Perturbation Condition and F1-Only Random Perturbation Condition had minimal effects on the formant production characteristics. Variability in the Perturbation and Return phases increased from baseline but only modestly and did so in similar fashions for all three experimental conditions equally for Perturbation and Return phases. While this might be due to the unpredictability of the feedback, our design in these studies does not permit this explanation to be distinguished from a generalized increase in production variance with the extended repetition of the same syllable. This will be examined in Experiment 2.

The average data showed two surprising patterns. First, both the F1/F2 Random Perturbation Condition and the F1-Only Random Perturbation Condition essentially remained at baseline levels. The second surprising result was that the F1/F2 Coupled Random Perturbation Condition, which perturbed the feedback randomly between 0 and +200/−250 Hz (F1/F2) with a mean of +100/−125 Hz (F1/F2), yielded results consistent with a static perturbation of +100/−125 Hz. The observed compensations are approximately 40–50% of the perturbation magnitude, which is consistent with many studies who have used a step perturbation (e.g., Munhall et al., 2009; MacDonald et al., 2011). The results suggest that the compensatory system is integrating feedback error over a sequence of utterances and thus, showing a sensitivity to an average error. In Experiment 2, the temporal consistency of the perturbations will be

TABLE 2 Coefficients from the Best-Fit Model used to predict speakers' normalized F2 values during Experiment 1.

Fixed effects	Estimate (*SE*)	95% CI	*t*-value	*P*-value	Random effects	*SD*
Intercept (F1/F2 condition baseline)	−0.03	[−16.36, 16.30]	−0.004	0.997	*Speaker*	
F1-only condition	−0.17	[−18.73, 18.38]	−0.02	0.985	Intercept (F1/F2 random condition)	19.71
Linear condition	0.37	[−15.12, 15.86]	0.05	0.961	F1-Only condition	19.66
Perturbation phase	3.10	[−2.73, 8.93]	1.04	0.297	Linear condition	17.94
Return phase	−0.36	[−6.18, 5.47]	−0.12	0.905	*Residual*	26.27
F1-only*Perturbation	**−17.76**	**[−26.01,−9.52]**	**−4.22**	**< 0.001**		
Linear*Perturbation	**34.81**	**[26.56, 43.06]**	**8.27**	**<0.001**		
F1-only*Return	**−22.37**	**[−30.62,−14.13]**	**−5.32**	**< 0.001**		
Linear*Return	**9.94**	**[1.69, 18.19]**	**2.36**	**0.018**		

Significant effects are bolded. 95% confidence intervals and *p*-values were computed using a Wald *t*-distribution with a Satterthwaite approximation. Number of observations = 7,247; Number of speakers = 18.

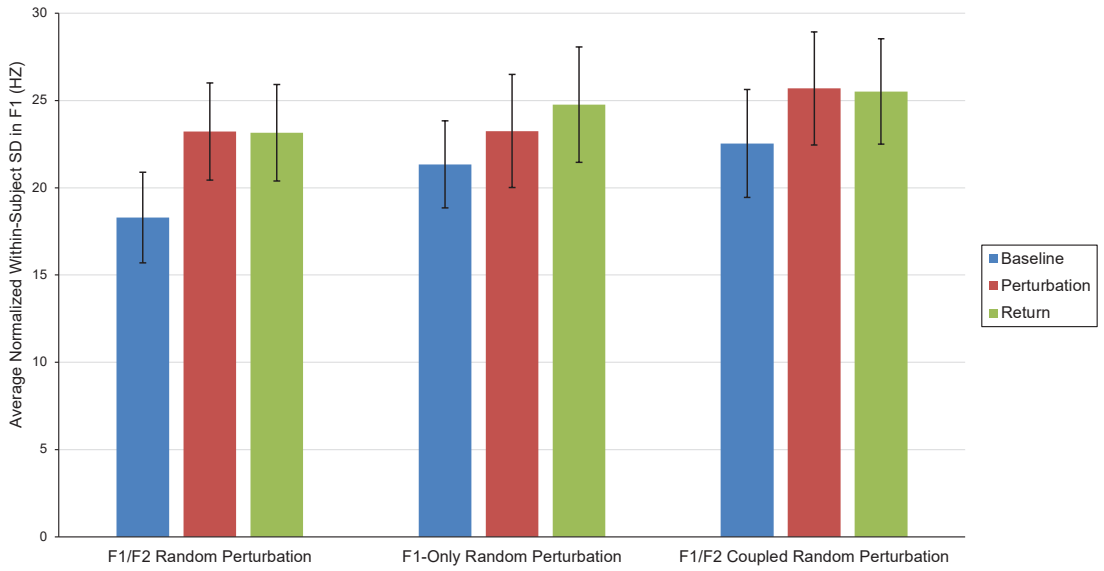


FIGURE 4
Average normalized F1 within-subject variability (i.e., SD) in the Baseline (blue), Perturbation (red), and Return (green) phases of the F1/F2 Random Perturbation Condition, F1-Only Random Perturbation Condition, and F1/F2 Coupled Random Perturbation Condition in Experiment 1. Error bars represent 95% confidence intervals.

manipulated to explore the nature of this integration of feedback error. A step perturbation will also be tested to compare the relative consistency of compensation to a static perturbation versus a variable one such as tested here.

Experiment 2

Our aim in this experiment was to examine whether the feedback system would show greater responsiveness to perturbations held constant for longer periods of time. Such findings would allow us to carry out a preliminary test of the temporal span over which the feedback integrates error information. This experiment also included a non-perturbation control condition where the feedback was held constant, and a step perturbation condition in which feedback was shifted from "head" to "had" during the Perturbation phase.

Materials and methods

The acoustic processing methods used for Experiment 2 were the same as reported above for Experiment 1. The design and procedure for Experiment 2 was similar to Experiment 1. The equipment was functionally similar to Experiment 1. As such, only differences will be described.

Participants

Twenty-two female speakers fluent in Canadian English who did not participate in Experiment 1 were recruited to participate in the study. Two participants were removed from the dataset due to technical issues with the formant perturbation system. The remaining 20 participants ranged in age from 19 to 32 years of age (M_{age} = 22.35; SD_{age} = 2.74) and reported having no speech or language impairments. Fourteen speakers reported being fluent in at least one other language in addition to English. All participants had normal audiometric hearing thresholds between 500 and 4,000 Hz ($\leq$20 dB hearing level) and provided their informed consent prior to participating. All experimental procedures were approved by the General Research Ethics Board at Queen's University.

Equipment

The equipment used for Experiment 2 was the same as previously reported in Nault and Munhall (2020). Participants sat in a different sound attenuated booth (Industrial Acoustic Co. model 1201a), and a different controller was used to produce formant shifts in real-time (National Instruments PXI-8176 embedded controller) than in Experiment 1. All other equipment was functionally the same as reported above for Experiment 1.

Design and procedure

Participants were asked to vocally produce the word "head" 80 times in five different conditions (Control, One, Three, Six, and Step conditions). In the Control Condition, participants received normal, unaltered auditory feedback for all 80 trials. In the four experimental conditions, there were three continuous phases that were not indicated to participants. During the

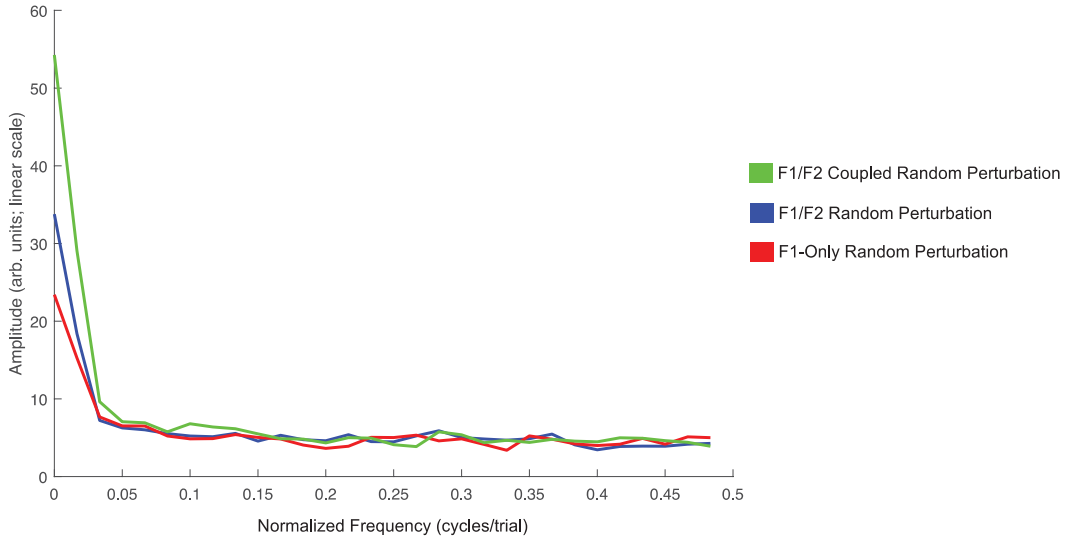


FIGURE 5
Average amplitude spectra across all 18 speakers for the F1/F2 Coupled Random Perturbation Condition (green), F1/F2 Random Perturbation Condition (blue), and F1-Only Random Perturbation Condition (red) in Experiment 1. The left-most frequency bin of 0 cycles/trial represents the DC-offset. It reflects the mean change in normalized F1 values in the Perturbation phase of each condition as compared to Baseline. Instances of peaks in amplitude at higher frequencies would represent prominent oscillation of F1 values across trials in the Perturbation phase. The spectra were created using a discrete Fourier transform with a Hanning window and sampling frequency set to one sample per trial.

Baseline phase (trials 1–20), speakers received normal, unaltered auditory feedback. Speakers' auditory feedback was then manipulated during the Perturbation phase (trials 21–50). In conditions One, Three, and Six, perturbations were applied in F1 and F2 with varying levels of temporal predictability (see **Figure 6**). As in the F1/F2 Coupled Random Feedback Condition in Experiment 1, the feedback perturbations for F1 and F2 were proportional in frequency. Thus, the feedback participants received varied in a linear fashion between the vowel /I/ in "hid" to /æ/ in "had" in F1/F2 space (see **Figure 2C**). In Condition One, a different perturbation was introduced on each trial. In Conditions Three and Six, perturbations were held constant for three and six trials, respectively. In all three conditions, the overall average of the F1 and F2 perturbation values was 0 Hz. During the Perturbation phase of the Step Condition, F1 and F2 perturbations of 200 and −250 Hz, respectively, were maintained for 30 trials (see **Figure 2B**). This is a standard perturbation often used in auditory feedback perturbation studies and it produces a shift across the vowel category boundary from /ɛ/ to /æ/. In all conditions, participants' natural auditory feedback was restored during the Return phase (trials 51–80). The order of conditions was counterbalanced across participants.

In between each condition, the experimenter entered the sound booth, and engaged in a few minutes of conversation with each participant. Participants were also asked to read "The Grandfather Passage" (Van Riper, 1963; Darley et al., 1975) aloud. This seminal 132-word passage is often used in clinical settings to elicit oral reading samples and to assess speech motor functioning and speech intelligibility (e.g., De Bodt et al., 2002) due to its semantic and syntactic complexity and diverse range of English phonemes. It was used in the current experiment to encourage speakers to return to baseline vowel production.

Results

The primary dataset for Experiment 2 included a total of 7,500 utterances (20 speakers * 5 conditions * 75 trials = 7,500). Issues with formant tracking led to the removal of 253 formant values (62 in F1; 191 in F2) from the final dataset. As in Experiment 1, we removed trials 1–5 from the dataset to reduce any possible influence on speech production of task familiarization and speaking while receiving feedback through headphones. We begin by providing a figure of the average normalized results for F1 and F2 in each condition. We then provide results from the Best Fit Models used to predict normalized speech production in F1 and F2. We also report results from within-subjects ANOVAs used to examine within-subject variability. We conclude our Results section with a visual depiction of the average amplitude spectra that were computed to examine oscillations in F1 compensatory behavior throughout the Perturbation phase of each condition.

The average normalized results for F1 and F2 across all three phases of each condition in Experiment 2 are shown in **Figure 7**. As shown, the Step Condition, on average, differed from all other conditions during the Perturbation phase. The Perturbation

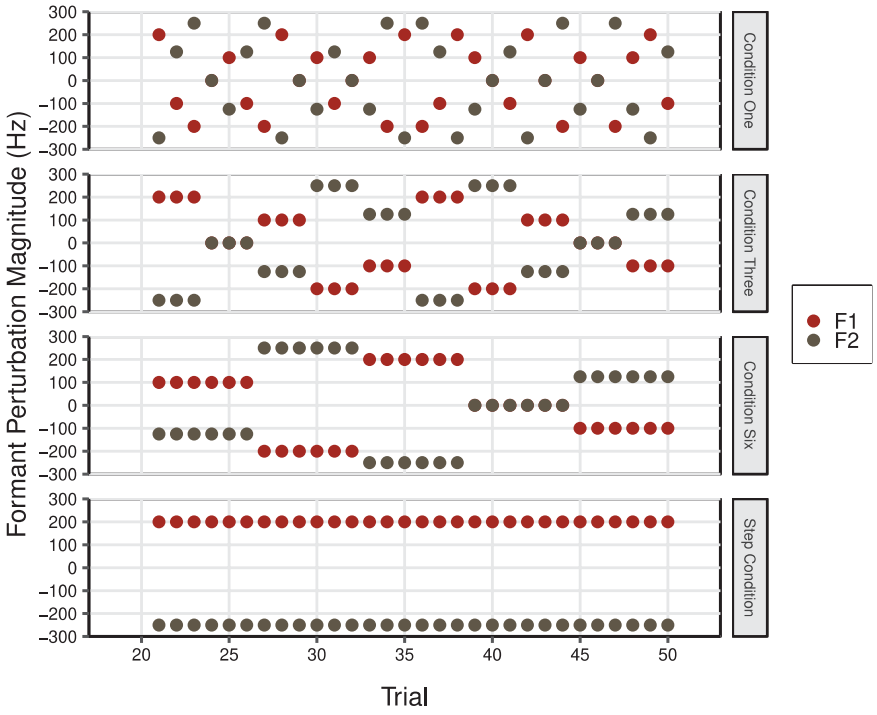


FIGURE 6
F1 (red) and F2 (gray) perturbation values in Hz during the Perturbation phase of Condition One, Condition Three, Condition Six, and the Step Condition of Experiment 2. The overall average F1 and F2 perturbation values in Condition One, Three, and Six was 0 Hz (F1 min = −200 Hz, F1 max = 200 Hz; F2 min = −250 Hz, F2 max = 250 Hz).

phase of Condition Six differed from the Control Condition indicating that sequential consistency of perturbations was required for compensatory behavior. The results for F2 were similar to F1. On average, the Step Condition produced more robust compensations than any of the other conditions. Compensatory behavior was, on average, more evident in Condition Six than it was during the Control Condition, which suggests that the consistency of perturbations across trials was important for compensation. The F2 results were generally more variable than those observed for F1.

The Best-Fit Model used to predict speakers' normalized F1 production values in Experiment 2 produced the best fit to the data and included a maximal random-effects structure with random intercepts for speakers. Including random slopes for condition and phase led to model convergence errors. The Best Fit-Model also included the fixed effects of Condition, Phase, and their interaction term. The Best-Fit Model significantly outperformed a null model that only included the maximal random-effects structure, $\chi 2(14) = 499.25$, $p < 0.001$, as well as alternative models that only included the fixed effect of Condition [$\chi 2(10) = 200.43$, $p < 0.001$] or Phase, $\chi 2(12) = 429.23$, $p < 0.001$. The Best-Fit Model was also a significantly better fit to the data than an alternative model that did not have the interaction term, $\chi 2(8) = 127.85$, $p < 0.001$.

Results from the Best-Fit Model indicated that there was a significant Phase effect. Pairwise comparisons using the Bonferroni correction indicated that speakers' normalized F1 values were significantly more negative during the Perturbation phase ($M = -9.18$, $SE = 2.43$) than during the Return phase ($M = -4.99$, $SE = 2.43$) and Baseline phase ($M = 1.00$, $SE = 2.52$), all $ps < 0.001$. The main effect of Condition was not significant. However, there was a significant interaction between Condition and Phase, which was mainly qualified by significant differences between phases of the Step Condition and phases of all other conditions. Notably, speakers' normalized F1 values were significantly more negative during the Perturbation phase of the Step Condition ($M = -34.34$, $SE = 2.78$) than they were during the Perturbation phases of the Control Condition ($M = 0.89$, $SE = 2.78$), Condition One ($M = -0.44$, $SE = 2.78$), Condition Three ($M = -5.24$, $SE = 2.78$) and Condition Six ($M = -6.79$, $SE = 2.78$), all $ps < 0.001$. Speakers' mean F1 values were also significantly more negative during the Perturbation phase of Condition Six than they were during the Perturbation phase of the Control Condition, $p = 0.039$. Pairwise differences between the Perturbation phases of all other conditions were not significant. Speakers' mean F1 values produced during the Return phase of the Step Condition ($M = -16.40$, $SE = 2.78$) were also significantly more negative than those produced during the Return phase of all other

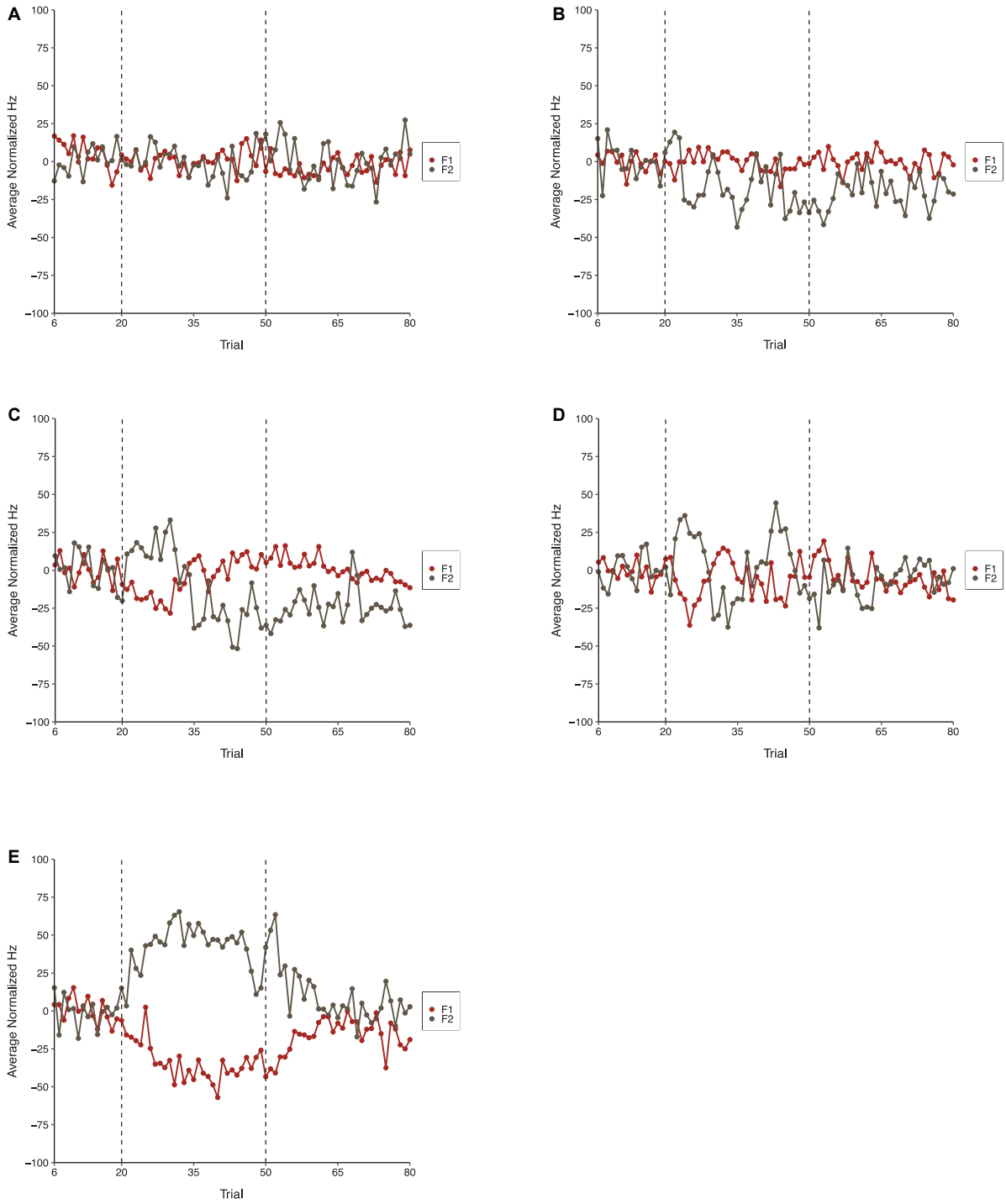


FIGURE 7
Average normalized F1 (red) and F2 (gray) speech production values in the Control Condition **(A)**, Condition One **(B)**, Condition Three **(C)**, Condition Six **(D)**, and the Step Condition **(E)** of Experiment 2. From left to right, dotted lines denote boundaries between the Baseline, Perturbation, and Return phases, respectively.

conditions, all *ps* < 0.001. A full list of pairwise comparisons is provided in the **Supplementary material**. Best-Fit Model coefficients are shown in **Table 3**.

The Best-Fit Model used to predict speakers' normalized F2 productions included a maximal random effects structure with random intercepts for speakers. It also included fixed

TABLE 3 Coefficients from the Best-Fit Model used to predict speakers' normalized F1 values during Experiment 2.

Fixed effects	Estimate (*SE*)	95% CI	*t*-value	*P*-value	Random effects	*SD*
Intercept (control baseline)	5.34 (3.19)	[−1.03, 11.71]	1.68	0.099	*Speaker* (intercept)	10.41
Condition one	−5.34 (3.06)	[−11.35, 0.66]	−1.74	0.081	*Residual*	37.29
Condition three	−5.33 (3.06)	[−11.34, 0.67]	−1.74	0.082		
Condition six	−5.52 (3.07)	[−11.53, 0.49]	−1.80	0.072		
Step condition	−5.49 (3.08)	[−11.52, 0.54]	−1.79	0.074		
Perturbation phase	−4.45 (2.66)	[−9.66, 0.76]	−1.68	0.094		
Return phase	**−8.95 (2.66)**	**[−14.16,−3.74]**	**−3.37**	**<0.001**		
One*Perturbation	4.02 (3.75)	[−3.33, 11.36]	1.07	0.284		
Three*Perturbation	−0.80 (3.75)	[−8.15, 6.55]	−0.21	0.831		
Six*Perturbation	−2.16 (3.75)	[−9.51, 5.19]	−0.58	0.564		
Step*Perturbation	**−29.73 (3.76)**	**[−37.10,−22.37]**	**−7.91**	**<0.001**		
One*Return	**9.04 (3.75)**	**[1.69, 16.38]**	**2.41**	**0.016**		
Three*Return	**9.32 (3.75)**	**[1.97, 16.67]**	**2.49**	**0.013**		
Six*Return	3.75 (3.75)	[−3.61, 11.11]	1.00	0.318		
Step*Return	−7.30 (3.76)	[−14.67, 0.07]	−1.94	**0.052**		

95% confidence intervals and *p*-values computed using a Wald *t*-distribution with a Satterthwaite approximation. Significant effects are bolded. Number of observations = 7,438; Number of speakers = 20.

effects of Condition, Phase, and their interaction term. The Best-Fit Model significantly outperformed a null model that only included the maximal random-effects structure, $\chi 2(14) = 584.43$, $p < 0.001$. It was also a significantly better fit to the data than alternative models that only included the fixed effect of Condition [$\chi 2(10) = 222.54$, $p < 0.001$] or Phase, $\chi 2(12) = 520.20$, $p < 0.001$. The Best-Fit Model significantly outperformed an alternative model that did not include the interaction between Condition and Phase, $\chi 2(8) = 156.21$, $p < 0.001$.

In the Best Fit Model for F2, the main effects of Condition and Phase were not significant. However, there was a significant interaction between these effects. As in the F1 model, the interaction was mainly explained by significant differences between phases of the Step Condition and phases of all other conditions. Importantly, speakers' average F2 values were significantly more positive during the Perturbation phase of the Step Condition ($M = 42.47$, $SE = 3.90$) than they were during the Perturbation phases of the Control Condition ($M = -0.25$, $SE = 3.91$), Condition One ($M = -16.62$, $SE = 3.92$), Condition Three ($M = -10.95$, $SE = 3.91$), and Condition Six ($M = 2.72$, $M = 3.91$), all $ps < 0.001$. Speakers' mean F2 values were significantly more negative during the Perturbation phase of Condition One than they were during the Perturbation phases of the Control Condition and Condition Six, both $ps < 0.001$. Speakers' mean F2 values were significantly more positive during the Perturbation phase of Condition Six than they were during the Perturbation phase of Condition Three, $p = 0.004$. As in the F1 model, there were also a number of significant mean differences between formant values produced during the Return phases of different conditions. A full list of pairwise comparisons is provided in the **Supplementary material**. Best-Fit Model coefficients for F2 are presented in **Table 4**.

Two repeated-measures ANOVAs (one for F1, one for F2) were conducted to examine whether within-speaker speech production variability (i.e., SD) differed by Condition (Control, One, Three, Six, and Step) and Phase (Baseline, Perturbation, and Return). One outlier in F1 that was more than three standard deviations from the mean was Winsorized and replaced with the next highest value in the dataset. The F1 model revealed that speakers' mean speech production variability did not significantly differ by Condition, $F(4,76) = 0.631$, $p = 0.642$, $\eta_p = 0.032$. However, there was a significant main effect of Phase, $F(2,38) = 4.85$, $p = 0.013$, $\eta_p = 0.203$. Pairwise comparisons showed that speakers' F1 productions were significantly more variable during the Perturbation phase ($M_{SD} = 31.64$) than they were during the Baseline phase ($M_{SD} = 28.47$), $p = 0.018$. There were no statistically significant differences in within-speaker variability between the Baseline and Return ($M_{SD} = 29.75$) phases ($p = 0.172$), nor between the Perturbation and Return phases, $p = 0.052$. The interaction between Condition and Phase was only marginally significant, $F(8,152) = 2.00$, $p = 0.050$, $\eta_p = 0.095$. Using the Bonferroni correction to adjust for multiple comparisons, it was determined that none of the interaction comparisons were significant, all $ps > 0.059$. Notably, there was no significant difference in within-subject variability between the Baseline ($M_{SD} = 30.42$), Perturbation ($M_{SD} = 29.16$), and Return ($M_{SD} = 28.65$) phases of the Control Condition, all $ps > 0.05$.

In the F2 model, the within-subjects effect of Condition and the interaction between Condition and Phase violated the sphericity assumption, Mauchly's Test of Sphericity, $ps < 0.05$.

TABLE 4 Coefficients from the Best-Fit Model used to predict speakers' normalized F2 values during Experiment 2.

Fixed effects	Estimate (*SE*)	95% CI	*t*-value	*P*-value	Random effects	*SD*
Intercept (control baseline)	1.35 (4.55)	[−7.72, 10.41]	0.30	0.768	*Speaker* (intercept)	197.1
Condition one	−1.23 (4.65)	[−10.34, 7.88]	−0.26	0.791	*Residual*	3163.3
Condition three	−1.37 (4.67)	[−10.53, 7.79]	−0.29	0.769		
Condition six	−1.29 (4.66)	[−10.43, 7.85]	−0.28	0.782		
Step condition	−1.54 (4.67)	[−10.69, 7.61]	−0.33	0.742		
Perturbation phase	−1.60 (4.03)	[−9.51, 6.31]	−0.40	0.692		
Return phase	−1.86 (4.04)	[−9.78, 6.07]	−0.46	0.646		
One*Perturbation	**−15.14 (5.70)**	**[−26.31,−3.97]**	**−2.66**	**0.008**		
Three*Perturbation	−9.33 (5.71)	[−20.53, 1.87]	−1.63	0.103		
Six*Perturbation	4.26 (5.71)	[−6.93, 15.44]	0.75	0.456		
Step*Perturbation	**44.25 (5.70)**	**[33.07, 55.43]**	**7.76**	**<0.001**		
One*Return	**−17.89 (5.69)**	**[−29.05,−6.73]**	**−3.14**	**0.002**		
Three*Return	**−22.14 (5.72)**	**[−33.35,−10.94]**	**−3.87**	**<0.001**		
Six*Return	−4.99 (5.71)	[−16.18, 6.21]	−0.87	0.383		
Step*Return	**11.34 (5.71)**	**[0.14, 22.54]**	**1.98**	**0.047**		

Significant effects are bolded. Number of observations = 7,309; Number of speakers = 20.

The Greenhouse-Geisser correction was thus used in making decisions about significance. As in the F1 model, the main effect of Condition was not significant, $F(2.68,50.96) = 0.499$, $p = 0.664$, $\eta_p = 0.026$. However, there was a significant main effect of Phase, $F(1.58,30.09) = 5.31$, $p = 0.016$, $\eta_p = 0.218$. Follow-up comparisons revealed that speakers were significantly more variable in F2 during the Perturbation phase ($M_{SD} = 51.44$) than they were during the Baseline phase ($M_{SD} = 45.12$), $p = 0.012$. Within-speaker production variability did not significantly differ between the Baseline and Return ($M_{SD} = 48.59$) phases ($p = 0.106$), nor between the Perturbation and Return phases, $p = 0.054$. The interaction between Condition and Phase was not significant, $F(4.70,89.22) = 2.25$, $p = 0.060$, $\eta_p = 0.106$. A visual depiction of the Phase effect in F1 and F2 is shown in **Figure 8**.

The spectra shown in **Figure 9** summarize the findings for F1 in Experiment 2. The DC-offset (seen at frequency 0 cycles/trial) shows the only major difference. The Step Condition is larger than the other conditions at this frequency. Condition Six is trending in the same direction. Otherwise, across conditions, there are no differences at higher frequencies in the spectra.

One possible explanation for compensation being significantly more pronounced in the Perturbation phase of the Step Condition and Condition Six than in Condition One and Condition Three is that the feedback error was held constant for a greater number of trials in these two conditions and thus, the error correction system was responding to more stable and predictable conditions.

We computed a series of bivariate correlations between F1/F2 perturbation values that were applied in the Perturbation phase of Condition One, Condition Three, and Condition Six and average normalized F1/F2 production values across all subjects from these three conditions[3]. Correlations could not be computed for the Step or Control Conditions due to the F1/F2 perturbation values being held constant throughout the entire Perturbation phases. Correlations were computed at four lags: zero (simultaneous), one, three, and five trials. Our reasoning was that a comparison between simultaneous and time-lagged correlations would provide insights into whether the error correction system was operating instantaneously, or whether it was integrating information over time.

A visual depiction of the average results from the bivariate correlations in F1 and F2 are shown in **Figure 10**. More negative correlation values indicate stronger compensatory responses.

As can be seen in each condition, the average simultaneous correlation values are much lower (i.e., closer to zero or more positive) than the average lag correlation values. This is particularly the case in Condition Three and Condition Six, where the feedback perturbations were applied in a more consistent and stable manner during the Perturbation phase.

Discussion

As in Experiment 1, only the introduction of perturbations that consistently deviated from baseline in direction and magnitude produced significant shifts across the Perturbation phase. The step change compensations resembled those observed in other studies that introduced such perturbations

3 We also computed correlations at the individual participant level, and they showed similar trends. Due to space limitations, these correlations were not included in the main text of the manuscript. They are publicly available on OSF here: osf.io/n4pgf.

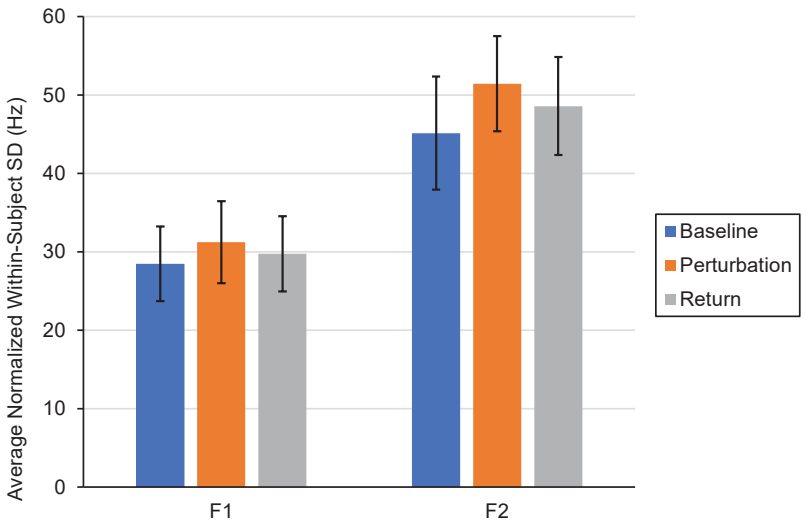


FIGURE 8
Average normalized F1 and F2 within-subject variability (i.e., SD) in the Baseline (blue), Perturbation (orange), and Return (gray) phases of Experiment 2. Error bars represent 95% confidence intervals.

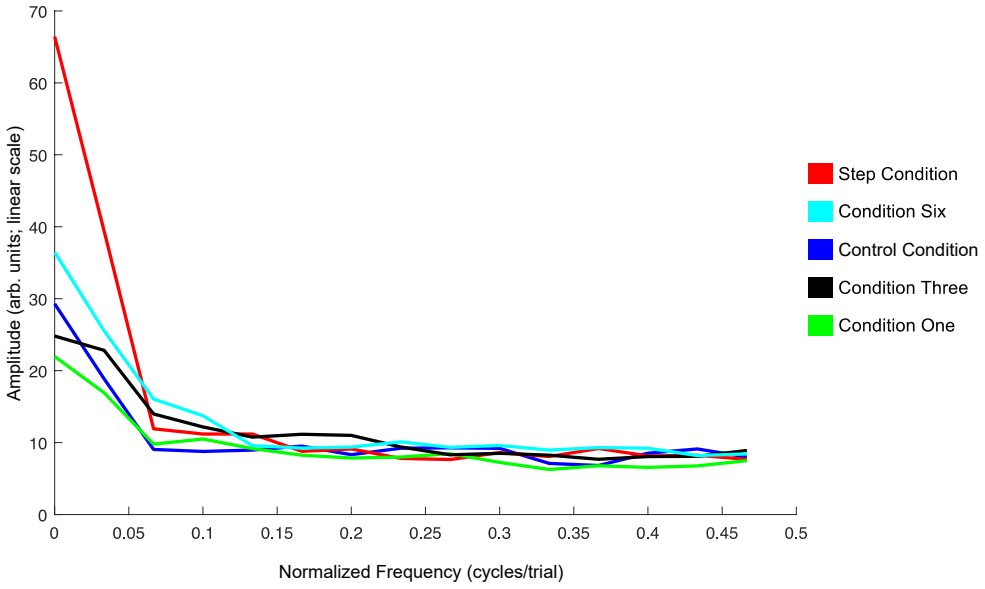


FIGURE 9
Average amplitude spectra across all 20 speakers for the Step Condition (red), Condition Six (cyan), Control Condition (blue), Condition Three (black), and Condition One (green) in Experiment 2. The left-most frequency bin of 0 cycles/trial represents the DC-offset. It reflects the mean change in normalized F1 values in the Perturbation phase of each condition as compared to Baseline. Instances of peaks in amplitude at higher frequencies would represent prominent oscillation of F1 values across trials in the Perturbation phase. The spectra were created using a discrete Fourier transform with a Hanning window and sampling frequency set to one sample per trial.

(e.g., Munhall et al., 2009; MacDonald et al., 2011). The different length of perturbations (1, 3, and 6 trials) did not significantly differ from each other, although the Six Condition was significantly different from the Control Condition. This finding is consistent with the idea that feedback deviations are compensated incrementally over trials and that six trials is within the span that is required for compensation to develop whereas one and three trials are too short for systematic change to develop in response to perceived errors. The lag correlation findings are consistent with this idea of a span of compensation.

For both F1 and F2, variability increased in the Perturbation phases of all conditions, and this was particularly true for F2.

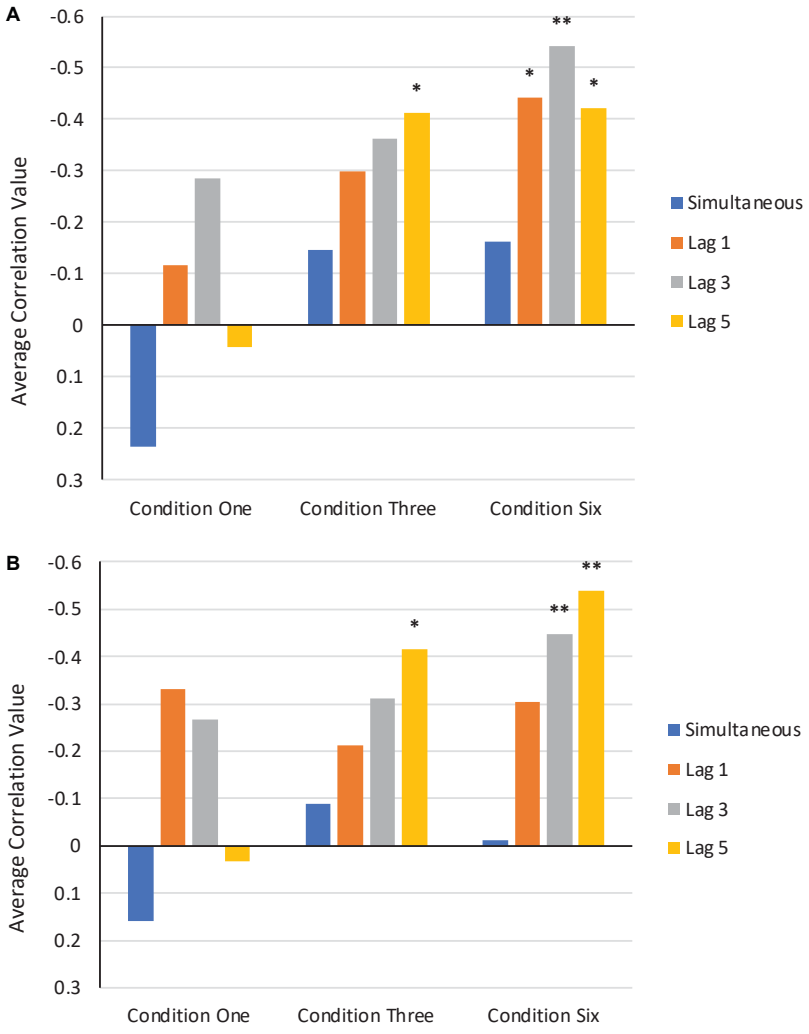


FIGURE 10
Average bivariate correlations in F1 **(A)** and F2 **(B)** between perturbation values applied in the Perturbation phase of Condition One, Condition Three, and Condition Six and participants' normalized production values. Simultaneous correlations are shown in blue. Correlations at trial lags of one, three, and five are shown in orange, gray, and yellow, respectively. *Correlation is significant at the 0.05 level. **Correlation is significant at the 0.01 level.

This finding contrasts with other studies that have not shown an increase in variability in perturbation phases (e.g., Nault and Munhall, 2020). Notably, the increase in within-subject variability during the Perturbation phase does not appear to be due to participant fatigue from being asked to say the same word repeatedly, as variability in F1 and F2 did not significantly differ in the three phases of the Control Condition. Rather, the increase in variability appears to be due to the unpredictability of the feedback in the experimental conditions.

General discussion

The experiments presented here are part of a broad literature in speech, limb, and eye movements that examine a subtype of motor learning called adaptation. Adaptive responses are designed to maintain the accuracy and stability of movements that are already learned when environmental conditions change, or when sensory perception is noisy. Adaptation is thus usually studied in paradigms that focus on reducing error following some form of perturbation. While compensatory response to auditory feedback perturbations is well documented, here we examined the speech compensation when the sensory feedback returned unpredictable errors.

Across a number of different interpretations of randomness in feedback, our results indicate that the use of auditory feedback in speech motor control is governed by the relevance of the feedback. Talkers acted, on average, like random feedback was irrelevant and average performance did not change from no perturbation conditions. The exceptions to this summary were

the three conditions that showed consistent error signals (Exp. 1: F1/F2 Coupled Random Perturbation Condition; Exp. 2: Step Condition, Six Condition). In each of these conditions, the error introduced to the feedback was relatively consistent over a span of utterances. This finding is consistent with other indicators that the auditory feedback system eschews correcting the error when the deviation is too large (MacDonald et al., 2010) or if the temporal delay is too great (Mitsuya et al., 2017). Perceptual and motor learning also requires information about the statistics of the environment, and non-stationary environments provide challenges to learning (e.g., Petrov et al., 2006; Narain et al., 2013). When sensory uncertainty exists, it is thought that subjects rely more on their prior estimates of the structure of the task (Körding and Wolpert, 2004). The detection of the uncertainty of the sensory information can be seen as equivalent to the relevance of feedback to performance of a task.

An outstanding issue is whether there is some flexibility in the use of auditory feedback in speech control. Lametti et al. (2012) suggested that individuals prioritized different sources of sensory information. Some people were more influenced by auditory feedback, while others were more reliant on somatosensory signals. In contrast to these individual differences in sensory processing are studies that indicate contextual modification of use of the auditory signal. There are indications that auditory errors can have reduced impact on speech if the signals seem irrelevant [see Wei and Körding (2009) for a study of feedback relevance in limb movements]. Daliri and Dittman (2019) used a 'clamping' technique in which the auditory feedback was not contingent on the talker's productions. The error was constant even when the talker compensated. This 'irrelevant' feedback, which was not contingent on the talkers' behavior, reduced the magnitude of adaptation.

The increase in variability in the Perturbation phases of the current experiments may be indicative of a destabilizing effect of the random perturbations. While our repeated measures designs and the repetitive nature of our protocols are possible explanations as well, within-subject variability did not significantly differ in F1 or F2 in the Control Condition in Experiment 2. However, the heightened variability in the Return phase of Experiment 1 is consistent with this possibility. While we are using the relative variability as a measure of the system's organization of auditory feedback processing, there are other possible contributions to changes in variability. Bays and Wolpert (2007) review a number of computational ways that the motor system can reduce the unpredictability of sensory information and thus counteract the potentially destabilizing effects of feedback uncertainty. One of these solutions is the integration of multisensory information to improve prediction. The importance of both somatosensory and auditory information in speech motor control is highlighted in theoretical accounts (e.g., Tourville and Guenther, 2011), although the experimental study of dynamic auditory and proprioceptive cues are technically difficult and infrequently attempted (cf., Lametti et al., 2012).

Auditory feedback processing as studied in the laboratory setting has many of the characteristics of phenomena that have driven concerns about the Reliability Paradox (see recent symposium at the Psychonomics Society 2021 meeting). There are a number of phenomena which are robust at a group average level but are not always apparent at the individual subject level (see Nault and Munhall, 2020). Test–retest reliability is also not strong in phenomena that are frequently included in clinical test batteries (e.g., the Stroop test, Implicit Association Test). The lack of robustness at the individual participant level of auditory feedback effects is somewhat unsettling. How can an error-correction system that is supposedly guiding speech motor control be so difficult to demonstrate? One answer is that auditory feedback is not necessary or sufficient for the control of learned speech sequences. Evidence from those who are deafened as adults can be interpreted as supporting this suggestion. While precision of some phonemes degrades, it does so slowly over time and not completely (Cowie et al., 1982). A second answer is that the precision and need for error-based correction of speech is overrated. Fluent speech is a remarkable motor skill, but its required precision is not as high as some manual skills (Uccelli et al., 2021), microsaccades (Poletti et al., 2020) and perhaps less than the bite force requirements of the mandible in chewing. In an analysis of the Switchboard Corpus, Greenberg (1999) reported that significant proportions of phonemes are substituted or deleted in this database. This indicates that intelligibility in communication does not always require the kind of error-correcting precision that the feedback paradigm might suggest.

Another contributing factor in formant-feedback processing is error in measurement (Shadle et al., 2016), particularly in speech produced with higher fundamental frequencies. This problem will have an impact on the data quality but can also have an impact on the quality of the perturbations. In addition to the difficulties associated with formant tracking, the data used to summarize performance makes assumptions about what feedback parameter is important for the talker. It is common, such as was done in the experiments presented here, to use an average formant frequency measured near the midpoint of the vowel. However, talkers may be using other aspects of vowels to control articulation than static indices of formant frequency. Vowels have inherent formant dynamics that vary with dialect, age, and gender of speakers (e.g., Stanley et al., 2021). These dynamics can influence compensatory behavior with participants correcting for changes in spectral trajectories (Jibson, 2020).

Overall, the present results are consistent with a control system that takes into account the statistics of the sensory environment. Two of the conditions point to this conclusion.

In Experiment 1, the F1/F2 Coupled Random Perturbation Condition had a mean perturbation value that differed from the baseline value across the 30 trials. This restricted or biased random error signal generated a compensatory response reflecting the average. In Experiment 2, keeping the perturbation constant for six trials also produced differential response from the pattern of responses for shorter perturbations. Our lag correlation analysis in Experiment 2 is also indicative of a control system that is not instantaneously responsive to introduced error. Rather, it appears to be sensitive to the consistency and reliability of the error, integrating information and initiating compensatory behavior over a longer time span. In the context that we are testing, more specific studies focused on the predictability shown in these conditions and how the nervous system computes the consistency are warranted (Burge et al., 2008).

Data availability statement

The raw data supporting the conclusions of this article are publicly available on OSF here: https://osf.io/n4pgf.

Ethics statement

The studies involving human participants were reviewed and approved by the Health Sciences Research Ethics Board at Western University (Experiment 1) and the General Research Ethics Board at Queen's University (Experiment 2). All participants provided their written informed consent prior to participating in these studies.

Author contributions

DRN collected the data for Experiment 2, performed data analyses for both experiments, and assisted with the writing and editing of this manuscript. TM collected the data for Experiment 1. DWP performed a portion of the data analysis, implemented the experimental system, and helped with manuscript editing. KGM contributed to all aspects of this manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by Discovery Grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada to KGM and DWP.

Acknowledgments

The authors thank Ruth Norman for assistance with some data collection.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnhum.2022.905365/full#supplementary-material

References

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang.* 68, 255–278.

Bates, D., Kliegl, R., Vasishth, S., and Baayen, H. (2015). Parsimonious mixed models. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1506.04967 (accessed March 3, 2022).

Bays, P. M., and Wolpert, D. M. (2007). Computational principles of sensorimotor control that minimize uncertainty and variability. *J. Physiol.* 578, 387–396. doi: 10.1113/jphysiol.2006.120121

Burge, J., Ernst, M. O., and Banks, M. S. (2008). The statistical determinants of adaptation rate in human reaching. *J. Vis.* 8:20. doi: 10.1167/8.4.20

Cowie, R., Douglas-Cowie, E., and Kerr, A. G. (1982). A study of speech deterioration in post- lingually deafened adults. *J. Laryngol. Otol.* 96, 101–112.

Daliri, A., and Dittman, J. (2019). Successful auditory motor adaptation requires task-relevant auditory errors. *J. Neurophysiol.* 122, 552–562. doi: 10.1152/jn.00662.2018

Darley, F. L., Aronson, A. E., and Brown, J. R. (1975). *Motor speech disorders*. Philadelphia, PA: WB Saunders Company.

De Bodt, M. S., Huici, M. E. H. D., and Van De Heyning, P. H. (2002). Intelligibility as a linear combination of dimensions in dysarthric speech. *J. Commun. Disord.* 35, 283–292. doi: 10.1016/S0021-9924(02)00065-5

Denes, P. B., and Pinson, E. N. (1973). *The Speech Chain: The Physics and Biology of Spoken Language*. Garden City, NY: Anchor Press.

Dhawale, A. K., Smith, M. A., and Ölveczky, B. P. (2017). The role of variability in motor learning. *Annu. Rev. Neurosci.* 40, 479–498. doi: 10.1146/annurev-neuro-072116-031548

Faisal, A. A., Selen, L. P., and Wolpert, D. M. (2008). Noise in the nervous system. *Nat. Rev. Neurosci.* 9, 292–303. doi: 10.1038/nrn2258

Greenberg, S. (1999). Speaking in shorthand–A syllable-centric perspective for understanding pronunciation variation. *Speech Commun.* 29, 159–176. doi: 10.1016/S0167-6393(99)00050-3

Heald, S. L., and Nusbaum, H. C. (2015). Variability in vowel production within and between days. *PLoS One* 10:e0136791. doi: 10.1371/journal.pone.0136791

Heinks-Maldonado, T. H., and Houde, J. F. (2005). Compensatory responses to brief perturbations of speech amplitude. *Acoust. Res. Lett. Online* 6, 131–137. doi: 10.1121/1.1931747

Houde, J. F., and Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science* 279, 1213–1216. doi: 10.1126/science.279.5354.1213

Jibson, J. (2020). "Self-perception and vowel inherent spectral change," in *Proceedings of the 2020 Meetings on Acoustics 179ASA*, Vol. 42, (Acoustical Society of America), 060020. doi: 10.1121/2.0001501

Kawahara, H. (1995). "Hearing voice: Transformed auditory feedback effects on voice pitch control," in *Proceedings of the International Joint Conference on Artificial Intelligence: Workshop on Computational Auditory Scene Analysis*, Montreal, 143–148.

Körding, K. P., and Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature* 427, 244–247. doi: 10.1038/nature02169

Krakauer, J. W., Hadjiosif, A. M., Xu, J., Wong, A. L., and Haith, A. M. (2019). Motor learning. *Compr. Physiol.* 9, 613–663. doi: 10.1002/cphy.c170043

Lametti, D. R., Nasir, S. M., and Ostry, D. J. (2012). Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback. *J. Neurosci.* 32, 9351–9358. doi: 10.1523/JNEUROSCI.0404-12.2012

Lenth, R. V. (2019). *emmeans: Estimated Marginal Means, Aka Least-squares Means. R Package Version 1.7.0*.

MacDonald, E. N., Goldberg, R., and Munhall, K. G. (2010). Compensations in response to real-time formant perturbations of different magnitudes. *J. Acoust. Soc. Am.* 127, 1059–1068. doi: 10.1121/1.3278606

MacDonald, E. N., Purcell, D. W., and Munhall, K. G. (2011). Probing the independence of formant control using altered auditory feedback. *J. Acoust. Soc. Am.* 129, 955–965. doi: 10.1121/1.3531932

MATLAB (2020b). *Version 9.9.0 (R2010b)*. Natick, MA: The MathWorks Inc.

Miller, N. (1992). Variability in speech dyspraxia. *Clin. Linguist. Phonet.* 6, 77–85. doi: 10.3109/02699209208985520

Mitsuya, T., MacDonald, E. N., and Munhall, K. G. (2014). Temporal control and compensation for perturbed voicing feedback. *J. Acoust. Soc. Am.* 135, 2986–2994. doi: 10.1121/1.4871359

Mitsuya, T., Munhall, K. G., and Purcell, D. W. (2017). Modulation of auditory-motor learning in response to formant perturbation as a function of delayed auditory feedback. *J. Acoust. Soc. Am.* 141, 2758–2767. doi: 10.1121/1.4981139

Munhall, K. G., MacDonald, E. N., Byrne, S. K., and Johnsrude, I. (2009). Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate. *J. Acoust. Soc. Am.* 125, 384–390. doi: 10.1121/1.3035829

Narain, D., van Beers, R. J., Smeets, J. B., and Brenner, E. (2013). Sensorimotor priors in nonstationary environments. *J. Neurophysiol.* 109, 1259–1267. doi: 10.1152/jn.00605.2012

Nault, D. R., and Munhall, K. G. (2020). Individual variability in auditory feedback processing: Responses to real-time formant perturbations and their relation to perceptual acuity. *J. Acoust. Soc. Am.* 148, 3709–3721. doi: 10.1121/10.0002923

Orfanidis, S. J. (1988). *Optimum signal processing: An introduction*. New York, NY: Macmillan publishing company.

Parrell, B., Lammert, A. C., Ciccarelli, G., and Quatieri, T. F. (2019a). Current models of speech motor control: A control-theoretic overview of architectures and properties. *J. Acoust. Soc. Am.* 145, 1456–1481. doi: 10.1121/1.5092807

Parrell, B., Ramanarayanan, V., Nagarajan, S., and Houde, J. (2019b). The FACTS model of speech motor control: Fusing state estimation and task-based control. *PLoS Comput. Biol.* 15:e1007321. doi: 10.1371/journal.pcbi.1007321

Patri, J. F., Diard, J., and Perrier, P. (2019). Modeling sensory preference in speech motor planning: A Bayesian modeling framework. *Front. Psychol.* 10:2339. doi: 10.3389/fpsyg.2019.02339

Patri, J. F., Perrier, P., Schwartz, J. L., and Diard, J. (2018). What drives the perceptual change resulting from speech motor adaptation? Evaluation of hypotheses in a Bayesian modeling framework. *PLoS Comput. Biol.* 14:e1005942. doi: 10.1371/journal.pcbi.1005942

Petrov, A. A., Dosher, B. A., and Lu, Z. L. (2006). Perceptual learning without feedback in non- stationary contexts: Data and model. *Vis. Res.* 46, 3177–3197. doi: 10.1016/j.visres.2006.03.022

Poletti, M., Intoy, J., and Rucci, M. (2020). Accuracy and precision of small saccades. *Sci. Rep.* 10:16097. doi: 10.1038/s41598-020-72432-6

Purcell, D. W., and Munhall, K. G. (2006). Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. *J. Acoust. Soc. Am.* 120, 966–977. doi: 10.1121/1.2217714

R Core Team (2020). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Raharjo, I., Kothare, H., Nagarajan, S. S., and Houde, J. F. (2021). Speech compensation responses and sensorimotor adaptation to formant feedback perturbations. *J. Acoust. Soc. Am.* 149, 1147–1161. doi: 10.1121/10.0003440

Riley, M. A., and Turvey, M. T. (2002). Variability and determinism in motor behavior. *J. Motor Behav.* 34, 99–125. doi: 10.1080/00222890209601934

Shadle, C. H., Nam, H., and Whalen, D. H. (2016). Comparing measurement errors for formants in synthetic and natural vowels. *J. Acoust. Soc. Am.* 139, 713–727. doi: 10.1121/1.4940665

Sosa, A. V. (2015). Intraword variability in typical speech development. *Am. J. Speech Lang. Pathol.* 24, 24–35. doi: 10.1044/2014_AJSLP-13-0148

Stanley, J. A., Renwick, M. E., Kuiper, K. I., and Olsen, R. M. (2021). Back Vowel dynamics and distinctions in Southern American English. *J. Engl. Linguist.* 49, 389–418.

Sternad, D. (2018). It's not (only) the mean that matters: Variability, noise and exploration in skill learning. *Curr. Opin. Behav. Sci.* 20, 183–195. doi: 10.1016/j.cobeha.2018.01.004

Tourville, J. A., and Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Lang. Cogn. Process.* 26, 952–981. doi: 10.1080/01690960903498424

Uccelli, S., Pisu, V., and Bruno, N. (2021). Precision in grasping: Consistent with Weber's law, but constrained by "safety margins". *Neuropsychologia* 163:108088. doi: 10.1016/j.neuropsychologia.2021.108088

Van Riper, C. (1963). *Speech correction* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Villacorta, V. M., Perkell, J. S., and Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *J. Acoust. Soc. Am.* 122, 2306–2319. doi: 10.1121/1.2773966

Wei, K., and Körding, K. (2009). Relevance of error: What drives motor adaptation? *J. Neurophysiol.* 101, 655–664. doi: 10.1152/jn.90545.2008

Whalen, D. H., Chen, W. R., Tiede, M. K., and Nam, H. (2018). Variability of articulator positions and formants across nine English vowels. *J. Phon.* 68, 1–14. doi: 10.1016/j.wocn.2018.01.003

frontiers | Frontiers in Human Neuroscience

TYPE Original Research
PUBLISHED 15 September 2022
DOI 10.3389/fnhum.2022.885074

Check for updates

OPEN ACCESS

EDITED BY
Lucie Menard,
Université du Québec à Montréal, Canada

REVIEWED BY
Robin Karlin,
University of Wisconsin–Madison, United States
Shanqing Cai,
Boston University, United States

*CORRESPONDENCE
Miriam Oschkinat
Miriam.Oschkinat@phonetik.uni-muenchen.de

SPECIALTY SECTION
This article was submitted to Speech and Language, a section of the journal Frontiers in Human Neuroscience

RECEIVED 27 February 2022
ACCEPTED 05 August 2022
PUBLISHED 15 September 2022

CITATION
Oschkinat M, Hoole P, Falk S and Dalla Bella S (2022) Temporal malleability to auditory feedback perturbation is modulated by rhythmic abilities and auditory acuity. *Front. Hum. Neurosci.* 16:885074. doi: 10.3389/fnhum.2022.885074

COPYRIGHT
© 2022 Oschkinat, Hoole, Falk and Dalla Bella. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Temporal malleability to auditory feedback perturbation is modulated by rhythmic abilities and auditory acuity

Miriam Oschkinat[1]*, Philip Hoole[1], Simone Falk[2,3,4] and Simone Dalla Bella[2,4,5,6]

[1]Institute of Phonetics and Speech Processing, Ludwig Maximilian University of Munich, Munich, Germany, [2]International Laboratory for Brain, Music and Sound Research, Montreal, QC, Canada, [3]Department of Linguistics and Translation, University of Montreal, Montreal, QC, Canada, [4]Centre for Research on Brain, Language and Music, Montreal, QC, Canada, [5]Department of Psychology, University of Montreal, Montreal, QC, Canada, [6]University of Economics and Human Sciences in Warsaw, Warsaw, Poland

Auditory feedback perturbation studies have indicated a link between feedback and feedforward mechanisms in speech production when participants compensate for applied shifts. In spectral perturbation studies, speakers with a higher perceptual auditory acuity typically compensate more than individuals with lower acuity. However, the reaction to feedback perturbation is unlikely to be merely a matter of perceptual acuity but also affected by the prediction and production of precise motor action. This interplay between prediction, perception, and motor execution seems to be crucial for the timing of speech and non-speech motor actions. In this study, to examine the relationship between the responses to temporally perturbed auditory feedback and rhythmic abilities, we tested 45 adult speakers on the one hand with a temporal auditory feedback perturbation paradigm, and on the other hand with rhythm perception and production tasks. The perturbation tasks temporally stretched and compressed segments (onset + vowel or vowel + coda) in fluent speech in real-time. This technique sheds light on the temporal representation and the production flexibility of timing mechanisms in fluent speech with respect to the structure of the syllable. The perception tasks contained staircase paradigms capturing duration discrimination abilities and beat-alignment judgments. The rhythm production tasks consisted of finger tapping tasks taken from the BAASTA tapping battery and additional speech tapping tasks. We found that both auditory acuity and motor stability in finger tapping affected responses to temporal auditory feedback perturbation. In general, speakers with higher auditory acuity and higher motor variability compensated more. However,

we observed a different weighting of auditory acuity and motor stability dependent on the prosodic structure of the perturbed sequence and the nature of the response as purely online or adaptive. These findings shed light on the interplay of phonological structure with feedback and feedforward integration for timing mechanisms in speech.

KEYWORDS

temporal auditory feedback perturbation, feedforward malleability, auditory acuity, finger tapping, rhythmic abilities

Introduction

In speech production, speakers execute speech movements based on learned internal representations (feedforward system) while the sensory experience of the produced outcome, such as auditory or somatosensory feedback, serves to monitor and evaluate the process constantly (feedback system). The interaction of feedback and feedforward systems in speech production has been of significant interest in speech research and has mainly been probed with real-time auditory feedback perturbations. In auditory feedback perturbation paradigms, speakers hear their voice over headphones while one or more parameters in the acoustic signal are altered in (almost) real-time. In response, speakers were found to counteract the applied feedback shift (*compensate*) in production. Compensation was classified as purely an *online response* when adjustments occurred ∼120–200 ms after perturbation onset in the ongoing production process (Burnett et al., 1998; Purcell and Munhall, 2006). When speakers compensated in future productions of the same/similar unperturbed speech segments, they were said to *adapt*. Online responses hereby support the incorporation of auditory feedback into the control level, adaptation indicates an update of the underlying motor plan for the respective production. While auditory feedback perturbations in the spectral domain (e.g., formant or pitch shifts) have been extensively studied, a few recent studies have started to investigate the role of auditory feedback for speech *timing* (e.g., Cai et al., 2011; Mitsuya et al., 2014; Floegel et al., 2020; Oschkinat and Hoole, 2020, 2022; Karlin et al., 2021). Analogously to spectral feedback perturbations, the majority of speakers was found to compensate for temporal feedback shifts both in the online control as well as in future productions (adaptation), supporting the incorporation of auditory feedback into speech timing mechanisms on control and planning levels. However, there are crucial differences between responses to spectral and temporal auditory feedback perturbations. In spectral perturbations, online responses to altered feedback could be observed in either direction (e.g., with an increase or decrease of formant frequencies in production). For the temporal domain, this bidirectionality is not given naturally: While it is perhaps possible to lengthen a sequence in production that was perceived shorter (more specifically, terminated early) or react to a delay in the auditory feedback, it is not possible to shorten segments online as a reaction to longer percepts (as the sequence is already terminated in production when the auditory stretch is received). Therefore, every shortening response as a reaction to a stretched speech signal must be adaptive. Further, responses in reaction to the perturbation that are not necessarily compensatory and not at the perturbation site itself (e.g., lengthening of following segments as a reaction to a preceding stretched segment) were classified as *reactive feedback control*. These effects might aim at recovering relative durations within a higher prosodic timeframe (e.g., adjusting segment proportions within a syllable) or can be rather unspecific responses to a disturbance in the feedback that demands attention and time to process. Reactive feedback responses seem similar to responses elicited by generally delayed auditory feedback, where speakers were found to slow down their speech rate or prolong speech elements in response (Yates, 1963).

In our previous study, Oschkinat and Hoole (2020), response patterns differed dependent on the part of the syllable that experienced the temporal shift. While speakers compensated and, in some cases, adapted for a temporally manipulated nucleus and coda of a syllable, no significant effect was found for temporally stretched syllable onsets. The subsequent studies by Karlin et al. (2021) and Oschkinat and Hoole (2022) produced similar results regarding the responses: Speakers adjusted their productions (in absolute segment durations) for perturbed nuclei and codas, but not for onsets. We suggested that, at least for timing relations in speech, the prosodic structure of the syllable causes segments to be more or less malleable in their articulatory execution than others. This hypothesis was based on insights into the articulatory structure of the syllable elaborated in the Articulatory Phonology/Task-Dynamics framework. In modeling inter-gestural timing, the syllable segments are modeled as coupled oscillators with different coordinative relations. In some languages, such as

English or German, gestures couple mainly in-phase or anti-phase with the adjacent gestures, dependent on syllable position. Thereby, in syllable onsets, consonant gestures are coupled anti-phase with each other but in-phase with the following vowel, while in codas each gesture is coupled locally anti-phased with the preceding one. The more global coupling of onsets with the vowel constitutes a greater temporal/articulatory stability than the local anti-phase coupling of the coda segment with the vowel (Byrd, 1996; Browman and Goldstein, 2000; Goldstein and Pouplier, 2014). For detailed consideration of the evidence for differential coordination patterns related to syllable position specifically for German see Pouplier (2012). Hence, codas should be more malleable when it comes to an auditory perturbation of timing than the more articulatorily entrenched onset patterns. In our follow-up study, Oschkinat and Hoole (2022), differences in the response patterns were not only observed for different parts of the syllable, but also for syllables with different stress patterns and syllable position within the word. Both our previous studies (Oschkinat and Hoole, 2020, 2022) indicated that auditory feedback can be used for temporal corrections in the speech production process, but that prosodic structure of the perturbed segment plays a role. With regard to current speech production models, these findings support the incorporation of auditory feedback into the speech production process as modeled in the Directions into Velocities of Articulators model (DIVA model, Guenther, 2006) but for *speech timing*, combined with knowledge about the prosodic stability of segments (as elaborated in Articulatory Phonology/Task-Dynamics; cf. Browman and Goldstein, 1992).

The role of perception and the feedback system for speech acquisition and speech production has been considered crucially relevant. According to the DIVA model, speakers rely on spatio-temporal representations of speech elements (speech targets) in feedforward control. These speech targets are established via auditory and somatosensory feedback in speech acquisition (Guenther, 2016). Thereby, the size of the size of an acquired speech target is assumed to depend on individual auditory acuity and sensory error detection performance. Speakers with better auditory acuity establish smaller speech targets, resulting in more distinct productions of different speech sounds and less variability in production (Perkell et al., 2004a,b, 2008; Ghosh et al., 2010). Individual differences in auditory acuity became a further focus of interest in connection with auditory feedback perturbation studies. Villacorta et al. (2007) assessed auditory acuity in the discrimination of the first formant (F1) in vowels and set it in relation to reactions to upward and downward shifts of F1 in the same vowels. They found that the better the individual auditory acuity, the more the speaker compensated for the applied feedback alteration. This conclusion was also drawn by Brunner et al. (2011) for perturbed consonants. They found speakers with a higher auditory acuity to produce /s/ and /ʃ/ with a more distinct acoustical contrast and to use compensation strategies to a greater extent than low acuity speakers.

While individual abilities in feedback control have been considered to be crucial influencing factors in building and controlling speech targets, much less attention has been given to the thought that also feedforward mechanisms, more precisely motor execution abilities, are governed by limits of individual abilities. The study by Martin et al. (2018) investigated relations between responses to spectral auditory feedback perturbations and feedback capacities (auditory acuity), as well as general cognitive control skills as an indicator for feedforward abilities. They found auditory acuity relevant for predicting responses, but not the executive control tasks. Apart from general cognitive abilities, another aspect that could plausibly influence distinctiveness in speech production is the ability to execute motor commands for desired speech targets precisely in time and space. However, the role of temporal precision in speech production is relatively understudied. Nevertheless, a rather different strand of research has investigated temporal precision and rhythmic abilities in non-speech motor execution. An indication for the relevance of internal timing abilities in feedforward control has been provided by research on rhythmic finger tapping with or without auditory stimuli (Repp, 2005; Repp and Su, 2013; Dalla Bella et al., 2017). In typical tapping tasks, participants tap regularly at a self-chosen rate (unpaced tapping), or along with an accompanying beat or sound sequence or synchronize to music (paced tapping, Dalla Bella et al., 2017). Unpaced tapping tasks give the examiner insight into feedforward timing mechanisms and their stability in motor execution (see Drake et al., 2000). Tapping to a beat, on the other hand, tests for sensorimotor synchronization (see Repp and Su, 2013 for an overview).

A link between non-verbal sensorimotor timing abilities and speech production was found when testing finger tapping performance in non-impaired speakers and speakers with speech timing disorders. For example, Falk et al. (2015) found weaker synchronization abilities with a metronome or a musical stimulus in children and adolescents who stutter than in non-stuttering peers. Individuals who stutter showed worse rhythmic tapping performance, with a tendency to over-anticipate the pacing events, than individuals who do not stutter. Another study tested for the connection of rhythmic variability in different motor domains in patients with Parkinson's disease. Puyjarinet et al. (2019) found a link between rhythmic variability in paced finger tapping, variability in speech (oral diadochokinesis tasks), and variability in gait. They further found deficits in rhythm perception linked to deficits in rhythm production and concluded that rhythm impairments in different motor domains in patients with Parkinson's disease might be caused by an impaired central rhythm mechanism (Puyjarinet et al., 2019). Further research on speech and non-speech timing showed that in speech with finger tapping, emphasis in one domain affects the other domain as well, e.g.,

stressing a syllable is accompanied by more emphasized tapping (Parrell et al., 2014).

Altogether, these studies point toward a strong link between motor behavior in speech and non-speech actions. This link is noteworthy in particular when investigating the role of feedforward stability for timing mechanisms in fluent speech. Indeed, it can be hypothesized that temporal stability in non-speech motor behavior is connected to temporal stability in speech motor control. Thereby, it has to be taken into consideration that, domain-independently, different motor timing tasks might require different underlying neural mechanisms. Neuroscientific research outlined different such mechanisms dependent on the demand of the timing task. Grube et al. (2010) and Teki et al. (2011, 2012) distinguished between event-based timing, which occurs relative to a beat, and duration-based timing, which requires the absolute estimation of temporal intervals, both mechanisms being associated with different brain regions (Teki et al., 2011). In speech production, it is assumable that different parts of an utterance or even of a syllable follow different timing strategies. The prediction of onsets in speech, for example, was suggested to be comparable with recurrences of a musical beat (Nozaradan et al., 2012; Peelle and Davis, 2012). Further, the supposed beat in an isochronous flow of speech syllables is located in the transition between onset and vowel (p-center, Morton et al., 1976). Accordingly, onset timing might be more closely related to event-based timing mechanisms. This assumption was supported by interpreting the brain regions involved in the timing mechanisms: Event-based timing was more associated with brain regions comprising the supplementary motor area and the premotor cortex (Teki et al., 2011). Both areas were found relevant for internal planning of motor movements within a precise timing plan rather than relying on sensory information. In our previous study (Oschkinat and Hoole, 2020), we assumed greater reliance on feedforward predictions in onsets leads to less compensation in auditory feedback perturbation. Nucleus and coda of the syllable might rather be timed with underlying duration-based timing mechanisms based on a word or syllable time frame.

The previous section outlined how perceptual abilities and general motor behavior connect to speech production. Further, the introduced timing mechanisms contribute to the complexity that is assumed to underlie the planning and control of speech timing.

The main goal of the following study is to shed light on the contribution of auditory feedback and motor timing abilities to speech production. Therefore, we examine the contribution of general internal timing stability as predictor for temporal speech feedforward stability, and the importance of feedback and feedforward mechanisms for executing and planning the temporal structure of fluent speech. To follow this aim, the present study assessed individual capacities in paced and unpaced finger tapping tasks and beat-based and duration-based perception tasks, and set them in relation to behavior during temporal auditory feedback perturbation from the data collected in our previous study (Oschkinat and Hoole, 2020). In doing so, we foreground the influence of individual auditory acuity and individual motor timing stability on speech production. Thereby, we address both feedback and feedforward systems as key actors for successful speech production. As for the outcome, we have two main hypotheses.

First, concerning the contribution of perception and motor execution, we expect speakers with better perceptual abilities (auditory acuity) to compensate more for temporal auditory feedback perturbations as found analogously for spectral properties of speech. This hypothesis is based on the idea that the better an auditory mismatch is perceived, the more (precisely) speakers can counteract it. Moreover, we expect speakers with a worse performance in motor execution in finger tapping tasks (speakers with a higher motor variability) to compensate more. This hypothesis ties up with the findings of Oschkinat and Hoole (2020, 2022), where a structurally less stable system was more malleable in the face of a temporal perturbation. We expect the effect of *structural* motor stability on timing behavior to extend to *individual* abilities in motor stability, which may also shape timing mechanisms in speech.

Second, regarding the nature of responses to the auditory feedback perturbation as an online response or adaptive, we expect to find perceptual acuity equally relevant for both online reactions and adaption, since both types of reaction require the ability to perceive the auditory mismatch and identify the direction for a compensatory response in the first place. General motor stability, on the other hand, should be a greater predictor for adaptation, since a less stable feedforward system should provide a greater tolerance toward updating the less stable representations. This hypothesis is tied to expectations about the relevance of auditory feedback and motor stability for different parts of the syllable, since the coda showed adaptation while the onset did not (Oschkinat and Hoole, 2020).

Methods (procedure and data processing)

Participants

Forty-five native speakers of German performed three testing blocks (described further below) in one testing session of approximately 2.5 h. Participants were between 19 and 30 years of age (mean age: 23 years, 34 females) and received financial compensation for their participation. Musicality was assessed with a questionnaire. Thirty-two participants stated they have received musical education on various instruments. Five of them reached a semi-professional level, indicating that they could earn money with music. Musicality was not a main focus of interest in the current study. However, additional analyses about effects of musicality on the response data can be found in the

Appendix. None of the participants claimed to have any speech, voice, or hearing disorders. All of the participants started with the Auditory feedback perturbation block. After that, the order of the Tapping and Perception blocks was counterbalanced over participants. Participants were recruited in the Munich area and testing was approved by the ethics committee of the medical faculty of the Ludwig Maximilian University of Munich.

The following sections outline the three testing blocks Perturbation, Tapping, and Perception.

Temporal auditory feedback perturbation

The perturbation response data was taken from the perturbation experiment reported in Oschkinat and Hoole (2020), including the same participants and their data. The following section briefly summarizes the procedure and measures of interest. For more thorough insight into the experiment, we point the reader to the original paper.

Setup

The temporal auditory feedback experiment tested the sensitivity to temporal perturbations with a special interest in position within the syllable. In two experiment conditions, the temporal structure of either onset + vowel (*Onset condition*) or vowel + coda (*Coda condition*) of the first syllable in a three-syllabic word was temporally altered (Onset condition: /ˈ**pfa**nku:xən/, pancake; Coda condition: /ˈn**apf**ku:xən/, ring cake). Thereby, the first segment per condition was stretched in real-time (Onset condition: /pf/, Coda condition: /a/) and the following segment was compressed (Onset condition: /a/, Coda condition: /pf/) leading to an on-time signal after completion of both shifting directions. While this alteration results in perturbation being completely contained within a single syllable, it should be noted that the second segment of the perturbed part starts delayed by the amount of stretching of the first segment (plus the systems delay of approximately 25 ms that is needed for online manipulation).

Perturbations were achieved with the Audapter software package for formant and pitch shifts as well as time-warping developed by Cai et al. (2008, 2011) and Tourville et al. (2013). Participants received auditory feedback via E-A-RTone^{TM} 3A in-ear earphones with foam eartips (3M, Saint Paul, MN, United States) and spoke into a Sennheiser H74 headset microphone (Wedemark, Germany) placed three cm from the corner of the mouth. The foam eartips ensure that the manipulated feedback rather than the airborne sound is predominantly perceived. In four blocks speakers uttered the phrase "besser Pfannkuchen" (Onset condition) or "besser Napfkuchen" (Coda condition). The carrier word "besser" (*better*) allowed for an online status tracking of the signal by the Audapter software to trigger the intended part within the target word (for more information on the online status tracking please refer to Oschkinat and Hoole, 2020).

While the online status tracking triggered the perturbation from the acoustic signal of each individual trial, the duration of the perturbation section (hence the duration of the onset + vowel or vowel + coda sequence) was determined manually by the experimenter (using Praat; Boersma and Weenink, 1999) for each participant in a pretest. This pretest included 15–20 trials of the experiment without perturbation, the number of trials depending on how fast the participant established a stable speech rate. The mean duration of the intended sequence over the pretest trials was then inserted into the protocol for testing. In the testing session, perturbation was applied in blocks: The first block consisted of 20 trials without perturbation (Baseline). In the second block, perturbation increased stepwise over 30 trials (Ramp phase) followed by 30 trials with maximum perturbation (Hold phase). After that, perturbation was abruptly removed and normal feedback restored for another 30 trials (After-effect phase). **Figure 1A** visualizes the applied perturbation over the course of the experiment, **Figure 1B** depicts spectrograms of the spoken signal (H1) and the received perturbation (H2) in both perturbation conditions during the Hold phase.

Analyses

The different perturbation phases allowed the examination of compensation as a general measure of reaction to the perturbation (Hold phase), and the classification of the response as either online control (when productions revert to the Baseline immediately in the After-effect phase) or adaptive (when adjustments remained into the After-effect phase where unaltered feedback is provided). For the purposes of the current study, the mean productions in the Hold phase as compared to the Baseline per participant are examined as an indicator for the strength of reaction during maximum perturbation. This measure will then be set in relation to measures from the tapping and the perception blocks. The analysis of the After-effect phase to classify responses as an online reaction or adaptive response was performed previously in Oschkinat and Hoole (2020) and will not be considered in detail in the current study. However, the determination of the nature of responses from this earlier analysis will turn out to be of substantial relevance for the interpretation and discussion of the present study further below.

For analyses, durations of the segments of interests were segmented manually in Praat. Production differences in word-normalized durations in the Hold phase (with maximum perturbation) relative to the Baseline for each segment of interest (CC /pf/ and V /a/) in each perturbation condition (Onset and Coda condition) were examined. Accordingly, four compensation measures are considered in the following calculations: Compensation to the onset segment in the Onset condition (Onset CC), compensation to the vowel in the Onset

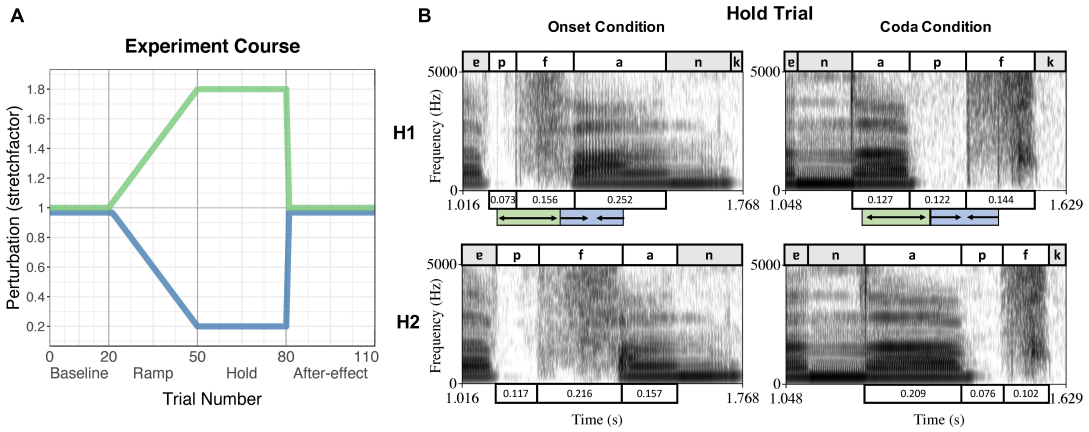


FIGURE 1
(A) Experiment course of the perturbation experiment, indicating the trial numbers and experiment phases on the *x*-axis and the stretchfactor of the perturbation on the *y*-axis. The green curve indicates the stretching of the first segment of the sequence, the blue curve indicates the compression of the second segment of the sequence. **(B)** Example of a Hold Trial per condition (Onset condition – left panel, Coda condition – right panel) produced by the same speaker. The upper panels show the spoken signal (H1), the lower panels the received perturbed auditory feedback (H2). Boxes above/underneath the spectrograms label the segments and their durations. The green/blue boxes below the upper panels indicate the stretching and compressing of the segments as triggered by the online status tracking, leading to the durations in the panels below (H2). Reproduced from Oschkinat and Hoole (2020), with the permission of the Acoustical Society of America.

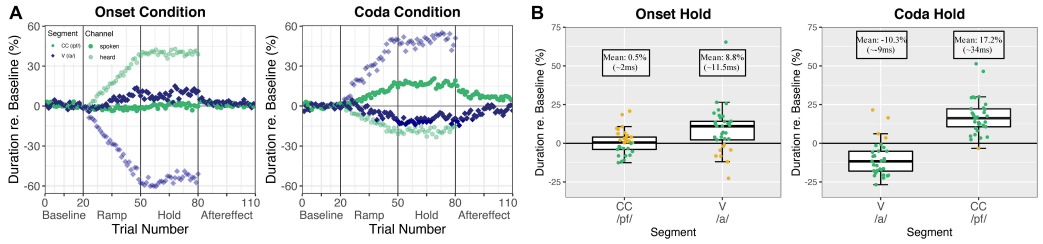


FIGURE 2
(A) Normalized relative durations averaged over all participants (*n* = 34 for the Onset condition, *n* = 33 for the Coda condition) per trial. The vowel /a/ is shown in blue rhombuses and CC /pf/ is shown in green round dots. The spoken signal is shown in solid colors and the perturbed (heard) signal is shown with higher transparency. The left panel visualizes the Onset condition and the right panel visualizes the Coda condition. **(B)** Normalized relative durations in the hold phase relative to the baseline mean (0) for vowel /a/ and CC /pf/ in the Onset condition (34 participants, left panel) and Coda condition (33 participants, right panel). Boxes correspond to the first and third quartiles and bars represent the median. Whiskers extend from the hinge to the highest/smallest value but no further than 1.5 interquartile range (IQR). Data beyond the whiskers are outliers. Individual participants are represented with colored dots where green dots mark compensatory behavior and golden dots mark a following of the perturbation direction. Reproduced from Oschkinat and Hoole (2020), with the permission of the Acoustical Society of America.

condition (Onset V), compensation to the vowel in the Coda condition (Coda V), and compensation to the coda segment in the Coda condition (Coda CC). These measures are the same as reported in Oschkinat and Hoole (2020) and will further be referred to as Onset CC, Onset V, Coda V, and Coda CC. The order represents the chronological appearance of the segments within the condition and consequently also represents the structure of the perturbation section, whereby the first segment per condition is stretched and the second segment is compressed.

Figure 2A shows the produced and heard (perturbed) durations per segment and experiment condition over the course of the experiment. **Figure 2B** summarizes the production difference in the Hold phase as compared to baseline productions per segment of interest.

The perturbation data per participant and trial were scanned for correct triggering of the perturbation section at the intended part within the utterance. Since the online status tracking is based on previously determined intensity and duration thresholds, in some cases due to inter- and intra-speaker variability in speaking rate and style, the perturbation did not cover the intended speech sequence. Single trials were removed from calculations when perturbation did not cover the onset + vowel or vowel + coda section as intended. Participants who had less than 16 trials in the hold phase with accurate location of the perturbation were removed from

further calculations. Linear regressions confirmed no effect of number of trials on compensation magnitude after removing the failure trials. Based on this exclusion criterion, 34 participants remained in the Onset condition and 33 participants in the Coda condition (as reported in Oschkinat and Hoole, 2020).

Responses to temporally perturbed auditory feedback

The boxplots in **Figure 2B** (and the corresponding analyses performed in Oschkinat and Hoole, 2020) indicate that in the Onset condition, speakers did not change productions of the CC segment, but compensated for the vowel perturbation by lengthening the vowel in production. In the Coda condition, speakers compensated for the vowel perturbation by shortening the vowel, and for the Coda CC perturbation by lengthening Coda CC in production. The transitions from the Hold to the After-effect phase in **Figure 2A** indicate that in the Onset condition, the lengthening of the vowel (left panel, blue solid dots) was mainly an online response, as reactions revert to baseline shortly after perturbation is removed in the After-effect phase. The responses in the Coda condition (right panel, solid dots), on the other hand, were both indicative of adaptive behavior, visible in continuing adjustive responses from the Hold to the After-effect phase for both segments of interest (for statistics please refer to Oschkinat and Hoole, 2020).

The temporal adjustments in the Hold phase relative to the baseline mean (visualized in **Figure 2B**) will be taken as the measure for *compensation* to the perturbation per condition and segment. Note that the Onset CC and the Coda V were stretched in perturbation so that an opposing reaction is indicated by a shortening of productions (negative estimates relative to the baseline mean, see **Figure 2**). Hence, an opposing response to Onset CC and Coda V is necessarily adaptive. For the analyses in the current study, the values of the Onset CC responses and the Coda V responses were multiplied by –1. Thus, an opposing response is always indicated by a positive value and following the perturbation direction by a negative value.

Before turning to the motor and perceptual tests that will be related to the perturbation response patterns, we introduce here some brief analysis not included in the original Oschkinat and Hoole (2020) paper. Its purpose is to give additional preliminary motivation that consideration of individual behavior patterns should be fruitful by examining linear relationships between the four compensation measures introduced above. Conceptually, it would belong better with the motivations for the current investigation considered in the introduction, but can only be succinctly presented now that the reader has been given detailed information on the design of the previous experiment. Linear models were calculated between the four compensation measures (Onset CC, Onset V, Coda V, and Coda CC, with the outcome visualized in **Table 1**). The analyses revealed a significant linear relationship between Coda V and Coda CC (adjusted R-squared = 0.10, df = 31, p = 0.04), revealing that speakers who compensated more for the vowel perturbation in the Coda condition (by shortening it in production) also compensated somewhat more for the Coda CC segment (by lengthening it in production). Regarding our hypothesis, we assume that for these segments, which both showed adaptive behavior, a certain level of motor malleability is given in speakers that compensate and adapt more. However, since both segments appeared within the same word/trial, the magnitude of perturbation might have contributed to equally strong within-participant responses. A relationship was also found between Onset V and Coda CC responses, which both were second segments in the perturbation section, hence compressed in the auditory feedback but lengthened in production (adjusted R-squared = 0.196, df = 25, p = 0.01). Further, both segments were displaced in time, due to the stretch of the first segment and assumable subject to online control effects triggered by the stretch of the previous segment. The relationship supports the hypothesis that there might be an individual auditory sensitivity in speakers to react to effects of delayed/shifted auditory feedback in the online control. Both relations reinforce the aim of the current study to find individual motor or auditory abilities that enhance or decrease articulatory timing malleability in the face of an auditory perturbation, and more generally in the speech production process.

Tapping battery

For the tapping test block, participants were seated in front of a Roland SPD-6 MIDI percussion pad linked via a Midi-Interface (Miditech, midiface, 4 × 4) to a computer controlled by MAX-MSP software (version 6.0). Loudspeakers delivered sound stimuli in free field at a fixed volume which was kept constant over participants. The experimenter instructed the participant to tap with their writing hand's index finger on the tapping pad. Practice trials preceded each of the following tasks, which could optionally be skipped when the following task was very similar to the preceding one. Tasks 1, 2, and 3 are adopted from the *Battery for the Assessment of Auditory Sensorimotor and Timing Abilities* (BAASTA, Dalla Bella et al., 2017). Tasks 4, 5, and 6 contain speech stimuli of different complexity implemented for this study's particular purposes. All tapping tasks required the participants to tap as regularly as possible without intended variation in inter-beat interval or tempo. Except for the unpaced tapping task, all tasks differed in stimulus and inter-onset-interval (IOI, or inter-beat-interval, IBI, in music stimuli) of the respective stimulus beat. Since not much is known about the connection between finger tapping tasks and responses to temporal auditory feedback perturbation, a spectrum of different tests should provide insight into the connection between motor execution performance in different rhythmic contexts. The unpaced tapping task (Task 1) captures internal timing mechanisms. The metronome tapping tasks

(Task 2) test for synchronization with a stimulus comprising a sequence of tones (metronome), the music tasks (Task 3) and speech tasks (Tasks 4–6) test beat detection in more complex stimuli. Thereby, the sentence tapping and music tapping require an identification of the beat in continuous sound flow, the syllable and wordlist tapping contain silence between each beat (here: word or syllable) analogously to the metronome tasks. The two music tasks differ in complexity. These tasks might provide further insight into timing abilities in different domains (speech/music). The single tapping tasks are summarized in **Table 2**, which gives a short explanation of the stimuli and the tempi performed by each participant.

All tapping data were pre-processed following the procedures as reported in Dalla Bella et al. (2017). The first ten taps were discarded in all tapping tasks, and artifacts (inter-tap intervals below 100 ms) and outliers were removed. For all tasks, including the unpaced tapping task, the mean inter-tap-interval (ITI) was calculated, and the coefficient of variation of the ITI (cv of the ITI, namely, the ratio of the standard deviation of the ITIs over the mean ITI) was taken as a measure for *motor variability*.

Perception tasks

The third block tested for individual perceptual abilities. Five adaptive staircase tasks assessed individual auditory acuity performances for temporal properties of various stimulus types. The listener was seated in front of a computer and provided with headphones. Volume was set to a comfortable level as tested and determined by the experimenter and was not changed between listeners unless requested. After the experimenter started the procedure in MATLAB, listeners performed the tasks by entering their responses directly into the testing computer. The first three staircase tasks captured duration discrimination abilities (hence duration-based timing mechanisms) in a 2-interval 2-alternative forced-choice paradigm. These tasks required judgments about the two perceived stimuli as identical or different. Task 1 required judgments about pure tone durations. Tasks two and three comprised monosyllabic words with temporal manipulations analogous to the auditory feedback perturbation paradigm described in section "Temporal auditory feedback perturbation." In the onset perception task, the onset of

TABLE 1 Correlation table providing the adjusted *R*-squared and *p*-values for the relationships of the four compensation measures (compensation in the hold phase relative to baseline for Onset CC, Onset V, Coda V, Coda CC).

	Onset_V		Coda_V		Coda_CC	
	Adj. R^2	*P*-value	Adj. R^2	*P*-value	Adj. R^2	*P*-value
Onset_CC	–0.026	0.658	0.051	0.135	–0.024	0.542
Onset_V			–0.04	0.99	**0.196**	**0.012**
Coda_V					**0.101**	**0.040**

Significant relations in bold.

TABLE 2 Overview of the performed finger tapping tasks.

Task Name	Tempo (ms)	Task/Explanation
(1) Unpaced	free	Regular Tapping for 60 s at a self-chosen tempo.
(2) Metronome	IOI 600, IOI 750 IOI 900	Tapping to a metronome (i.e., a sequence of tones with a frequency of 1319 Hz) for 60 s per tempo.
(3) Music	Rossini: IBI 600, Bach: IBI 600	Tapping to piano midi stimuli created from well-formed (regular) excerpts of the beginning of Bach's *Badinerie* and Rossini's *Wilhelm Tell*.
(4) Syllable	IOI 750	Tapping to the syllable "bla." Four instances of the syllable "bla," uttered by a German female speaker were randomly concatenated for 45 s with the IOI measured between the syllables' supposed p-centers (determined using the algorithm from Cummins and Port, 1998).
(5) Wordlist	IOI 900	Tapping to a spoken wordlist (recorded by a female German speaker) of real monosyllabic words (nouns and adjectives) with complex onsets [CCV(C)]. Words were concatenated for 55 s with the IOI measured from the supposed p-centers (Cummins and Port, 1998).
(6) Sentence	IOI 600	Tapping to short sentences for 45 s (arranged from stimuli taken from Falk et al., 2017), repeated three times. Sentences presented a regular alternating rhythm (one unstressed – one stressed syllable) with an inter-stress-interval of 600 ms measured from the supposed p-center of each stressed syllable suggesting tapping on every second syllable.

a word was stretched and the vowel compressed. In the coda perception task, the vowel was stretched and the coda compressed (see **Table 3** for details). With these tasks an opportunity was provided to measure individual perceptual acuity of manipulated sound durations within a syllable similar to the auditory feedback perturbation. In addition to the three discrimination tasks, two beat-alignment tasks (BAT) related to the sentence and music tasks (Tasks 5 and 6) of the tapping battery in section "Tapping battery" were performed. Tasks 4 and 5 required beat-alignment judgments (and therefore event-based timing mechanisms) in a 1-interval 2-alternative forced-choice paradigm. The decision required a binary judgment on whether the metronome superimposed onto the auditory stimulus was aligned with the accents/beats of the speech or music stimuli or whether it was regular but shifted away from the natural accent/beat.

For all five tasks, continua of stimuli between two endpoints were generated, whereby one endpoint consisted of the original stimulus and the second endpoint was a manipulated version. The manipulations target durations exclusively. For the three duration discrimination tasks (1–3), stimuli were presented in pairs (2-interval) in which one stimulus was always the original stimulus and the other stimulus varied in degree of manipulation between the two endpoints. Manipulations of segment durations were performed in Praat using *psola* (see task-specific description below). The presented stimuli were randomized, whereby the original stimulus was either in the first or in the second position. In the two beat-alignment tasks (4 and 5), one endpoint of the continuum was a stimulus with perfect beat-alignment, and the other endpoint a stimulus with the maximally shifted beat. In these tasks, always one stimulus from the continuum was presented while the degree of shift in alignment varied along the continuum. The difference between the two presented stimuli in Tasks 1, 2, and 3, or the degree of metronome shift in Tasks 4 and 5 is referred to as *delta* (in ms). In each task, the delta

TABLE 3 Overview of the performed perception tasks.

Task Name	Stimuli/Continuum	Design/Task Question
(1) Pure Tone	Stimulus: Two pure tones (frequency: 333.3 Hz) Continuum endpoints: (1) Tone duration of 600 ms (2) Tone duration of 1200 ms	Design: 2-interval 2-alternative forced choice duration discrimination task Question: Do both tones have the same duration or not?
(2) Onset	Stimulus: Monosyllabic CVC word Continuum endpoints: (1) "Schaf" (/ʃa:f/, sheep) and (2) "Schaf" manipulated, with /ʃ/ stretched by 200 ms and the following /a:/ compressed by 200 ms	Design: 2-interval-2-alternative forced choice duration discrimination task Question: Are the two words identical or different?
(3) Coda	Stimulus: Monosyllabic CVC word Continuum endpoints: (1) "Gas" (/ga:s/, gas) (2) "Gas" manipulated, with /a:/ stretched by 150 ms and the following /s/ compressed by 150 ms	Design: 2-interval-2-alternative forced choice duration discrimination task Question: Are the two words identical or different?
(4) Speech	Stimulus: Sentence with a metronome beat on every second (hence on every stressed) syllable at a stable tempo of 600 ms inter-beat-interval repeated three times (stimuli from Falk et al., 2017) Continuum endpoints: (1) Well-aligned metronome beat on every stressed syllable based on the p-center algorithm from Cummins and Port (1998) (2) Misplacement of the metronome beat by shifting it 200 ms later than in the original stimulus	Design: 1-interval-2-alternative forced choice beat-alignment task Question: Does the metronome match the stimulus or not?
(5) Music	Stimulus: Midi excerpt of Bach's Badinerie (taken from Dalla Bella et al., 2017) Continuum endpoints: (1) Perfectly aligned metronome beat (2) Misplacement of the metronome beat by shifting it 200 ms later than in the original stimulus	Design: 1-interval-2-alternative forced choice beat-alignment task Question: Does the metronome match the stimulus or not?

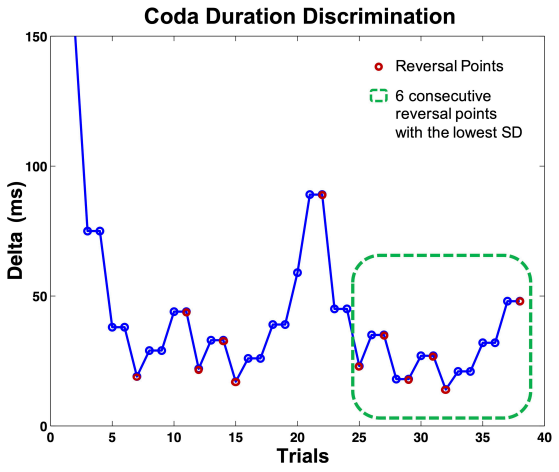


FIGURE 3
Visualization of the course of a perception staircase test (here: coda duration discrimination task). Circles indicate one presented stimuli pair. If the listener identified a pair correctly, the same pair was presented again, whereby the order of the two stimuli was randomized. After responding to the second presentation correctly, the delta of the following stimuli pair dropped; with every (single) wrong answer, the delta increased. The test ended after 12 reversals; each reversal point marked with a magenta circle. The green dashed square indicates the six consecutive reversal points (counting on from the reversal point at trial 25) with the lowest SD; the mean delta of these trials served as an auditory acuity score for further calculations. Catch-trials are not included in the figure.

could be varied in increments of 1 ms. The maximum/start delta defines the largest difference between the stimuli or between the metronome and the stimulus. Estimations of the lowest correctly identified delta (based on calculations described further below) will serve to measure each participant's individual auditory acuity.

All five staircase tasks had fixed but adaptive step intervals. In other terms, the next presented stimulus pair was triggered by the listener's response. The tasks always required two correct difference detections in a row to the same stimuli pair to mark a successful identification, but only one incorrect response to mark a false identification (2 down/1 up protocol). The two stimuli in both presented trials appeared in random order. Following a successful mismatch identification, the current delta was multiplied by 0.5; with every not detected mismatch, the delta was multiplied by 1.5 (see **Figure 3**). Whenever there was a change of response quality (successful identification/false identification), one reversal was counted (see **Figure 3**). Each task ended when a fixed number of reversals was reached (12 reversals for the discrimination tasks 1–3, eight reversals for the beat-alignment tasks 4 and 5). Each task contained four to six presentations of two identical stimuli or a perfectly aligned beat (*catch-trials*). Listeners who did not identify more than 50% of the presented catch-trials correctly or did not reach a score below 70% of the start delta of a test were suspected of answering by chance or classified as incapable of performing the task. However, none of the participants had to be excluded from further calculations based on these criteria. To measure individual performance, for every participant and task, an individual auditory acuity score (i.e., a differential threshold) was assessed by calculating a mean over the delta of the most stable consecutive six reversal points in each task (the six reversals with the lowest SD, indicating a stable pattern of response), visualized in **Figure 3**. The most stable consecutive six reversals per test were chosen analogously to the procedure in Brunner et al. (2011).

Analyses

The data provided by the Tapping and Perception blocks capture many motor and perceptual abilities for rhythmic speech and non-speech tasks. The following sections aim at summarizing the data, since the single tasks per block are expected to be highly collinear. In perception, the expected underlying mechanisms of the tasks as event-based or duration-based timing provide motivation to group tasks together based on timing mechanisms. However, it is also possible that dividing the tasks into speech and non-speech tasks reduces collinearity to a greater extent and explains the largest variance in the dataset. Further, in tapping, unpaced tapping can be accomplished successfully by engaging either beat-based mechanisms (i.e., generating an underlying pulse) or duration-based mechanisms (i.e., repetition of a single interval, Teki et al., 2011). Therefore, to reduce the complexity of the measures and find the most important dimensions in the data, we conduct one manual split: The unpaced tapping task will be treated as a separate measure, since it is the only tapping task that gives insight into pure feedforward timing mechanisms against eight paced tapping tasks. Further than that, principal component analyses for the (remaining) tapping and perception tasks are conducted. Summarizing the individual tasks per block into principal components should provide a mean performance measure per participant over tests that are highly correlative. In doing so, the responsibility for grouping the data in a meaningful way is passed on to the PCA. If there are substantial differences between the single tasks per block, they are expected to turn out as different underlying dimensions in the principal component analysis.

Principal component analysis

The perception and tapping data were submitted to principal component analysis (PCA) using R's *mclust* package (Scrucca et al., 2016). The PCA reduces the number of independent variables to single components by extracting the

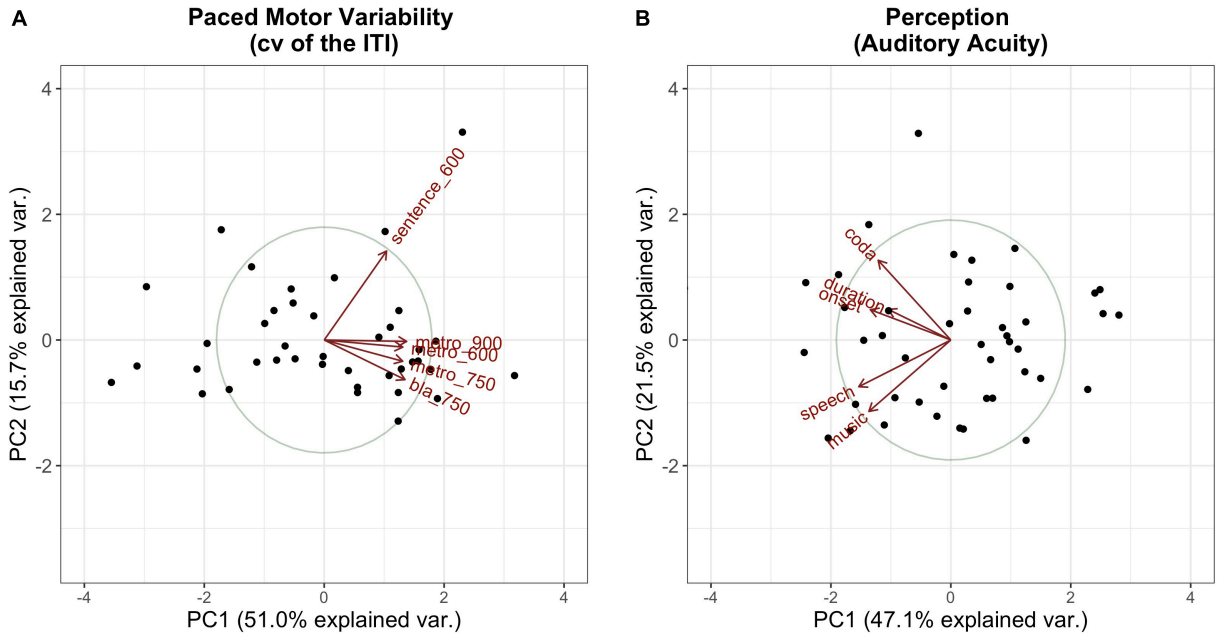


FIGURE 4
Visualization of the first (*x*-axis) and the second (*y*-axis) principal components along with their amount of explained variance per PCA (**A**: Tapping PCA, **B**: Perception PCA). Dots mark the single participants; vectors represent the factor loadings of the single tasks on each of the components. Tasks are abbreviated by type (e.g., "metro" for metronome tapping) followed by the IOI (if relevant).

underlying dimension for variables that highly correlate with each other. The extracted underlying dimensions (principal components) of a PCA do not correlate with each other and describe the dataset's maximum variance. In the following, the main components are extracted and used for further calculations. PCAs were calculated separately for all perceptual tasks (auditory acuity measures) and for all the tapping tasks (motor variability measures), except for the unpaced tapping task. The unpaced tapping task differed from the other tapping tasks in modality, as it was the only task without a pacing event. It gives insight into intrinsic motor timing without a guiding stimulus. Motor variability of the unpaced tapping task was individually taken into account in addition to the principal components. Unpaced tapping Motor variability was *z*-normalized before calculations. With this data partitioning, we hoped to keep one general underlying dimension per measure of interest (Perception, paced Tapping, unpaced Tapping) and further expected the PCA to give more insight into the nuances of the individual tasks.

Data pre-processing

As PCAs cannot deal with incomplete data, participants with no data in more than one out of the nine tapping tasks (including unpaced tapping) were not submitted to the PCA. Based on this criterion, five of 45 participants were removed before submitting the data to the PCA. For those participants who had missing data in one task, the missing value was filled with the k-nearest-neighbor imputation method (knn-imputation, Beretta and Santaniello, 2016). For one participant a missing value was imputed in the music Badinerie tapping task, for another participant in the metronome IOI 900 tapping task. Although not submitted to the PCA, this procedure was also applied to the unpaced tapping task. Two missing values in unpaced tapping were filled per knn-imputation.

Before submitting the paced tapping tasks to the PCA, the Kaiser–Meyer–Olkin measure (Kaiser, 1970) verified the measure sampling adequacy (MSA) overall per measure block and single task. The measure represents the *ratio of the squared correlation* between the single tasks to the *squared partial correlation* between the tasks. An MSA value of 0 indicates that the sum of partial correlations is large relative to the sum of correlations, suggesting too much diffusion in the data for factor analysis/PCA. An MSA of 1 indicates that the patterns of correlation are compact, indicating that the factors can distinguish the data reliably (Field et al., 2012, p. 769). An MSA value above 0.5 qualified the single measures for submitting them to the PCA, and the overall MSA measure classified the whole task block as suited for PCA if the overall MSA was > 0.5. In tapping motor variability, overall MSA was 0.68. However, the single MSAs for the Badinerie and Rossini music tapping tasks and the wordlist tapping task were < 0.5 and hence not submitted to the PCA. One could potentially leave these tasks aside and analyze them separately. However, this

study aimed for an approach that optimized the coherence of the study as a whole, so these tasks were excluded from further calculations.

The same outlier and knn-imputation treatment as for the tapping data was applied to the perception data. However, no participant had missing values in the perception data. Hence no participant was removed and no data filled with knn-imputation. For the perception tasks, the overall MSA was 0.58 and all five tasks were kept in the PCA. Values were centered and scaled when submitted to the PCA.

The PCs that explained the most variance, whereby a substantial amount of explained variance is given for PCs with an eigenvalue > 1 ("Kaiser criterion," Kaiser, 1960), were kept for further calculations. Those comprised the first principal component for the Tapping PCA and the first two principal components for the Perception PCA. Hence, PC2 in Tapping, which separates the sentence tapping from the metronome and syllable tapping, was dropped for further calculations. **Figure 4** visualizes the first two components for each of the measures of interest.

Interpreting PC scores

Tables 4, **5** summarize the factor loadings of PC1 (and PC2) per PCA for each of the single tasks. Variables that have a larger loading than would be the case if all variables contributed equally, namely square root of 1 divided by the number of variables, will be regarded as important contributors to the respective principal component. For the Tapping PCA (**Table 4**), all presented tasks show fairly similar loadings on PC1. PC1 is therefore interpreted as a general measure for paced tapping motor variability. Correlation plots indicated the relation between the raw values submitted to the PCA and the PCs provided by the PCA. For *PC1 Paced Motor Variability*, better performances, meaning low motor variability values, are associated with lower PC1-scores. The same directionality applies to *Unpaced Tapping Motor Variability*, whereby lower values indicate low motor variability (hence a better performance).

For the Perception PCs, all of the tasks correlate negatively with PC1, with a higher PC score indicating a better perception (meaning a lower auditory acuity threshold). Therefore, PC1 reflects general auditory acuity and will further be referred to as *PC1 Auditory Acuity*. The music BAT perception task correlates negatively with PC2, as does the speech BAT task, although less intensely. The coda discrimination task correlates positively with PC2, and so do the onset and duration discrimination tasks, but less intensely. This clustering indicates that PC2 distinguishes beat-alignment judgments (event-based timing mechanisms) in speech and music from duration discrimination (duration-based timing mechanisms) and will further be named *PC2 BAT Perception*. Hereby, higher (positive) PC-scores are associated with better beat-alignment perception (especially in music) but worse duration discrimination (especially in the coda task). In view of this interpretation of PC2 BAT Perception, we expect this measure to be more closely connected to compensation in onsets, since we previously assumed onsets to rely more on event-based timing mechanisms. Duration-based perception tasks might be more closely coupled with compensation in the coda, since coda timing requires more likely absolute estimations of the time-lag from the onsets. Concerning the vowels, we have no precise hypothesis, since in the transition from onset to the vowel the p-center might play a role. The end of the vowel, on the other hand, might be more likely estimated with an absolute (duration-based) timing mechanism. The perturbation data and predictors provided by the Tapping and Perception blocks are summarized in **Table 6**.

TABLE 4 Factor loadings for each of the tapping tasks on PC1 for Motor Variability.

Task	PC1 paced motor variability
metro_600	0.46
metro_750	0.46
music_badine	–
music_ross	–
metro_900	0.48
bla_750	0.47
sentence_600	0.36
wordlist_900	–

High factor loadings on a component (|value| > 0.45) are shaded in grey.

TABLE 5 Factor loadings for each of the perception tasks on the PCs for the Perception PCA.

Task	PC1 perception	PC2 perception
Onset	−0.46	0.24
Coda	−0.41	0.64
Duration	−0.36	0.25
BAT speech	−0.52	−0.38
BAT music	−0.47	−0.57

High factor loadings on a component (|value| > 0.45) are shaded in gray.

Outlier treatment (summary)

From the complete set of 45 participants in the beginning, the predominant basis for participant exclusion came from the perturbation data. Thirty-four participants remained in the Onset condition and 33 in the Coda condition after scanning the data for correct triggering of the perturbation (see section "Analyses"). The full set of participants was submitted to the PCAs for Tapping and Perception to get more reliable scores based on a larger dataset. Excluded from the full set were

TABLE 6 Overview of the measures of each of the three testing blocks along with the interpretation of the single PCs from the principal component analyses.

Test block	Quality	Measure 1	Measure 2
Perception	Auditory Acuity	PC1: Auditory Acuity	PC2: Beat-alignment (BAT) Perception
Tapping	Motor Variability	PC1: Paced Motor Variability	Unpaced Motor Variability
Perturbation	Onset compensation	Onset CC compensation	Onset V compensation
	Coda compensation	Coda V compensation	Coda CC compensation

The four perturbation measures (shaded in gray) will, due to their difference in articulation, position within the syllable, and perturbation direction, always be treated as different dependent variables. Measures 1 and 2 from Tapping and Perception will serve as predictors in model fitting.

participants who had missing values in more than one of the tasks (five participants in Tapping, none for Perception), while data was imputed in the Tapping block for missing data in maximally one task per participant (applying to two participants). Data was also imputed for the Unpaced Tapping task (for two participants).

After calculating the PCAs, outliers of the generated principal components and the unpaced tapping task (data outside the 95% confidence intervals) were removed to reduce noise in the data. The same outlier treatment was applied to the perturbation response data.

The data were scanned for missing values based on outlier exclusion in the four Tapping/Perception measures (PC1 Auditory Acuity, PC1 BAT Perception, PC1 Paced Motor Variability, Unpaced Motor Variability). No participant had more than one missing value in the data. The single missing values were replaced with knn-imputation as performed on the raw data prior to the PCA (one participant: PC1 Auditory Acuity, two participants: Unpaced Motor Variability). Since the perturbation response data serve as the dependent variable in this study, none of the missing values (excluded participants as described in section "Temporal auditory feedback perturbation" and outliers) were imputed for the perturbation measures. The data was then divided into four datasets, each comprising one perturbation measure as the dependent variable (Onset CC, Onset V, Coda V, Coda CC) and the four Tapping/Perception measures as predictors (see **Table 6** for an overview of measures).

After data exclusion and imputation, the remaining data comprised 28 participants for Onset CC perturbation, 29 for Onset V perturbation, 28 for Coda V perturbation, and 26 for Coda CC perturbation. Note that in visual presentation, outliers and imputed data after calculation of the PCAs are included but marked as such.

Statistical modeling and results

The preceding section prepared the data for the following statistical analyses. These aim at testing our two main hypotheses that we firstly expect better auditory acuity and higher motor variability to be connected with more compensation. Secondly, we expect the perception measures to be more relevant for predicting effects in the online control, present in the Onset V and Coda CC, which were the second perturbed segments per condition. Further, we expect motor variability to be more connected with segments that were adapted for due to the perturbation, which was found in both segments in the Coda condition (Coda V and Coda CC). After interpreting the PCs in section "Interpreting PC scores," we further assumed the PC2 BAT Perception to be more related to Onset CC compensation, since syllable onset timing in speech production has been suggested to rely on event-based timing mechanisms. However, since we did not find a significant effect of compensation for the Onset CC in the first place, this effect might not show in the analyses. We do not have a precise hypothesis regarding timing mechanisms of the tapping tasks, since the distinction is less pronounced than for the perception block.

To examine the most relevant predictors for responses per perturbed segment (CC or V) and perturbation condition (Onset vs. Coda condition), we make use of a machine learning technique by fitting regression trees to the data. Regression trees should provide insight into the most relevant predictors for splitting the data. Subsequently, linear models are fitted to the data with the predictors provided by the regression trees including their interactions to examine how well these predictors describe the variance in the data. In doing so, the regression tree analysis can be seen as an exploratory approach used for describing the most prominent qualitative features in the data. The linear models are then used as a confirmatory analysis to assess the robustness of the subdivisions into groups and provide the explained variance and statistical significance.

Regression trees

Classification and Regression trees (CART, Breiman et al., 1984) are forms of decision trees that divide a dataset into further subgroups based on given discrete (classification) or continuous (regression) predictors. For each (sub)group, a

simple model is fitted to predict the average outcome of the dependent variable. CART models are represented as a binary tree, whereby at each split one predictor is chosen by an algorithm detecting the least modeling error. For the purpose of our study, we adapted the method to process a relatively small dataset ($n \approx 29$) to extract the most salient splitting criteria (predictors) of our data and to further model linear relationships between the predictors and the response data as suggested by the tree. Regression trees were fitted with the *rpart* function using the *rpart* package by Therneau and Atkinson (2019). Trees were visualized with the *rpart.plot* function/package (Milborrow, 2021). With this approach, the most descriptive of the four given predictors were extracted, and overfitting of the data avoided, which could occur due to the large number of predictors compared to the number of observations. The *rpart* function applies automated 10-fold cross-validation when choosing the best splitting predictor at each splitting point (*node*) and therefore reduces the risk of overfitting the data. For each node, the variable and a threshold along this variable are chosen to reduce the variance in the child nodes. At each splitting point, the variables are scanned for the error between the predicted and the measured values, and squared to get the sum of squared errors (SSE). The lowest SSE defines the splitting variable and the splitting point within the variable. This approach has some similarities with previous studies in which participants were split into groups of better performers vs. worse performers for a given variable. For example, Ghosh et al. (2010) divided their participants into "low- and high-acuity groups," based on their median for auditory and somatosensory acuity, as previously done in Perkell et al. (2004b). In contrast to a traditional median split, the tree in the CART procedure first helps to decide on the best variable for dividing the data into groups, and crucially also lets the function choose the best threshold for splitting the participants along this selected predictor.

Four regression trees were fitted (one to each dataset) with the perturbation response data (compensation) as the dependent variable and PC1 Auditory Acuity, PC2 BAT Perception, PC1 Paced Motor Variability, and Unpaced Motor Variability as predictors, setting rpart's *method* parameter to *anova*. The minimal number of participants for each split was set to four (*minsplit*), including the final splits (*minbucket*). Further, a cost complexity parameter (*cp*) has to be set to define the complexity of the tree. The *cp* decisively shapes the complexity of the tree and is a tuning parameter that should provide the best tree for predicting future data (balancing over- and underfitting of the tree model). A cp of 0 fits the most complex tree by predicting each observation (overfitting); a large *cp* might reduce the cross-validation error but increases the relative error and might underfit the data. To avoid underfitting of the small dataset, the *cp* per model was chosen based on the cross-validation error, but with a slightly greater tolerance toward a greater cross-validation error than in machine learning approaches. This approach was further motivated by the fact that in our case finding relationships in the data is of higher priority than actually predicting future data. Overfitting, in turn, was avoided by suppressing recursive splits of the same predictors that lead to a non-linear relation between the predictor and the outcome variable (*cp* per model: Onset CC: 0.1, Onset V: 0.15, Coda V: 0.15, Coda CC 0.1).[1]

The four tree models are presented in **Figure 5**. Per tree, each blue box shows an average of the outcome variable (compensation) and the number of participants that fall into this category/split. Below the first box, the variable is presented that first splits the data into two subgroups, along with the estimated threshold that splits the data along this variable. Participants who fulfill this criterion (dependent on the operator are above or below this threshold) are assigned to the branch on the left side ("yes"), those who do not are assigned to the branch on the right side ("no").

Interpreting regression trees

The quality of the predictions will be examined in turn for each of the four perturbed segments by looking at the tree models individually, supported by a more detailed visualization of each tree in **Figure 5** using the strip plots on the right. While the upper strip(s) visualize partitioned data of the tree, the lowest strip visualizes the overlap of the groups by including all participants along the compensation scale. In particular, quite an informative impression of the prediction quality can be gained by looking at the extent to which the groups defined by the regression tree overlap.

The tree for Onset CC compensation shows the average amount of compensation (0.9%) for 28 participants in the blue box on the top (**Figure 5A**). The first splitting parameter is PC2 BAT Perception; this splits the dataset into participants with a PC2 BAT Perception above 0.77 (right branch), who on average have a compensation value of 6.5%, this applying to seven participants. Those who have a PC2 BAT Perception score lower than 0.77 have a mean compensation value of -1%, which accounts for 21 participants. For these 21 participants, PC1 Auditory Acuity was chosen as the most important parameter to further split the data, whereby participants with a PC1 Auditory Acuity score higher than 0.77 had larger compensation values (mean 2.9%; seven participants) and those with a score below 0.77 compensated less, or even more likely followed the perturbation (mean −3%, 14 participants).

1 While we believe that non-linear relationships between auditory acuity and compensation, and particularly between motor variability and compensation are readily conceivable, substantially larger numbers of observations would be needed to fit and assess such models.

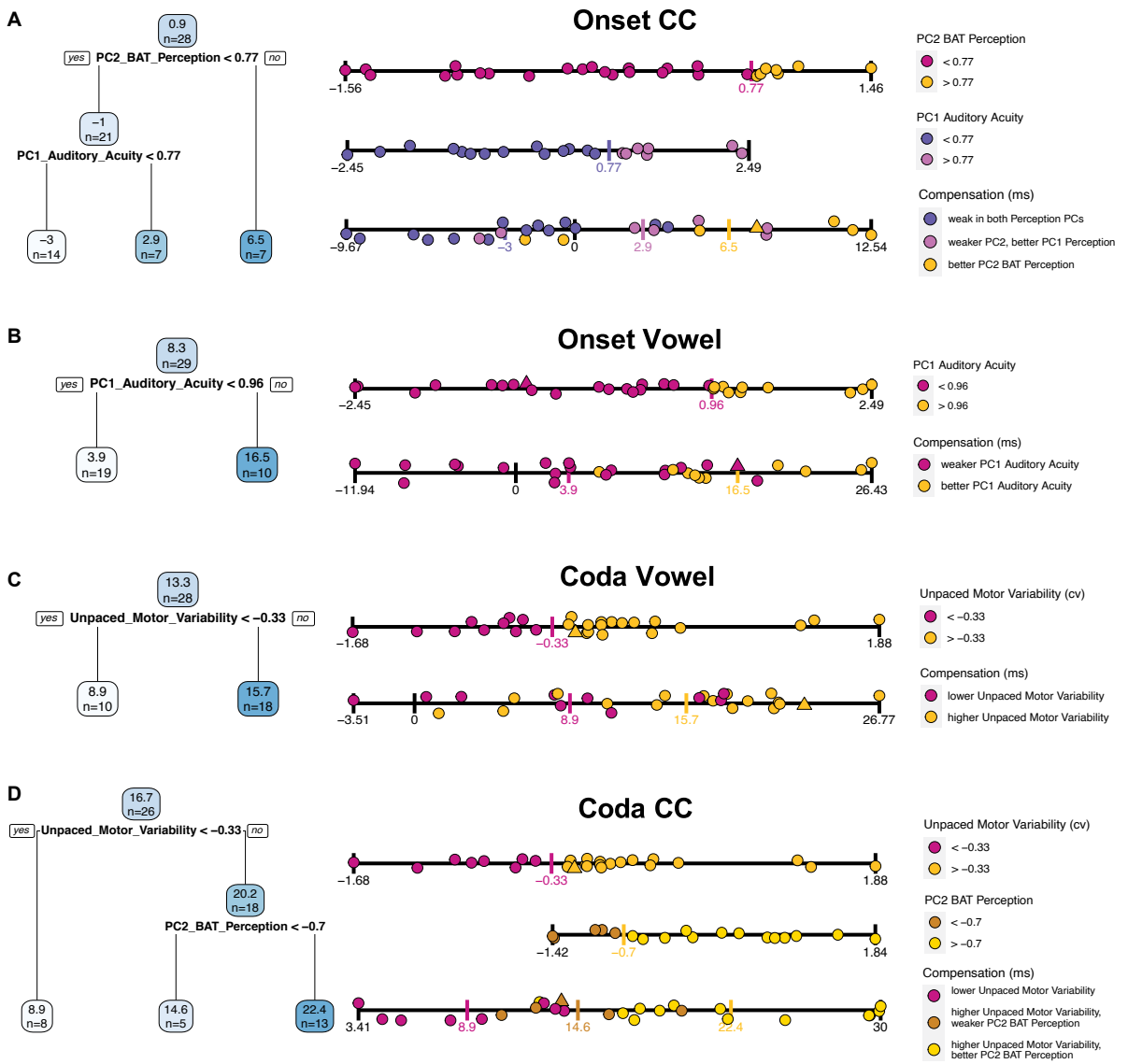


FIGURE 5
Fitted regression trees for each of the individual compensation measures in blue on the left side [from top to bottom: **(A)** Onset CC, **(B)** Onset Vowel, **(C)** Coda Vowel, **(D)** Coda CC]. Blue boxes are nodes and show predicted average compensation values at each splitting point, followed by the number of included participants. Darker blue boxes indicate higher compensation values, lower compensation values in lighter color. Below the blue boxes the variable that splits the data, along with an operator (> or <) and the estimated threshold for splitting. Participants who fulfill this criterion are included into the left branch ("yes"), participants who do not are assigned to the right branch ("no"). On the right next to the regression trees a further visualization of the splitting includes each participant's performance. Each strip displays the variable according to the legend on the right next to it. The upper strips are splitting predictors; single participants are colored based on the thresholds as provided by the tree (threshold marked with a colored vertical tick). The lowest strip per plot displays compensation values, including the participants color-coded by the splitting thresholds per predictor variable. Triangles are imputed data for the respective predictor (upper strips), or of at least one of the predictors in the lowest (compensation) strip. Recall that lower motor variability corresponds to better performance in finger tapping.

In summary, participants with better perception of beat-alignment compensated more. Those who were less good at perception of beat-alignment but better in general auditory acuity compensated a bit less, and those who showed weak abilities in both Perception PCs compensated least, or followed the perturbation direction. Recall here that compensation necessarily means the response is adaptive, since an opposing response is realized by shortening the Onset CC in production. A following of the perturbation direction due to weak perceptual skills could be explained by the lack of precisely detecting the direction of the auditory shift or determining the directionality of a response that would counteract the perceived shift. The relationships are further visualized on the right-hand side of Figure 5. In Figure 5A, the third strip plot (compensation)

allows the overlap of the strongest (yellow dots) and weakest (slateblue dots) compensators to be visualized: The group of high performers in perception of beat-alignment (yellow dots) shows almost no overlap with the group of low performers in perception of beat-alignment and with low auditory acuity (slateblue dots, lowest strip), indicating a good prediction of compensation based on the two Perception PCs.

For vowel compensation in the Onset condition, only one split was achieved, namely with PC1 Auditory Acuity (**Figure 5B**). Participants with a PC1 Auditory Acuity higher than 0.96 typically compensated more (mean 16%; 10 participants), while those with a PC1 Auditory Acuity score lower than 0.96 compensated less (mean 3.9%; 19 participants). Therefore, participants with better auditory acuity (PC1 Auditory Acuity) compensated more for the vowel in the Onset condition (see lower "Compensation" strip plot in **Figure 5B**). For strong compensatory responses (above ∼17%) high PC1 Auditory Acuity performance (yellow dots) is found, while for very weak compensators and followers (below ∼4% compensation) only low PC1 Auditory Acuity performance (magenta dots) is found. But in the mid-range of compensation substantial overlap of the perception groups is shown, indicating the limits on prediction accuracy. Note, however, that none of the participants with high auditory acuity actually followed the perturbation (i.e., compensation < 0).

For vowel compensation in the Coda condition (**Figure 5C**), only Unpaced Motor Variability emerged as a splitting factor, whereby participants with higher Unpaced Motor Variability (>-0.33) showed stronger compensatory responses (mean 15.7%; 18 participants), and participants with lower Unpaced Motor Variability compensated less (mean 8.9%; 10 participants). Overall prediction accuracy appears quite weak, since the two groups overlap substantially ("Compensation" strip plot, **Figure 5C**). Nonetheless, it seems worth pointing out that the nine participants with the strongest response ($>$∼17% compensation) belong to the high motor variability group (yellow dots).

Finally, the tree for CC compensation in the Coda Condition (**Figure 5D**) was first split with regard to Unpaced Motor Variability, whereby participants with Unpaced Motor Variability higher than -0.33 compensated more (mean 20.2%; 18 participants), and speakers below this score (less Unpaced Motor Variability) compensated less (mean 8.9%; 8 participants). For those with higher Unpaced Motor Variability (above –0.33), PC2 BAT Perception split the data further into subgroups, whereby participants with better perception of beat-alignment (>-0.7) compensated to a greater extent (mean 22.4%; 13 participants), while those with a lower perception score compensated less (mean 14.6%; five participants). To summarize, higher motor variability in unpaced tapping leads to stronger compensatory responses. For participants with higher motor variability, better beat-alignment perception abilities enhance compensatory responses and lower beat-alignment perception performance weakens responses. As visualized in the "Compensation" strip plot of **Figure 5D**, the group of high Unpaced Motor Variability and high PC2 BAT Perception performers (yellow dots) does not overlap (except one participant) with the group of low Unpaced Motor Variability performers (magenta dots), indicating quite precise prediction of compensation by the tree model. The in-between group with higher Unpaced Motor Variability and weaker perception of beat-alignment (brown dots) shows medium strong responses. This group selection supports the idea that better perceptual abilities lead to stronger compensation, but only if a certain malleability in the motor system is given.

In summary, the most relevant predictors for compensation in the Onset condition were the perception PCs, while for the Coda condition Unpaced Motor Variability was most relevant. Vowel prediction was generally less pronounced than CC prediction, as indicated by the overlap of the groups in the "Compensation" strips in **Figure 5**. The division into "good" and "worse" performers in Unpaced Motor Variability or perception was typically computed by dividing into thirds rather than at the median or mean along the respective variable. Note here that PC1 Paced Motor Variability did not achieve a split in the data for any segment/condition, and will therefore be dropped in the following analyses.

Linear models

To further quantify the quality of the subdivisions achieved by the trees, linear models were fitted with the predictors derived from the regression tree analysis. Analyses were performed using the *lm* function of R's included stats package (v4.1.2). These aimed at providing an indication of the accuracy of the modeled trees without a test and a training set, but by including the predictors provided by the respective regression tree and their interaction into a linear regression model. The prediction of each linear model, i.e., predicted compensation values versus measured compensation values, is visualized in **Figure 6**.

The linear model for CC compensation in the Onset condition was therefore modeled with compensation as the dependent variable, and both perception PCs and their interaction as predictors. The result indicated that the model was significant [$F(3,24) = 3.28$, $p = 0.038$] and explained 20.2% of the variance in the data (adjusted R-squared). The model revealed a significant contribution of PC2 BAT Perception to modeling the data ($t = 2.15$, $p = 0.041$), but no significant contribution of PC1 Auditory Acuity ($t = 0.430$, $p = 0.670$) nor the interaction between both Perception PCs ($t = -1.624$, $p = 0.117$).

The linear model for vowel compensation in the Onset condition was computed with compensation as the dependent variable and PC1 Auditory Acuity as predictor. The model explained 11.8% of the variance in the data and was significant,

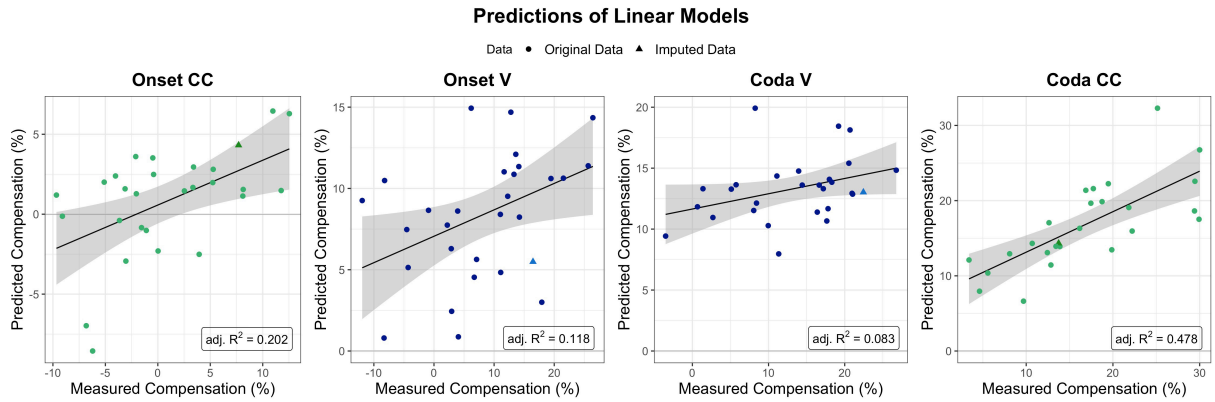


FIGURE 6
Predictions as provided by the linear models per segment and condition of interest (with the four perturbation measures ordered chronologically from Onset CC on the left to Coda CC on the right). Model fits for Onset CC, Onset V, and Coda CC are significant, Coda V non-significant. The Coda CC model was the one explaining most variance in the data (adj. *R*-squared = 0.478). Dots mark single participants, triangles single participants for which data was imputed for (at least one of) the predictors.

$F(1,27) = 4.74$, $p = 0.038$. The model revealed a significant contribution of PC1 Auditory Acuity to modeling the data ($t = 2.178$, $p = 0.038$).

Vowel compensation in the Coda condition was modeled with Unpaced Motor Variability as a predictor. Overall model fit was quite weak (adjusted R-squared = 0.083), and non-significant, $F(1,26) = 3.44$, $p = 0.075$.

CC compensation in the Coda condition was modeled with Unpaced Motor Variability and PC2 BAT Perception as predictors as well as their interaction. The model was significant, $F(3,22) = 8.617$, $p < 0.001$, and accounted for 47.8% of the variance. Unpaced Motor Variability contributed significantly to the model ($t = 4.617$, $p < 0.001$), and so did PC2 BAT Perception ($t = 2.126$, $p = 0.045$). The interaction between both predictors did not contribute significantly ($t = 1.210$, $p = 0.239$).

Speech motor variability and compensation

While the previous section indicated that non-verbal motor abilities relate to responses to temporal auditory feedback perturbation, one could ask if similar effects can be seen for speech motor variability and perturbation. The following section briefly examines temporal speech variability in the baseline of the perturbation experiment and its relation to compensatory behavior. For the assessment of speech motor variability, data from the perturbation experiment was examined. The coefficient of variation (standard deviation divided by the mean) of the word-normalized segment durations (V and CC) produced in the baseline phase per experiment condition (Onset/Coda) per participant was calculated. The first nine trials of the baseline were not included into calculations, as they were excluded in analyses in Oschkinat and Hoole (2020). The coefficient of variation (cv) of baseline productions per segment and condition was then correlated with the respective compensation measure. Unlike tapping motor variability, the cv of the baseline segment durations were not significantly related to the amount of compensation. Figure 7 visualizes the relationship between compensation and baseline variability per segment of interest, accompanied by statistical outcome of the correlations.

Discussion

Main findings

The present study investigated the connection of individual perceptual and general motor abilities to responses to temporal auditory feedback perturbation. Our results support the idea that individual perceptual abilities and individual capacities in precise motor execution both shape the speech production process. The extracted qualities were summarized by their underlying dimensions obtained from a principal component analysis and served as predictors in statistical modeling. The analyses followed an exploratory-confirmatory approach: Regression trees selected the most relevant predictors, which subsequently were included in linear modeling. In model prediction, beat-alignment perception, general auditory acuity, and motor variability of an unpaced tapping task explained variance in the perturbation response data. In general, the perceptual dimension generated the most prominent predictors for describing variance in response to temporal onset perturbation of syllables (applying to both complex onsets and the following vowel). In contrast, motor variability of unpaced tapping was most relevant to predict responses to temporally perturbed auditory feedback of syllable codas (significant for consonant coda clusters, non-significant for the preceding vowel). This relationship supports our second hypothesis, suggesting that greater motor variability allows for

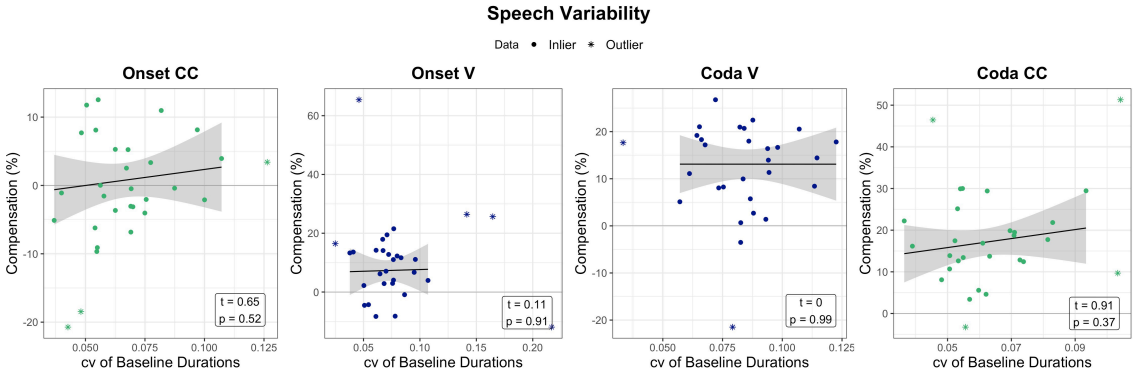


FIGURE 7
Relation between baseline variability in the perturbation experiment (coefficient of variation of baseline durations per segment) in relation to compensation. Effects are non-significant. Green and blue dots mark the data that has been used for model prediction. Stars are outliers that have been excluded in statistical modeling. The black line visualizes the regression of the data (dots) without outliers (stars).

more adaptation (whereby adaptation was only found in the Coda condition). Auditory acuity, on the other hand, was suggested to be relevant for responses in the online control as well as adaptation. Indeed, auditory acuity explained variance in compensation to the segments in the Onset condition, and the Coda CC in the Coda condition. Thereby, the vowel in the Onset condition and CC in the Coda condition were the second perturbed segment and therefore exposed to online effects induced by the stretching of the preceding segment. This result is in line with the findings by Martin et al. (2018), who found auditory acuity more relevant for online effects than for long-term adaptation of motor commands. The vowel in the Coda condition showed the weakest predictability, as the only segment that comprised exclusively adaptive responses and was compensatorily shortened in production. Better perceptual abilities and higher motor variability were linked to stronger compensation, as expected in our first hypothesis (but see the contribution of PC2 BAT Perception for Coda CC compensation discussed further below).

Concerning the motor variability component, there is, to our knowledge, until now no study that explicitly investigated the connection between individual motor timing capacities and responses to perturbed feedback. However, the current study's results align with our main conclusion drawn from Oschkinat and Hoole (2020): Greater variability leads to a less stable system that is more malleable in the face of auditory feedback perturbation. In Oschkinat and Hoole (2020), this assumption was directed to *structural* effects: Syllable onsets are articulatorily more stable than syllable codas and therefore less malleable in the face of an auditory feedback perturbation. The data of the current study suggest that structurally given malleability can be further modulated by *individual* motor stability in temporal auditory feedback perturbation. Especially, perturbation to coda segments (the prediction accuracy of R-squared 0.47% was by far the highest of our four conditions) showed that low motor variability (better temporal stability) was linked to less compensatory responses, more precisely to less *adaptation*. This assumption is further supported by the correlation of the compensation measures in Coda condition with each other. Both segments seem to share a certain malleability, although this has to be interpreted carefully because they were also manipulated within one perturbation frame. For speakers with higher motor variability, better perceptual performance increased compensatory responses, and weaker perceptual performance weakened compensatory responses. This interplay indicates a tradeoff between perceptual and motor abilities, and supports the finding that better perceptual abilities do indeed lead to more compensation/adaptation, but that adaptation only occurs if a certain system malleability is given. For future paradigms that aim at capturing strong adaptive responses, it might be particularly revealing to focus on speakers with this specific combination of high auditory acuity and high motor variability, and provoke adaptive shortening responses rather than lengthening, since the latter might always contain not just an adaptive component but also an online response.

Timing mechanisms

Regarding the perceptual components, these findings are in line with previous studies that examined the link between auditory acuity and responses to spectral feedback alterations. For example, in Villacorta et al. (2007) and Brunner et al. (2011), speakers with higher auditory acuity compensated more for a spectral shift in the auditory feedback of vowels. Still, the comparability with these studies is not naturally given: While Villacorta et al. (2007) assessed auditory discrimination ability for F1 when F1 was perturbed in the experiment, the perceptual correlate of perturbed speech timing is not self-evident (see Oschkinat and Hoole, 2022 for discussion). In our data, both duration discrimination and perceptual beat-alignment abilities were linked to temporal feedback

perturbation. The regression tree structure for the Onset CC condition (**Figure 5A**) suggests that beat-alignment judgments (event-based timing mechanisms, represented by PC2 BAT Perception) are most relevant for predicting compensatory behavior for temporal perturbation of the complex onset. This relation supports our minor hypothesis raised after interpreting the PCs in section "Interpreting PC scores" that event-based timing mechanisms might be more relevant in predicting behavior in onsets. Speakers who more precisely detect a shift of the p-center in the auditory feedback (as introduced with the stretched Onset CC in perturbation), may adjust more for it. However, PC2 BAT Perception further explained variance in compensation to the Coda CC segment, which is not as expected or explainable with the p-center concept. In this case, recall that higher PC2 BAT Perception scores were further associated with weaker perceptual abilities in discriminating duration differences in codas (coda perception task). Therefore, the prediction of beat-alignment timing being more closely associated with onsets and duration-based timing with codas could not be fully supported. The Onset CC regression tree analysis further suggested that good duration discrimination abilities (duration-based timing, PC1 Auditory Acuity) can partially counteract worse PC2 BAT Perception abilities. Moreover, bad performance in both perception domains leads to poor compensation, or more precisely, mainly to a following of the perturbation (negative responses). The predictability of following responses from poor perceptual skills might result from the inability of speakers to precisely locate either the direction of the shift in the auditory feedback or the direction in response that would oppose the perceived shift direction. The present findings further add to the discussion about what leads to following responses in so many perturbation studies (see, e.g., Katseff et al., 2012; Franken et al., 2018). The aforementioned perceptual abilities explained 20.2% of the variance in our data (Onset CC condition). Of the substantial remaining variance, some of it might be explained, for example, by how speakers balance auditory against somatosensory errors (Katseff et al., 2012). Indeed, we have suggested that somatosensory feedback may be particularly relevant in syllable onsets, since somatosensory feedback is accessible earlier than auditory feedback. In syllable onsets, auditory feedback cannot be used to estimate relative durations within the syllable as it is possible in codas, where onset and vowel durations have been already perceived. Therefore, somatosensory feedback could be more informative for error correction in timing (Oschkinat and Hoole, 2020).

Speech motor variability

While motor variability in unpaced tapping correlated with responses in the Coda condition, a similar relationship could not be found for measures of (temporal) *speech* motor variability. Certainly, these results need to be interpreted cautiously. While the unpaced tapping task tests pure, task-unspecific internal timing stability with a low-complexity motor task, speech production requires a complex coproduction of muscles, each of them allowing for variability/degrees of freedom in the execution. Further, speech variability measures were assessed from only 11 trials in the baseline phase, which might not give a solid mean for such analyses. Regarding the motor variability measure from finger tapping, we nevertheless admit that in using this measure as an indication for internal variability/malleability, we cannot precisely disentangle imprecision in motor execution from imprecision in the internal motor plan. However, PC1 Paced Motor Variability did not seem relevant in explaining the data of this study and was dropped as a predictor for compensation in the regression tree analyses. One might conclude here that the difference between paced and unpaced timing tasks is not the ability to precisely execute motor commands according to an internal plan, but rather the internal rhythmic representation that is needed in unpaced tapping but externally provided in paced tapping. By closely examining previous studies, the results turn out to be partially in line with investigations on spectral speech variability and compensation to spectral perturbation. In previous studies, compensation to spectral perturbations correlated with the variability of contrast of different speech sounds in production (Ghosh et al., 2010; Brunner et al., 2011). However, compensation did not correlate with the variability of one individual parameter (e.g., F1) in repeated phoneme productions (MacDonald et al., 2010, 2011). Mixed findings were also provided by Nault and Munhall (2020), who conducted a study on inter-and intraspeaker variability. They measured the standard deviation of the first two formants of vowels produced in the baseline phase of a spectral perturbation experiment and found a relation between F1 variability in the baseline and F1 compensation in the hold phase, but no contribution of baseline variability of F2 as a predictor for compensation to perturbed F2. Another recent study neither found relations between adaptation and vowel spacing in the baseline phase nor correlations between adaptation and variability in productions of single baseline phonemes (Parrell and Niziolek, 2021). In this view, the non-existent relationship between speech variability and compensation in our data is in line with the findings by MacDonald et al. (2010, 2011), Parrell and Niziolek (2021), and partially Nault and Munhall (2020). However, analogously to the variety of spectral measures, there is still considerable room to ponder about the best parameter to measure variability in production of speech timing. Further, it has to be kept in mind that temporal information of speech is different from spectral information: While spectral properties of fricatives and vowels serve to distinguish similar sounds from each other, duration's primary purpose is not to distinguish sounds but to give their spectral evolvement

a stage[2]. Individual variability in production might not be relevant when high variability cannot result in another category. The distinctive function of duration is much less pronounced than the distinctive function of spectral properties of speech. Duration and timing are certainly not arbitrary but follow different goals, such as enabling fluency and intelligibility and realizing prosodic aspects of speech.

Conclusion

The findings of the current study gave insights into how feedback and feedforward mechanisms in speech and non-speech are connected and how their interaction shapes timing in speech production. We also believe that the study provides a valuable foundation for guiding future studies in the selection of more targeted perception and tapping tests for similar research approaches. In particular, motor variability in tapping tasks seems worth further exploration; especially unpaced tapping as a measure of general internal motor stability should be considered. Regarding the significance of the current and similar studies, it should be noted that to date not much is known about the reproducibility of reactions to (temporal) auditory feedback perturbation. Saying this, there is certainly a need for establishing a firm understanding of how compensatory responses to the same perturbations vary within participant across multiple sessions. Further, in future investigations, it might be worth looking into groups of participants with different levels of auditory acuity and motor stability, such as musicians and non-musicians. In follow-up investigations, the perception staircase paradigm could be improved by using a 4-interval 2-alternative AABA design, which would probably provide more reliable threshold estimations than the 2-interval paradigm with catch-trials or an ABX paradigm (Gerrits and Schouten, 2004). Finally, it should be noted that in our PCA approach, some tasks were dropped in calculations, and relationships between the single motor and perception tasks and compensation to temporal perturbation might have been blurred. Nevertheless, we see the PCA-driven analyses as clearly crucial here in making the large number of individual tasks tractable for analysis, and indeed provided interesting insights, e.g., by distinguishing perception tasks based on timing mechanisms.

Data availability statement

The measurement data supporting the conclusions of this article will be made available by the authors, without undue reservation.

2 We are not, of course, ignoring the presence of quantity distinctions.

Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Committee of the Medical Faculty of the Ludwig Maximilian University. The participants provided their written informed consent to participate in this study.

Author contributions

MO, PH, SF, and SDB: study design, data processing, and writing. MO: data acquisition and analyses. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the German Research Foundation, Deutsche Forschungsgemeinschaft (DFG), under Grant No. HO 3271/6-1 and the German Academic Exchange Service (DAAD) from grants of the Bundesministerium für Bildung und Forschung (BMBF).

Acknowledgments

We thank Michele Gubian for his help with regression trees and two reviewers for their valuable comments on a previous version of the manuscript. We also thank Sebastian Böhnke, Valeria Meißner, and Paul Bachmann for their help in running the tests and our participants for taking part in this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Beretta, L., and Santaniello, A. (2016). Nearest neighbor imputation algorithms: A critical evaluation. *BMC Med. Inform. Decis. Mak.* 16(Suppl. 3):74. doi: 10.1186/s12911-016-0318-z

Boersma, P., and Weenink, D. (1999). *PRAAT, a system for doing phonetics by computer (version 5.3.78) [computer program]*. Available Online at: http://www.praat.org [accessed June 14, 2021].

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Cart. Classification and regression trees*. Monterey, CA: Wadsworth and Brooks/Cole.

Browman, C. P., and Goldstein, L. M. (1992). Articulatory phonology: An overview. *Phonetica* 49, 155–180. doi: 10.1159/000261913

Browman, C. P., and Goldstein, L. M. (2000). Competing constraints on intergestural coordination and self-organization of phonological structures. Les Cahiers de l'ICP. *Bull. Commun. Parlée* 5, 25–34.

Brunner, J., Ghosh, S., Hoole, P., Matthies, M., Tiede, M., and Perkell, J. (2011). The influence of auditory acuity on acoustic variability and the use of motor equivalence during adaptation to a perturbation. *J. Speech Lang. Hear. Res.* 54, 727–739. doi: 10.1044/1092-4388(2010/09-0256)

Burnett, T. A., Freedland, M. B., Larson, C. R., and Hain, T. C. (1998). Voice F0 responses to manipulations in pitch feedback. *J. Acoust. Soc. Am.* 103, 3153–3161. doi: 10.1121/1.423073

Byrd, D. (1996). Influences on articulatory timing in consonant sequences. *J. Phon.* 24, 209–244. doi: 10.1006/jpho.1996.0012

Cai, S., Boucek, M., Ghosh, S. S., Guenther, F. H., and Perkell, J. S. (2008). "A system for online dynamic perturbation of formant trajectories and results from perturbations of the Mandarin triphthong /iau/," in *Proceedings of the 8th international seminar on speech production ISSP*, Strasbourg, 65–68.

Cai, S., Ghosh, S. S., Guenther, F. H., and Perkell, J. S. (2011). Focal manipulations of formant trajectories reveal a role of auditory feedback in the online control of both within-syllable and between-syllable speech timing. *J. Neurosci.* 31, 16483–16490. doi: 10.1523/JNEUROSCI.3653-11.2011

Cummins, F., and Port, R. (1998). Rhythmic constraints on stress timing in English. *J. Phon.* 26, 145–171. doi: 10.1006/jpho.1998.0070

Dalla Bella, S., Farrugia, N., Benoit, C.-E., Begel, V., Verga, L., Harding, E., et al. (2017). BAASTA: Battery for the assessment of auditory sensorimotor and timing abilities. *Behav. Res. Methods* 49, 1128–1145. doi: 10.3758/s13428-016-0773-6

Drake, C., Jones, M. R., and Baruch, C. (2000). The development of rhythmic attending in auditory sequences: Attunement, referent period, focal attending. *Cognition* 77, 251–288. doi: 10.1016/s0010-0277(00)00106-2

Falk, S., Müller, T., and Dalla Bella, S. (2015). Non-verbal sensorimotor timing deficits in children and adolescents who stutter. *Front. Psychol.* 6:847. doi: 10.3389/fpsyg.2015.00847

Falk, S., Volpi-Moncorger, C., and Dalla Bella, S. (2017). Auditory-motor rhythms and speech processing in French and German listeners. *Front. Psychol.* 8:395. doi: 10.3389/fpsyg.2017.00395

Field, A., Miles, J., and Field, Z. (2012). *Discovering statistics using R*. Thousand Oaks, CA: Sage publications.

Floegel, M., Fuchs, S., and Kell, C. A. (2020). Differential contributions of the two cerebral hemispheres to temporal and spectral speech feedback control. *Nat. Commun.* 11:2839. doi: 10.1038/s41467-020-16743-2

Franken, M. K., Acheson, D. J., McQueen, J. M., Hagoort, P., and Eisner, F. (2018). Opposing and following responses in sensorimotor speech control: Why responses go both ways. *Psychon. Bull. Rev.* 25, 1458–1467. doi: 10.3758/s13423-018-1494-x

Gerrits, E., and Schouten, M. (2004). Categorical perception depends on the discrimination task. *Percept. Psychophys.* 66, 363–376. doi: 10.3758/BF03194885

Ghosh, S. S., Matthies, M. L., Maas, E., Hanson, A., Tiede, M., Ménard, L., et al. (2010). An investigation of the relation between sibilant production and somatosensory and auditory acuity. *J. Acoust. Soc. Am.* 128, 3079–3087. doi: 10.1121/1.3493430

Goldstein, L., and Pouplier, M. (2014). "The temporal organization of speech," in *The Oxford handbook of language production*, eds M. Goldrick, V. Ferreira, and M. Miozzo (New York, NY: Oxford University Press), 210–227.

Grube, M., Cooper, F., Chinnery, P., and Griffiths, T. (2010). Dissociation of duration-based and beat-based auditory timing in cerebellar degeneration. *Proc. Natl. Acad. Sci. U.S.A.* 107, 11597–11601. doi: 10.1073/pnas.0910473107

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *J. Commun. Disord.* 39, 350–365. doi: 10.1016/j.jcomdis.2006.06.013

Guenther, F. H. (2016). *Neural control of speech*. Cambridge, MA: MIT Press.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educ. Psychol. Meas.* 20, 141–151.doi: 10.1177/001316446002000116

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika* 35, 401–415. doi: 10.1007/BF02291817

Karlin, R., Naber, C., and Parrell, B. (2021). Auditory feedback is used for adaptation and compensation in speech timing. *J. Speech Lang. Hear. Res.* 64, 3361–3381. doi: 10.1044/2021_JSLHR-21-00021

Katseff, S., Houde, J., and Johnson, K. (2012). Partial compensation for altered auditory feedback: A tradeoff with somatosensory feedback? *Lang. Speech* 55, 295–308. doi: 10.1177/0023830911417802

MacDonald, E. N., Goldberg, R., and Munhall, K. G. (2010). Compensations in response to real-time formant perturbations of different magnitudes. *J. Acoust. Soc. Am.* 127, 1059–1068. doi: 10.1121/1.3278606

MacDonald, E. N., Purcell, D. W., and Munhall, K. G. (2011). Probing the independence of formant control using altered auditory feedback. *J. Acoust. Soc. Am.* 129, 955–965. doi: 10.1121/1.3531932

Martin, C. D., Niziolek, C. A., Duñabeitia, J. A., Perez, A., Hernandez, D., Carreiras, M., et al. (2018). Online adaptation to altered auditory feedback is predicted by auditory acuity and not by domain-general executive control resources. *Front. Human Neurosci.* 12:91. doi: 10.3389/fnhum.2018.00091

Milborrow, S. (2021). *rpart.plot: Plot 'rpart' models: An enhanced version of 'plot.rpart'. R package version 3.1.0*. Available Online at: https://CRAN.R-project.org/package=rpart.plot (accessed April 25, 2022).

Mitsuya, T., MacDonald, E. N., and Munhall, K. G. (2014). Temporal control and compensation for perturbed voicing feedback. *J. Acoust. Soc. Am.* 135, 2986–2994. doi: 10.1121/1.4871359

Morton, J., Marcus, S., and Frankish, C. (1976). Perceptual centers (P-centers). *Psychol. Rev.* 83, 405–408. doi: 10.1037/0033-295X.83.5.405

Nault, D. R., and Munhall, K. G. (2020). Individual variability in auditory feedback processing: Responses to real-time formant perturbations and their relation to perceptual acuity. *J. Acoust. Soc. Am.* 148, 3709–3721. doi: 10.1121/10.0002923

Nozaradan, S., Peretz, I., and Mouraux, A. (2012). Selective neuronal entrainment to the beat and meter embedded in a musical rhythm. *J. Neurosci.* 32, 17572–17581. doi: 10.1523/JNEUROSCI.3203-12.2012

Oschkinat, M., and Hoole, P. (2020). Compensation to real-time temporal auditory feedback perturbation depends on syllable position. *J. Acoust. Soc. Am.* 148, 1478–1495. doi: 10.1121/10.0001765

Oschkinat, M., and Hoole, P. (2022). Reactive feedback control and adaptation to perturbed speech timing in stressed and unstressed syllables. *J. Phon.* 91:101133. doi: 10.1016/j.wocn.2022.101133

Parrell, B., and Niziolek, C. A. (2021). Increased speech contrast induced by sensorimotor adaptation to a nonuniform auditory perturbation. *J. Neurophysiol.* 125, 638–647. doi: 10.1152/jn.00466.2020

Parrell, B., Goldstein, L., Lee, S., and Byrd, D. (2014). Spatiotemporal coupling between speech and manual motor actions. *J. Phon.* 42, 1–11. doi: 10.1016/j.wocn.2013.11.002

Peelle, J. E., and Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Front. Psychol.* 3:320. doi: 10.3389/fpsyg.2012.00320

Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Stockmann, E., Tiede, M., et al. (2004a). The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts. *J. Acoust. Soc. Am.* 116, 2338–2344.

Perkell, J. S., Matthies, M. L., Tiede, M., Lane, H., Zandipour, M., Marrone, N., et al. (2004b). The Distinctness of Speakers' /s/ - /ʃ/ Contrast is related to their auditory discrimination and use of an articulatory saturation effect. *J. Speech Lang. Hear. Res.* 47, 1259–1269. doi: 10.1044/1092-4388(2004/095)

Perkell, J. S., Lane, H., Ghosh, S., Matthies, M. L., Tiede, M., Guenther, F., et al. (2008). "Mechanisms of vowel production: Auditory goals and speaker acuity," in *Proceedings of the 8th international seminar on speech production*, Strasbourg, 29–32.

Pouplier, M. (2012). "The gestural approach to syllable structure: Universal, language-and cluster-specific aspects," in *Speech planning and dynamics*, eds S. Fuchs, M. Weirich, D. Pape, and P. Perrier (Frankfurt: Peter Lang), 63–96.

Purcell, D. W., and Munhall, K. G. (2006). Compensation following real-time manipulation of formants in isolated vowels. *J. Acoust. Soc. Am.* 119, 2288–2297. doi: 10.1121/1.2173514

Puyjarinet, F., Bégel, V., Gény, C., Driss, V., Cuartero, M. C., Kotz, S. A., et al. (2019). Heightened orofacial, manual, and gait variability in Parkinson's disease results from a general rhythmic impairment. *NPJ Parkinsons Dis.* 5:19. doi: 10.1038/s41531-019-0092-6

Repp, B. H. (2005). Sensorimotor synchronization: A review of the tapping literature. *Psychon. Bull. Rev.* 12, 969–992. doi: 10.3758/BF03206433

Repp, B. H., and Su, Y.-H. (2013). Sensorimotor synchronization: A review of recent research (2006–2012). *Psychon. Bull. Rev.* 20, 403–452. doi: 10.3758/s13423-012-0371-2

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R J.* 8, 289–317. doi: 10.32614/RJ-2016-021

Teki, S., Grube, M., and Griffiths, T. D. (2012). A unified model of time perception accounts for duration-based and beat-based timing mechanisms. *Front. Integr. Neurosci.* 5:90. doi: 10.3389/fnint.2011.00090

Teki, S., Grube, M., Kumar, S., and Griffiths, T. D. (2011). Distinct neural substrates of duration-based and beat-based auditory timing. *J. Neurosci.* 31, 3805–3812. doi: 10.1523/JNEUROSCI.5561-10.2011

Therneau, T., and Atkinson, B. (2019). *rpart: Recursive partitioning and regression trees. R package version 4.1-15*. Available Online at: https://CRAN.R-project.org/package=rpart (accessed April 25, 2022).

Tourville, J. A., Cai, S., and Guenther, F. (2013). Exploring auditory-motor interactions in normal and disordered speech. *Proc. Meet. Acoust.* 19:060180 . doi: 10.1121/1.4800684

Villacorta, V. M., Perkell, J. S., and Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *J. Acoust. Soc. Am.* 122, 2306–2319.

Yates, A. J. (1963). Delayed auditory feedback. *Psychol. Bull.* 60, 213–232. doi: 10.1037/h0044155

Appendix

Effects of musicality

Along with the three experiment blocks, musical education of the participants was assessed by a questionnaire asking about whether they received musical education, where they received it, for how long, on what instruments/singing, and whether they learned an instrument without education. Since musicality was not the main interest in this study, for a simple overview, participants were grouped into those who received musical education (at least for 2 years), and those who did not receive any musical education (non-musical group).

Additional analyses showed that when comparing the two groups with a two-sampled Welch test, no group differences were observed for PC1 Paced Motor Variability nor for Unpaced Tapping Motor Variability, PC1 Auditory Acuity nor PC2 BAT Perception (non-musical group: $n = 10$, musical group: $n = 25$). Note here, however, that the group of musically educated participants was much larger than the group of non-musically-educated participants, and that there was high variability in duration of musical education within the group of musicians (from 2 years up to 13 years on one instrument), and in the start age of musical education on the first instrument (5–13 years of age).

Further, there were no group differences in response to the temporal perturbation for the Onset CC (non-musical group: $n = 9$, musical group: $n = 19$), no difference for Onset V (non-musical group: $n = 9$, musical group: $n = 20$), and the Coda CC (non-musical group: $n = 7$, musical group: $n = 19$). For the Coda V, the group of musically educated participants adapted less than the non-musical group ($t = 2.447$, df = 15.99, p-value = 0.026, non-musical group: $n = 8$, musical group: $n = 20$). Since the Coda V was the only segment that exhibited significant adaptation effects, this connection suggests that musicians showed more resistance in adapting to perceived errors than non-musically trained participants. Note however, that this relationship was not found for Unpaced Motor Variability or PC1 Motor Variability, perhaps due to a different subset of participants based on outlier exclusion. The effect of musical training on adaptation should give a direction for future studies and suggests that highly trained musicians as compared to non-trained musicians might be a group worth investigating more closely.

frontiers | Frontiers in Human Neuroscience

TYPE Original Research
PUBLISHED 02 November 2022
DOI 10.3389/fnhum.2022.929687

Check for updates

OPEN ACCESS

EDITED BY
Jeffery A. Jones,
Wilfrid Laurier University, Canada

REVIEWED BY
David Jackson Morris,
University of Copenhagen, Denmark
Nishant Rao,
Haskins Laboratories, United States

*CORRESPONDENCE
Elaine Kearney
ekearney@bu.edu

SPECIALTY SECTION
This article was submitted to Speech and Language, a section of the journal Frontiers in Human Neuroscience

RECEIVED 27 April 2022
ACCEPTED 04 October 2022
PUBLISHED 02 November 2022

CITATION
Kearney E, Nieto-Castañón A, Falsini R, Daliri A, Heller Murray ES, Smith DJ and Guenther FH (2022) Quantitatively characterizing reflexive responses to pitch perturbations. *Front. Hum. Neurosci.* 16:929687. doi: 10.3389/fnhum.2022.929687

COPYRIGHT
© 2022 Kearney, Nieto-Castañón, Falsini, Daliri, Heller Murray, Smith and Guenther. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Quantitatively characterizing reflexive responses to pitch perturbations

Elaine Kearney[1]*, Alfonso Nieto-Castañón[1,2], Riccardo Falsini[1], Ayoub Daliri[3], Elizabeth S. Heller Murray[4], Dante J. Smith[5] and Frank H. Guenther[1,6,7]

[1]Department of Speech, Language, and Hearing Sciences, Boston University, Boston, MA, United States, [2]The McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, United States, [3]College of Health Solutions, Arizona State University, Tempe, AZ, United States, [4]Department of Communication Sciences and Disorders, Temple University, Philadelphia, PA, United States, [5]Gradutate Program for Neuroscience, Boston University, Boston, MA, United States, [6]Department of Biomedical Engineering, Boston University, Boston, MA, United States, [7]The Picower Institute for Learning and Memory, Massachusetts Institute of Technology, Cambridge, MA, United States

Background: Reflexive pitch perturbation experiments are commonly used to investigate the neural mechanisms underlying vocal motor control. In these experiments, the fundamental frequency–the acoustic correlate of pitch–of a speech signal is shifted unexpectedly and played back to the speaker via headphones in near real-time. In response to the shift, speakers increase or decrease their fundamental frequency in the direction opposing the shift so that their perceived pitch is closer to what they intended. The goal of the current work is to develop a quantitative model of responses to reflexive perturbations that can be interpreted in terms of the physiological mechanisms underlying the response and that captures both group-mean data and individual subject responses.

Methods: A model framework was established that allowed the specification of several models based on Proportional-Integral-Derivative and State-Space/Directions Into Velocities of Articulators (DIVA) model classes. The performance of 19 models was compared in fitting experimental data from two published studies. The models were evaluated in terms of their ability to capture both population-level responses and individual differences in sensorimotor control processes.

Results: A three-parameter DIVA model performed best when fitting group-mean data from both studies; this model is equivalent to a single-rate state-space model and a first-order low pass filter model. The same model also provided stable estimates of parameters across samples from individual subject data and performed among the best models to differentiate between subjects. The three parameters correspond to gains in the auditory feedback controller's response to a perceived error, the delay of this response, and the gain of the somatosensory feedback controller's "resistance" to this correction. Excellent fits were also obtained from a four-parameter model with an additional auditory velocity error term; this model was

better able to capture multi-component reflexive responses seen in some individual subjects.

Conclusion: Our results demonstrate the stereotyped nature of an individual's responses to pitch perturbations. Further, we identified a model that captures population responses to pitch perturbations and characterizes individual differences in a stable manner with parameters that relate to underlying motor control capabilities. Future work will evaluate the model in characterizing responses from individuals with communication disorders.

KEYWORDS

computational modeling, motor control, speech production, pitch, auditory feedback

Introduction

Auditory perturbation paradigms have become an important experimental approach in uncovering the neural mechanisms underlying vocal motor control. First described by Elman (1981), these paradigms involve manipulating the frequency spectrum of someone's speech and playing it back to them via headphones in near real-time, such that they–often subconsciously–detect an error in their production. In pitch perturbation experiments specifically, the frequency spectrum is perturbed so that the fundamental frequency (f_o; the acoustic correlate of pitch) is higher or lower than produced. In response to this manipulation, speakers will change their f_o in the direction opposite the perturbation, which makes what they hear in the headphones closer to what they intended to produce. When the perturbations are unexpected (for example, when applied randomly on a small percentage of trials or when applied at a random time during each trial), the compensatory response is referred to as reflexive; that is, the response is evident within a given perturbed trial but has a limited effect on subsequent trials. This contrasts with perturbations sustained over many trials that elicit both reflexive within-trial responses as well as adaptive across-trial responses (Daliri, 2021). The current work focuses on reflexive responses to pitch perturbations; we will use the term *pitch shift reflex* (PSR) to refer to such responses (Kiran and Larson, 2001).

There is a long history of utilizing reflexive responses as a diagnostic tool for probing neural function. For example, the pupillary light reflex was used by Claudius Galenus in the 2nd century to evaluate the visual capabilities of candidates for cataract surgery (see Thompson, 2003 for a historical review). Since that time, scientists have characterized the pupillary light reflex in ever-increasing detail, and modern investigations often utilize pupillography to accurately measure the time course of the pupil's reaction to changes in light input. These studies have led to the parameterization of the temporal profile of the pupillary light reflex (e.g., Hall and Chilcott, 2018) as well as parameterized mathematical models of the dynamics of the pupillary light reflex that capture individual differences (Pamplona, 2008). The different parameters in these characterizations correspond to different neural processes; thus, an individual's pupillary light reflex can be used to differentiate damage to one part of the nervous system from damage to another, in turn allowing clinicians to make informed decisions regarding treatment options. The dynamics of the pupillary light reflex are now used to gauge neural function in a wide range of disorders extending beyond impairment of the visual system, including concussion (Master et al., 2020), schizophrenia (Bär et al., 2008), Alzheimer's disease (Tales et al., 2001), Parkinson's disease (Stergiou et al., 2009), autism spectrum disorders (Fan et al., 2009), and alcoholism (Rubin, 1980). Against this background, a primary goal of the current study is to mathematically characterize the pitch reflex response using mathematical models with parameters that reflect the function of different neural subsystems involved in the control of voice.

Since the early application of the pitch perturbation paradigm, over 140 studies have used this paradigm to investigate various aspects of vocal motor control and across different populations. These studies have revealed several properties of the PSR. First, responses are typically in the direction opposite the perturbation, while a small percentage of responses occur in the same direction as the perturbation (e.g., Burnett et al., 1998; Franken et al., 2018). Second, the compensation is usually incomplete, likely reflecting an interaction between the auditory and somatosensory control systems (Smith et al., 2020). Third, the responses occur in a variety of speech stimuli (Natke and Kalveram, 2001; e.g., sustained vowels, syllables, running speech; Chen et al., 2007; Smith et al., 2020). In addition, investigations of the PSR in speakers of tonal languages, such as Mandarin, show an interaction between the linguistic intent of an utterance

and perturbations, with larger responses evident when the perturbation changes the meaning of a word (Xu et al., 2004). Musicians and singers, who have higher-than-average experience controlling pitch, are also able to ignore large pitch perturbations (∼200 cents) while they compensate more completely for smaller and shorter perturbations (∼25 cents) (Zarate et al., 2010; Behroozmand et al., 2014; Parkinson et al., 2014).

While the majority of pitch-perturbation studies to date have focused on neurotypical adult speakers, a growing number of studies have examined responses in children and individuals with communication disorders. Reflexive perturbation responses in children are evident as young as age 3 years (Russo et al., 2008; Scheerer et al., 2013, 2016; Heller Murray and Stepp, 2020) but are associated with longer response latencies and greater variability compared to adult responses. Studies have also investigated responses in individuals with Parkinson's disease (Kiran and Larson, 2001; Liu et al., 2012; Abur et al., 2021a), Alzheimer's disease (Ranasinghe et al., 2017), cerebellar degeneration (Houde et al., 2019; Li et al., 2019), apraxia of speech (Ballard et al., 2018), aphasia (Behroozmand et al., 2018, 2022), hyperfunctional voice disorders (Abur et al., 2021b), 16p11.2 deletions (Demopoulos et al., 2018), autism (Russo et al., 2008), and in those who stutter (Loucks et al., 2012; Sares et al., 2018, 2020). Collectively, these studies shed light on the development of vocal motor control and the mechanisms underlying speech and voice disorders. In the future, these findings may inform novel treatments that directly target these mechanisms.

Compensatory responses to pitch perturbations rely on neural processes that compare the target pitch for a given utterance to the pitch as sensed through audition and apply corrections if and when an error is detected. We can use computational models to explicate these internal processes by specifying the processes with mathematical equations and evaluating how well the equations (i.e., the models) explain existing experimental data. There are several candidate model classes that may be used to model reflexive pitch perturbation data, including *Proportional-Integral-Derivative (PID), State-Space (SS), and Directions Into Velocities of Articulators (DIVA) models*.

The PID model class was originally designed to mimic the steering strategy used by expert ship helmsmen (Minorsky, 1922) and is now commonly used in a wide range of engineering applications. This model class includes proportional (P) models, where the corrective command is proportional to the error signal; proportional-derivative (PD) models, where the proportional command is supplemented with a command that is formed by multiplying the derivative of the error signal by a gain; proportional-integral (PI) models, where the proportional command is supplemented with a command formed by multiplying the integral of the error signal by a gain; and finally, PID models that combine all three error terms. SS models also originated in control engineering and have been widely applied in studies of limb motor control (Thoroughman and Shadmehr, 2000; Smith et al., 2006; Galea et al., 2015; Huberdeau et al., 2015). SS models model physical systems as a set of input, output, and state variables using first-order differential equations. The DIVA model is a prominent neural network model of speech motor control (Guenther, 2016; Kearney and Guenther, 2019). It is organized around three control subsystems, namely feedforward, auditory feedback, and somatosensory feedback control, and has been used to explain a wide number of speech phenomena. Although the SS and DIVA models have different theoretical motivations, they are closely related mathematically (as we will demonstrate) and will be treated together throughout this paper.

To the best of our knowledge, only one study to date has utilized a computational model to simulate responses to a reflexive pitch perturbation paradigm (Larson et al., 2000). Larson et al. (2000) implemented a model in which the f_o error was computed as the difference between the target f_o and actual f_o (following a 130 ms processing delay), partially integrated via a low-pass filter, and applied to the output. The model simulations were compared graphically to experimental data and approximated the overall timing and shape of the observed responses. The authors acknowledged that the model was likely an over-simplification of the underlying processes but nonetheless showed promise and feasibility for computational modeling of reflexive perturbation data. The current study extends this work by investigating a variety of models that utilize different numbers of free parameters to quantitatively fit pitch shift responses measured experimentally.

Both SS and DIVA models have been successfully used to simulate responses to adaptive perturbation paradigms (Daliri and Dittman, 2019; Kearney et al., 2020). Daliri and Dittman (2019) implemented an SS model with two parameters (an internal estimate forgetting factor and a sensory error weighting factor) that showed good fits to experimental data. Kearney et al. (2020) developed SimpleDIVA–a simplified version of the DIVA model–with three parameters that correspond to gains in the key subsystems involved in speech motor control (auditory feedback, somatosensory feedback, and feedforward control). SimpleDIVA also provides good fits to experimental data and is able to account for a number of variations in the sensorimotor adaptation paradigm (e.g., perturbing more than one dimension or using masking noise). An additional benefit of SimpleDIVA is that the model's parameters provide a mechanistic explanation of behavioral responses in terms of the neural control systems believed to be involved in controlling speech production. These adaptive models, however, are not immediately applicable to reflexive response data as the mechanisms underlying the responses are not the same. Specifically, because we do not expect trial-to-trial learning in a reflexive experiment (Daliri et al., 2020; cf. Hantzsch et al., 2022), we examine the within-trial responses averaged over

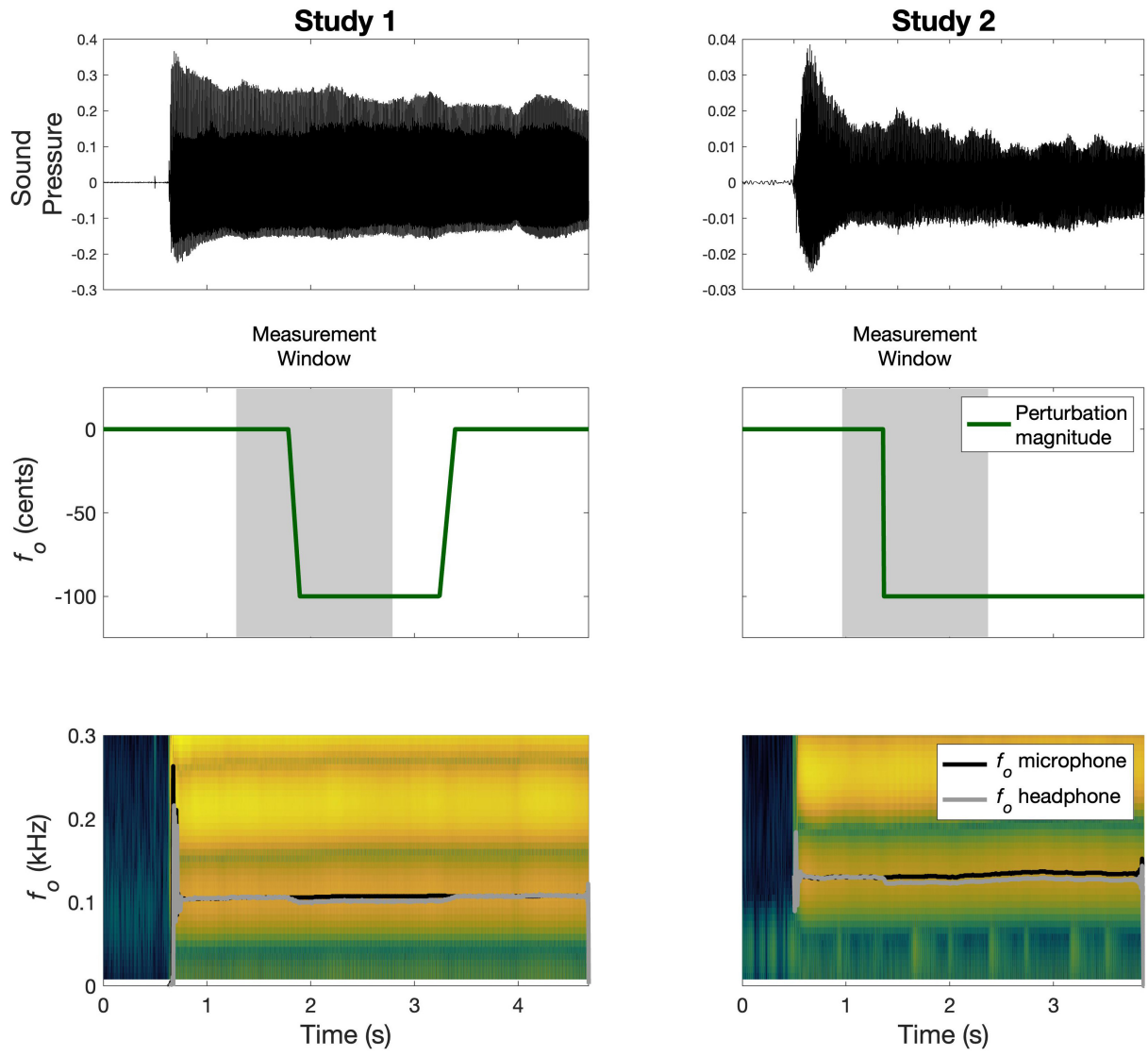


FIGURE 1
Schematic of sample perturbed trials from Study 1 and Study 2. The **top panel** shows the recorded microphone signal; the **middle panel** illustrates the time-course of the perturbation magnitude in green along with the corresponding measurement window in gray; the **bottom panel** shows the spectrogram for the microphone signal with the measured f_o traces overlaid in black (microphone) and gray (headphone).

all perturbed trials in an experiment. Examining within-trial responses also means that we need to account for latencies associated with processing delays.

Several earlier PSR studies have observed that the compensatory response could occur on more than one time scale, resulting in a complex or multi-peaked response (Burnett et al., 1997, 1998; Larson, 1998; Hain et al., 2000). The first peak was described as a short-latency, rapid response occurring around 100-225 ms, and the second as a long-latency, slow response occurring around 250–600 ms. The simplest form of the DIVA/SS model produces only a single response peak. For this reason, we also investigate generalized versions of the DIVA/SS model that are better able to capture multi-component responses.

To address our primary goal of developing a quantitative model of the PSR, we established a model framework that allows the specification of several model variations based on PID and SS/DIVA model classes. The performance of the different models was then compared by fitting them to datasets from two prior PSR studies (Heller Murray and Stepp, 2020; Smith et al., 2020). We operationally defined model validity in terms of the ability to capture population-level responses to pitch perturbation experiments as well as individual differences in sensorimotor control processes. That is, a valid model should be able to (1) explain group mean responses to pitch perturbations,

(2) have parameters that are stable across samples from an individual subject, and (3) have parameters that differentiate between individual subjects.

Materials and methods

Our overall approach is to mathematically define a number of control models that each involve optimizable parameters. Each model generates a time series of f_o values, denoted by the variable $f(t)$, where t ranges from 0 to the trial length of the experiment being modeled. A particle swarm optimization procedure is used to find the optimal parameter values [in terms of minimizing root-mean-square error (RMSE)] for each model when fitting a particular data set, and the resulting fit is characterized in terms of RMSE, Akaike information criterion (AIC), and cross-validated classification scores. Model fits were performed separately for two datasets from different studies involving unpredictable perturbations of f_o (Heller Murray and Stepp, 2020; Smith et al., 2020) applied during extended vowel productions of young healthy adult speakers.

Datasets

In Study 1 (Smith et al., 2020), a group of English speakers (N = 18; aged 18–34) completed 80 trials, during which they sustained the vowel /a/ for four seconds. On a quarter (20) of all trials, an auditory perturbation of −100 cents was applied at a jittered point in time, 1,000–1,500 ms after the beginning of the trial. The perturbation was implemented as a time-domain/formant-adjusted shift using Audapter software (Cai et al., 2008); this process shifts only f_o while preserving the produced formants. The perturbation onset was characterized by a linear ramp that took 110 ms to reach the full perturbation magnitude. The perturbation remained on for a further 1,000–1,500 ms. The order of perturbed and control trials was pseudorandomized, with no consecutively perturbed trials. f_o trajectories (Hz) were extracted for the duration of the vowel using Praat (Boersma and Weenink, 2018), and then time-aligned to the beginning of the perturbation and parsed from −500 to +1500 ms in MATLAB. A schematic of a sample perturbed trial and corresponding data is shown in **Figure 1**. The data were normalized to the average of each subject's baseline. On average, subjects compensated for 48.8% (*SD*: 20.8) of the perturbation, calculated as change from baseline to the last 250 ms of a trial and expressed as a percentage of the maximum perturbation magnitude.

In Study 2, a group of English speakers (N = 20; aged 18–28) completed 60 trials, during which they sustained the vowel /i/ for 3 s (Heller Murray and Stepp, 2020). On each trial, an auditory perturbation of +100 cents or −100 cents was applied at a jittered point in time, 500–1,000 ms after voice onset. The perturbation was implemented as a full-spectrum shift by shifting the values and spacing of the vocal harmonics using Eventide Eclipse hardware (Eventide Inc., Little Ferry, NJ, USA; Heller Murray et al., 2019), thus shifting f_o. The perturbation onset was characterized by a step function (or sudden onset) and, once applied, the perturbation remained on for the rest of the trial. All trials in the experiment were perturbed, and the direction of the perturbation was pseudorandomized to ensure that no more than five consecutive trials were perturbed in the same direction. The intertrial interval was also jittered between 500 and 1000 ms to reduce anticipation of the next trial. f_o trajectories (Hz) were extracted for the duration of the vowel using Praat (Boersma and Weenink, 2018), and then time-aligned to the beginning of the perturbation and parsed from −400 to +1400 ms in MATLAB. A schematic of a sample perturbed trial and corresponding data is shown in **Figure 1**. To fit the models to data from both perturbation directions together, all data were normalized by dividing by each subject's baseline average, and then flipping the upshift data around the x-axis. On average, subjects compensated for 17.1% (*SD*: 14.4) of the perturbation, calculated as change from baseline to the last 250 ms of a trial and expressed as a percentage of the maximum perturbation magnitude.

Assumptions and definitions for all control models

We use the variable f_T to represent the value of f_o that the controller is attempting to achieve; we assume this target is constant for a given speaker rather than a function of time since the experimental task being modeled involves attempting to maintain a constant pitch, and we equate f_T to the average f_o of the speaker prior to the onset of the perturbation (i.e., during the *baseline period* between 0 and 500 ms for Study 1 and 0 to 400 ms for Study 2). Next, we assume that the output of the controlled *plant* (corresponding to the vocal tract articulators and musculature) is updated based on the signal provided by the controller at each time point[1] as follows:

$$f(t) = f_T + \int_{\delta=0}^{t} \dot{f}_C(\delta)\, d\delta \quad (1)$$

where $f(t)$ is the plant output (i.e., the actual f_o produced by the subject) at time t, $\dot{f}_C(t)$ is the controller output at time t, and δ is a dummy variable for integration. This controller output represents a corrective command in response to the perceived error at time t. During the baseline period, $\dot{f}_C(t)$ is set to 0 for all models, and the baseline period is accordingly not included in RMSE calculations.

1 We use continuous time notation here for clarity, although the simulations utilize a discrete time representation with one time point per data sample. The data modeled here were sampled at 200 Hz; accordingly, the simulations utilize 5 ms time steps.

The auditory feedback of the produced sound available to the controller, corresponding approximately to the auditory cortical representation of the pitch/f_o of the produced sound, is defined as follows:

$$f_A(t) = f(t-\tau_A) \cdot (1+P(t-\tau_A)) \quad (2)$$

where τ_A is a delay parameter that is optimized (along with other model parameters) to fit a particular dataset, and $P(t)$ is the size of the perturbation applied at time point t, expressed as a percentage of $f(t)$ in decimal form (e.g., $P = 0.06$ corresponds to a 6% upward perturbation of f_o). The delay τ_A represents the combined delay of the perturbation processing software and hardware and the total neural processing delay from the auditory periphery to the corresponding motor output in the auditory feedback control system.

The DIVA model also includes a somatosensory representation of f_o, assumed to derive from laryngeal mechanoreceptors, which is related to the actually produced f_o as follows:

$$f_S(t) = f(t-\tau_S) \quad (3)$$

where τ_S is a delay parameter (corresponding roughly to the transmission delay from the somatosensory periphery to somatosensory cortex) that can be optimized (along with other model parameters) to fit a particular dataset. This somatosensory representation can be shown to be closely related to the parameter A in a typical state-space model, which weights the degree to which the current state of the system contributes to the next state (see *Basic DIVA equation* below).

Proportional-integral-derivative equation

A PID controller is defined by the following equation:

$$\begin{aligned}\dot{f}_C(t) &= \alpha_P \cdot (f_T - f_A(t)) + \alpha_I \\ &\cdot \int_{\delta=0}^{t} (f_T - f_A(\delta))\, d\delta + \alpha_D \\ &\cdot \frac{d}{dt}(f_T - f_A(t))\end{aligned}$$

which simplifies to:

$$\dot{f}_C(t) = \alpha_P \cdot (f_T - f_A(t)) + \alpha_I \cdot \int_{\delta=0}^{t} (f_T - f_A(\delta))\, d\delta - \alpha_D \cdot \dot{f}_A(t) \quad (4)$$

where α_P, α_I, and α_D are optimizable gains for the position, integral, and derivative terms. We will simulate four models using this equation: a proportional model (P) in which α_I and α_D are fixed at 0, a proportional-integral (PI) model in which α_D is fixed at 0, a proportional-derivative (PD) model where α_I is fixed at 0, and a proportional-integral-derivative (PID) model in which all parameters are optimized.

Basic directions into velocities of articulators/state-space equation

The DIVA model's feedback controller consists of both auditory and somatosensory feedback control components. The standard formulation of the DIVA model's feedback controller is:

$$\dot{f}_C(t) = \alpha_A \cdot (f_T - f_A(t)) + \alpha_S \cdot (f_T - f_S(t)) \quad (5)$$

where α_A and α_S are parameters denoting the gains of the auditory and somatosensory feedback control systems, respectively, and τ_S is a delay parameter corresponding to the delay between an action and the corresponding somatosensory feedback signal in somatosensory cortex. When τ_S is set to 0 [and therefore $f_S(t) = f(t)$; see EQ3], EQ5 is mathematically equivalent to the following SS model[2]:

$$\dot{f}_C(t) = A \cdot f_c(t) + B \cdot (f_T - f_A(t))$$

where $f_c(t) = f(t) - f_T$ (see Eq. 1), B is equal to α_A in EQ5, and A is equal to -α_S in EQ5. Preliminary simulations of the two models verified this mathematical equivalence and also indicated nearly identical performance for generalized versions of the DIVA/SS models described below. The model of EQ5 is also equivalent to the low-pass filter or "leaky integrator" model proposed by Larson et al. (2000), which is a special case of EQ5 with $\alpha_A = \alpha_S$ and the time constant of the low-pass filter equal to our time step size (0.005 s) times $1/\alpha_S$. For simplicity, we will use the DIVA-based formulations for simulations herein as it provides a more direct physiological interpretation of model parameters than the SS or Larson et al. (2000) formulations.

Generalized directions into velocities of articulators/state-space equations

The model of EQ5 can be generalized to include an f_o velocity target in addition to the f_o position target as follows:

$$\begin{aligned}\dot{f}_C(t) &= \alpha_A \cdot (f_T - f_A(t)) + \alpha_{Av} \cdot (\dot{f}_T - \dot{f}_A(t-\tau_{Av})) + \alpha_S \\ &\cdot (f_T - f_S(t)) + \alpha_{Sv} \cdot (\dot{f}_T - \dot{f}_S(t-\tau_{Sv}))\end{aligned}$$

2 This model is also equivalent to the SS model of sensorimotor adaptation posed by Daliri and Dittman (2019) with the parameter *a* from that model equal to 1 - α_S, parameter *b* equal to α_A, and removal of τ_A from the current model since sensorimotor adaptation data were modeled on a trial-by-trial basis rather than a timepoint-by-timepoint basis by Daliri and Ditman. The model is also equivalent to a leaky integrator of the error signal with the leak rate parameter equal to α_S.

where $\dot{f}_T$ is the target velocity, α_{Av} and α_{Sv} are the auditory and somatosensory feedback control gains of the velocity-based response component, respectively, and τ_{Av} and τ_{Sv} represent the differential delays between the position and velocity components. Because subjects in the experiments being modeled were instructed to maintain a constant pitch, $\dot{f}_T$ is set to 0 and this equation reduces to:

$$\dot{f}_C(t) = \alpha_A \cdot \left(f_T - f_A(t)\right) - \alpha_{Av} \cdot \dot{f}_A(t-\tau_{Av}) + \alpha_S \cdot (f_T - f_S(t)) - \alpha_{Sv} \cdot \dot{f}_S(t-\tau_{Sv}) \quad (6)$$

This characterization is approximately equivalent (though not identical) to a two-state (position and velocity error) SS model.

Alternatively, the model of EQ5 can be generalized to allow two different position-error-based responses that operate at different delays:

$$\dot{f}_C(t) = \alpha_A \cdot (f_T - f_A(t)) + \alpha_{As} \cdot (f_T - f_A(t-\tau_{As})) + \alpha_S \cdot (f_T - f(t)) + \alpha_{Ss} \cdot (f_T - f_S(t-\tau_{Ss})) \quad (7)$$

where α_A and α_S are the auditory and somatosensory feedback control gains of the faster response component, α_{As} and α_{Ss} are the auditory and somatosensory feedback control gains of the slower response component, and τ_{As} and τ_{Ss} represent the differential delay between the fast and slow components ($\tau_{As}, \tau_{Ss} = 0$). In effect, this model is a quantification of the idea that the response to a pitch perturbation includes a relatively fast, automatic component (captured by the terms involving α_A and α_S) and a slower component (captured by the terms involving α_{As} and α_{Ss}) that may be under more conscious control than the faster component (Burnett et al., 1997, 1998; Larson, 1998; Hain et al., 2000). This characterization is also approximately equivalent to a two-state (fast and slow position error) SS model.

Model versions used in simulations

A total of 19 different models were tested: 4 based on PID control (models P, PI, PD, and PID) and 15 based on the DIVA model and equivalent or near-equivalent state-space formulations (D1–D15). **Table 1** lists the equations and optimized parameters for all models. All unused parameters from an equation were set to 0.

Model parameter optimization

To fit a model to a particular dataset, a particle swarm optimization procedure was used to find optimized values of the free parameters of the model to fit a given dataset. The particle swarm optimization routine was chosen because it rapidly finds solutions in high-dimensional workspaces such as those utilized here and makes no assumptions regarding differentiability of the optimization problem. In this procedure, the system is initialized with a population of 10,000 random sets of parameter values ("particles") and iterated until convergence to obtain an optimized parameter set. In each iteration, all parameter sets are evaluated by computing the RMSE of their fits to the data, and a fraction of all sets is replaced by random linear combinations of those parameter sets currently producing the best fits. The procedure stops when all 10,000 parameter sets converge within a 1% range of the optimal solution or after 100 consecutive iterations without any improvement in the optimal fit to the data. When the procedure stops, the optimal parameter set among the 10,000 sets from the last iteration is selected as the solution. For each model fit, the optimization procedure was run 10 times in order to evaluate any potential residual variability due to initial conditions or local optima. The resulting parameter estimates were highly robust to the initial conditions of the swarm procedure, indicative of reaching the global minimum of the RMSE measure. The minimum-RMSE solution across all 10 repetitions was chosen as the optimized parameter set, and Pearson's *r* was calculated for this solution to characterize fit quality.

The particle swarm optimization procedure requires upper and lower bounds for the optimized parameters in order to efficiently search the parameter space. The parameter ranges for the current simulations were chosen to be big enough that they

TABLE 1 List of models included in the simulations.

Name	EQ	# Parameters	Optimized parameters
P	EQ4	2	α_P, τ_A
PI	EQ4	3	α_P, α_I, τ_A
PD	EQ4	3	α_P, α_D, τ_A
PID	EQ4	4	α_P, α_I, α_D, τ_A
D1	EQ5	3	α_A, τ_A, α_S
D2	EQ5	4	α_A, τ_A, α_S, τ_S
D3	EQ6	3	α_A, τ_A, α_{Av}
D4	EQ6	4	α_A, τ_A, α_{Av}, τ_{Av}
D5	EQ6	4	α_A, τ_A, α_S, α_{Av}
D6	EQ6	5	α_A, τ_A, α_S, τ_S, α_{Av}
D7	EQ6	6	α_A, τ_A, α_S, τ_S, α_{Av}, τ_{Av}
D8	EQ6	6	α_A, τ_A, α_S, α_{Av}, τ_{Av}, α_{Sv}
D9	EQ6	7	α_A, τ_A, α_S, τ_S, α_{Av}, τ_{Av}, α_{Sv}
D10	EQ6	8	α_A, τ_A, α_S, τ_S, α_{Av}, τ_{Av}, α_{Sv}, τ_{Sv}
D11	EQ7	4	α_A, τ_A, α_{As}, τ_{As}
D12	EQ7	6	α_A, τ_A, α_S, τ_S, α_{As}, τ_{As}
D13	EQ7	6	α_A, τ_A, α_S, α_{As}, τ_{As}, α_{Ss}
D14	EQ7	7	α_A, τ_A, α_S, τ_S, α_{As}, τ_{As}, α_{Ss}
D15	EQ7	8	α_A, τ_A, α_S, τ_S, α_{As}, τ_{As}, α_{Ss}, τ_{Ss}

did not exclude any reasonable solutions[3] but small enough to allow for relatively rapid convergence to the optimal solution. With this goal, the allowable range for all gain parameters was –0.1 to 1.1 with the exception of α_I in the PI and PID models, which used a range of –0.001 to 0.001 (the α_I parameter corresponds to the gain of the auditory error integral, which determines how much the corrective response increases as the error accumulates over the duration of the perturbation; preliminary simulations resulted in very tiny values for this parameter that did not always stabilize when using the larger range). A negative gain indicates a response that exacerbates, rather than corrects, the corresponding error; the negative gains allow us to model following responses. A gain of 1 corresponds to immediate full compensation for the corresponding error; gains significantly above 1 are therefore prone to instabilities and highly unlikely to represent optimal solutions. Delay parameters were limited to 0–500 ms except for the differential delays τ_{Av} and τ_{Sv}, which were limited to –100 to 500 ms to allow for the possibility that the velocity error response is faster than the position error response. Preliminary simulations indicated that none of the optimized parameters were at one of the ends of the allowable range for any model; in other words, solutions were not artificially limited by the chosen bounds.

Akaike information criterion calculations

Because adding more parameters will inevitably improve RMSE (even to the point of overfitting the data), for model comparisons we focus on AIC, which is designed to meaningfully compare models with different numbers of free parameters using the information theoretic criterion of minimum information loss[4]. The AIC for each model is defined by the equation $AIC = 2k - 2ln(L)$, where k is the number of free parameters in the model and L is the maximum likelihood of the model. We estimated the optimal model parameters for each model by minimizing the residual mean square error between the model fit and the observed traces. Assuming that the trace residuals were normally distributed but potentially correlated across timepoints, the model log-likelihood could be approximated as $ln(L) = N/2\left(-ln(MSE) - 1 - ln(2\pi)\right)$, where MSE is the mean square error of the model, and N is the effective degrees of freedom of the trace residuals (equal or smaller to the number of samples in the data). The degrees of freedom were computed using Satterthwaite–Welsh approximation (Satterthwaite, 1946) from the observed autocorrelation of the data before the onset of the perturbation (common to all models). Last, in order to facilitate comparisons of the resulting AIC measures across different datasets or with different studies, we reported corrected-AIC measures, dividing AIC by the data's effective degrees of freedom, leading to the combined equation:

$$cAIC = AIC/N = 2k/N + ln(MSE) + 1 + ln(2\pi) \quad (8)$$

When comparing two models, the relative likelihood of the two models can be computed from the difference in AIC values as $exp((AIC_{min} - AIC)/2)$. To identify statistically significant differences in cAIC, we calculated the cAIC threshold necessary to support a 20:1 relative likelihood between the two models using the formula:

$$thr_{cAIC} = 2ln(1/0.05)/N \quad (9)$$

A model whose cAIC is less than another model's cAIC by more than this threshold is, with 95% likelihood, the superior model.

Cross-validated classification simulations

The last set of simulations further tested the models' abilities to characterize stable properties of each subject by optimizing the models using a subset of data from each subject (training trials) and then testing the models on the remaining trials (test trials). Specifically, for each model and each subject, 10 cross-validation iterations were performed, each involving a different random subset of 10 test trials (from a total of 13–20 trials per subject in Study 1 and 19–57 trials per subject in Study 2) used for testing, with the remaining trials for that subject used as the training set for optimizing model parameters (i.e., model parameters were optimized to fit the average trace of the training trials). The optimized model was then compared to the test trials to compute a combined cAIC value, using the same cAIC formula above as in the individual-trace analyses but setting MSE to the average of the MSE values across all of the individual test trials, and setting the data samples N to the average effective degrees of freedom across all the test trials multiplied by the total number of trials for that model/subject combination. This led to a single cAIC value for each model and each subject, characterizing the model's ability to predict the behavior of out-of-sample trials for an individual subject. The average of these cAIC values was then calculated across subjects for each model.

In addition, we wanted to evaluate, for each model, whether a subject's model parameter values could be used to uniquely

3 For example, it does not make much sense within the DIVA model for the auditory feedback gain to be less than 0 (which would exacerbate rather than correct auditory errors) or greater than 1 (which would overcompensate for auditory errors). The bounds used here are slightly larger than these to allow for random variation that may occur in any particular dataset.

4 We chose AIC here over the closely related Bayesian Information Criterion (BIC) because we anticipate that the model training datasets will generally be small; in such cases BIC tends to choose models that are too simple due to its use of a stronger penalty term for the number of model parameters (Burnham and Anderson, 2002).

identify this subject's traces from different trials compared to the traces of other subjects. The models' abilities to correctly identify a subject were assessed from these same cross-validation iterations by first computing RMSE values comparing the average traces of one subject's test trials to the model traces obtained from fitting the training trials of the same (or a different) subject. From these comparisons we then determined *overall* and *pairwise* classification scores for each model, from a classification procedure that chose the subject with minimal RMSE as the most likely subject to have generated that mean test trace. All classification scores represent the percent correct identifications of a subject based on the mean of 10 test trials, averaged across the 10 cross-validation iterations and all appropriate between-subject comparisons. The *overall* classification scores represent the percentage of times the correct subject (i.e., the one who generated the test trials) had the lowest RMSE when compared to all other subjects for that same model, and they were computed as:

$$p_{overall} = \frac{1}{10N}\sum_{m=1}^{N}\sum_{i=1}^{10}\prod_{n\neq m}^{N}\frac{1}{10}\sum_{j=1}^{10}\left[RMSE_{imim} < RMSE_{imjn}\right]$$

where N is the number of subjects, and $\mathrm{RMSE}_{i,m,j,n}$ represents the RMSE value obtained when comparing the mean trace from the test trials of the i-th cross-validation iteration of subject m to the model traces obtained from fitting the training trials of the j-th cross-validation iteration of subject n. The *overall* classification scores for each model are reported in the "Overall" columns of **Table 2**. Study 1 involved 18 subjects and Study 2 involved 20, so chance performance on the classification task was 5.6% for Study 1 and 5% for Study 2.

Pairwise classification scores represent the percentage of times the correct subject had lower RMSE than another (randomly selected) incorrect subject for that same model, and they were computed as:

$$p_{pairwise} = \frac{1}{10N(N-1)}\sum_{m=1}^{N}\sum_{i=1}^{10}\sum_{n\neq m}^{N}\frac{1}{10}\sum_{j=1}^{10}\left[RMSE_{imim} < RMSE_{imjn}\right]$$

Classification accuracies for a given model were averaged across all pairs of subjects to obtain the scores listed in the "Pairwise" columns of **Table 2**; chance performance on this classification task is 50%.

Additionally, intraclass correlation coefficients (ICC) were calculated to quantify the reliability/stability of model parameters across the 10 cross-validation iterations. ICC values were calculated as:

$$ICC = \frac{{\sigma_{between}}^2}{{\sigma_{between}}^2+{\sigma_{within}}^2}$$

where $\sigma_{between}$ is the between-subject standard deviation for a given parameter and σ_{within} is the within-subject standard deviation for a given parameter. ICC values fall between 0 and 1, with values < .5 indicating poor reliability, values 0.5–0.75, 0.75–9, and 0.9–1 indicating moderate, good, and excellent reliability, respectively (Koo and Li, 2016).

Results

Table 2 summarizes the fit statistics for all models and simulations. The following subsections describe these results by simulation set: fits to study group means, fits to individual subject means, and cross-validated classification simulations.

Fits to group means

The group mean trace for each study was formed by first calculating the mean f_o value at every time point for each individual subject (averaged across all that subject's trials), then averaging these individual subject means to form the group mean trace. The group mean traces are indicated by the solid blue lines in **Figure 2A** (Study 1) and **Figure 3A** (Study 2), with standard error of the mean (SEM) indicated by blue shading. Full compensation, the inverse of the perturbation magnitude, is shown in green. Full compensation illustrates what a 100% compensation for the perturbation would look like, although this is rarely achieved in auditory perturbation studies.

The columns labeled "Group" in **Table 2** indicate the RMSE for each model's fit to the group mean trace as well as the cAIC value resulting from comparing the model fit to the individual subject mean traces. The lowest RMSE and cAIC values for each study are indicated in boldface. Blue shading indicates cAIC values that are within the cAIC threshold of the lowest cAIC value; in other words, the models with no shading are inferior to the best (boldfaced) model according to the cAIC criterion, whereas the models with blue shading are not significantly different (at the $p < 0.05$ false positive level) from the best model. For both studies, the three-parameter model D1 provided the best fit according to cAIC, with the three-parameter model PI also falling within the cAIC threshold, along with several four-parameter models (PID, D2, D5, and D11), a five-parameter model (D6), and several six-parameter models (D7, D8, D11, D12, and D13). For the remainder of this article, we will refer to models within the cAIC threshold of the best model collectively as the "best models."

When multiple models fall within the AIC threshold of the top model, there is not enough empirical evidence to support the selection of an individual model among them. In these cases, and until more evidence becomes available, it is reasonable to give preference to the model with the fewest parameters amongst these models. Thus, according to the cAIC criterion, the models providing the best fits to the group mean data are the three-parameter models D3 and PI, followed by the 4-parameter

TABLE 2 Fit statistics for all simulated models.

	Study 1 group		Study 2 group		Study 1 subject		Study 2 subject		Study 1 Xval classification			Study 2 Xval classification		
Model	**RMSE**	**cAIC**	**RMSE**	**cAIC**	**RMSE**	**cAIC**	**RMSE**	**cAIC**	**cAIC**	**Overall**	**Pair**	**cAIC**	**Overall**	**Pair**
P	0.00312	−6.33001	0.00198	−7.1133	0.00381	−6.1104	0.00237	−5.4655	−1.6882	23.39%	84.23%	−1.3595	18.25%	72.11%
PI	0.00083	−6.41743	0.00073	−7.1847	0.00201	−6.1983	0.00130	−5.4930	−1.7201	35.94%	88.10%	−1.3708	24.14%	74.59%
PD	0.00312	−6.32608	0.00197	−7.1110	0.00355	−6.1208	0.00234	−5.4644	−1.6870	26.82%	85.53%	−1.3587	18.21%	72.11%
PID	0.00065	−6.41588	0.00071	−7.1820	0.00158	−6.2102	0.00120	−5.4926	−1.7206	38.76%	88.77%	−1.3701	24.16%	74.58%
D1	0.00059	**−6.42106**	0.00039	**−7.1932**	0.00187	−6.2049	0.00112	−5.4950	−1.7228	35.70%	88.15%	−1.3713	22.69%	73.66%
D2	0.00059	−6.41672	0.00039	−7.1901	0.00174	−6.2023	0.00100	−5.4960	−1.7187	35.72%	88.28%	**−1.3715**	24.72%	74.68%
D3	0.00312	−6.32608	0.00197	−7.1110	0.00355	−6.1208	0.00234	−5.4644	−1.6870	26.82%	85.53%	−1.3587	18.21%	72.11%
D4	0.00116	−6.40619	0.00083	−7.1780	0.00183	−6.1964	0.00145	−5.4848	−1.7157	37.75%	88.42%	−1.3669	22.04%	73.25%
D5	0.00053	−6.41737	0.00026	−7.1921	0.00151	−6.2129	0.00105	−5.4941	**−1.7233**	38.55%	88.74%	−1.3706	23.07%	73.78%
D6	0.00052	−6.41321	0.00021	−7.1895	0.00137	−6.2106	0.00092	−5.4951	−1.7195	38.73%	88.89%	−1.3708	25.56%	74.83%
D7	0.00046	−6.40946	0.00017	−7.1867	0.00113	−6.2103	0.00081	−5.4942	−1.7147	39.49%	89.00%	−1.3697	25.65%	74.93%
D8	0.00052	−6.40886	0.00017	−7.1866	0.00117	−6.2093	0.00092	−5.4918	−1.7142	38.53%	88.78%	−1.3688	23.70%	73.94%
D9	0.00052	−6.40451	0.00017	−7.1835	0.00103	−6.2065	0.00076	−5.4930	−1.7110	40.20%	89.18%	−1.3689	26.05%	75.06%
D10	0.00048	−6.40055	0.00013	−7.1806	0.00100	−6.2014	0.00071	−5.4922	−1.7070	40.37%	89.14%	−1.3679	26.12%	75.13%
D11	0.00104	−6.40902	0.00040	−7.1900	0.00147	**−6.2129**	0.00097	**−5.4964**	−1.7227	38.57%	88.93%	−1.3713	25.38%	75.08%
D12	0.00051	−6.40890	0.00027	−7.1857	0.00111	−6.2110	0.00078	−5.4948	−1.7129	39.06%	88.94%	−1.3699	25.70%	75.24%
D13	0.00051	−6.40889	0.00027	−7.1857	0.00116	−6.2099	0.00082	−5.4944	−1.7160	39.92%	89.03%	−1.3698	25.07%	75.15%
D14	0.00051	−6.40454	0.00013	−7.1838	0.00105	−6.2064	0.00076	−5.4933	−1.7110	**40.71%**	89.18%	−1.3690	25.01%	75.16%
D15	0.00035	−6.40165	0.00012	−7.1807	0.00093	−6.2027	0.00066	−5.4924	−1.7058	40.57%	**89.22%**	−1.3682	**26.32%**	**75.27%**

Xval, cross-validation; RMSE, root-mean-square error; cAIC, corrected Akaike information criterion; Overall, overall accuracy (%); Pair, pairwise accuracy (%).
Boldface type indicates the model with lowest RMSE and cAIC for each study.
Blue shading indicates cAIC values that are within the cAIC threshold of the lowest cAIC value.

10.3389/fnhum.2022.929687

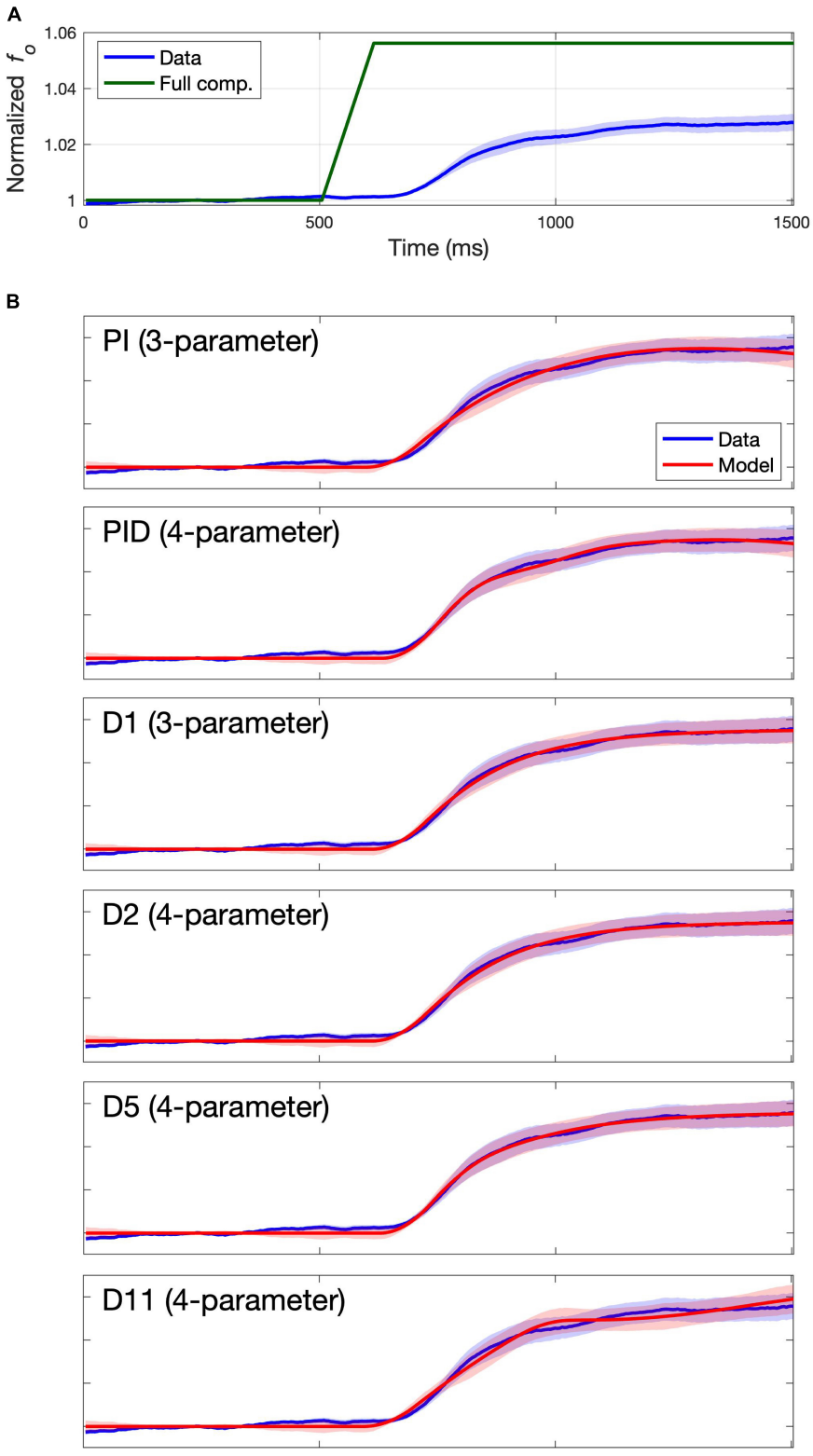


FIGURE 2
Group mean data and model fits for Study 1. Group mean data and standard error of the mean are shown with a blue line and shading. **(A)** Group mean data shown relative to full compensation in green. Full compensation is the inverse of the perturbation magnitude and illustrates what 100% compensation would look like. **(B)** Group mean data shown relative to model fit (red line) and standard error of the model fit (red shading) for models providing best fits to the group mean data.

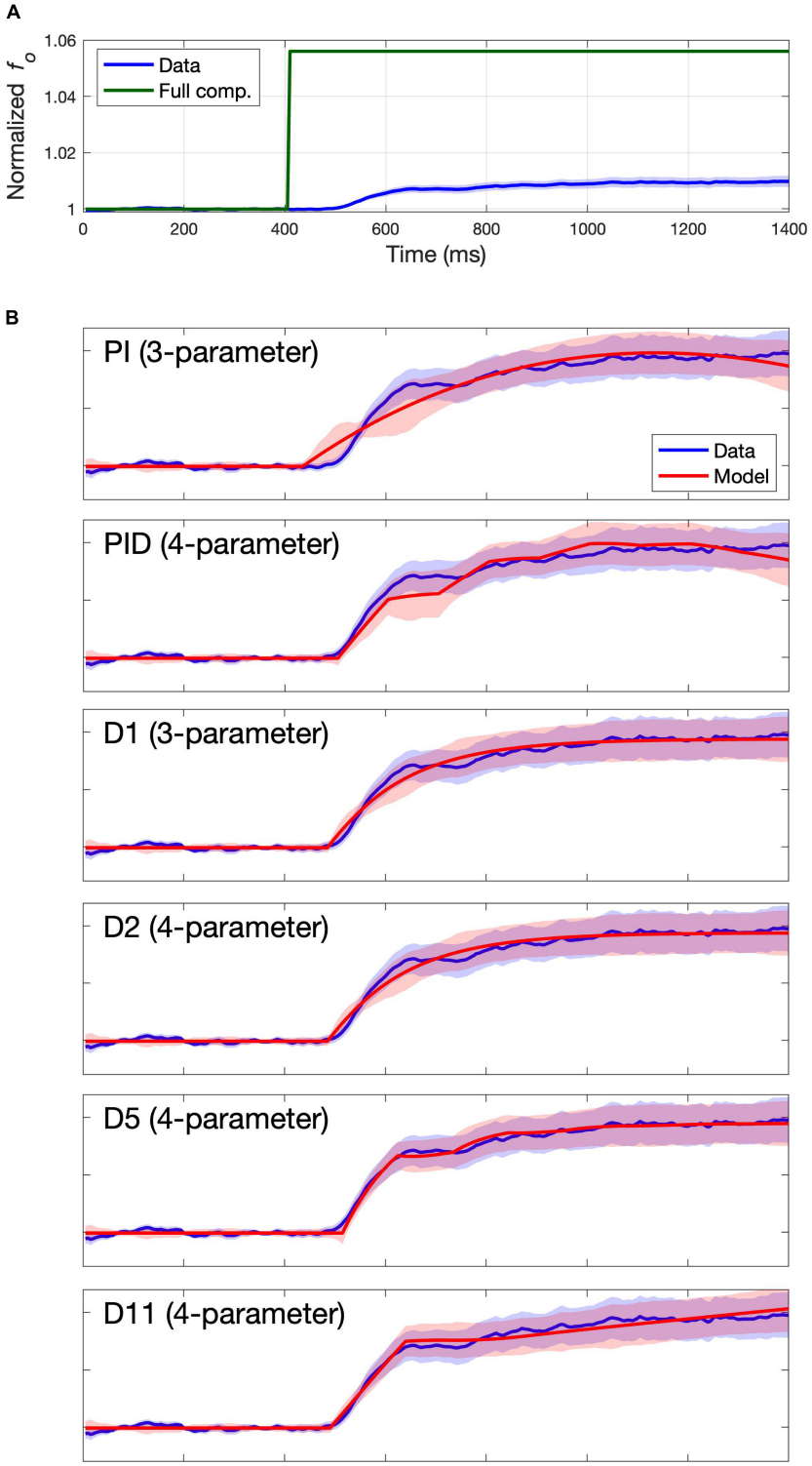


FIGURE 3
Group mean data and model fits for Study 2. Group mean data and standard error of the mean are shown with a blue line and shading. **(A)** Group mean data shown relative to full compensation in green. Full compensation is the inverse of the perturbation magnitude and illustrates what 100% compensation would look like. **(B)** Group mean data shown relative to model fit (red line) and standard error of the model fit (red shading) for models providing best fits to the group mean data.

models PID, D2, D5, and D11. **Figures 2B**, **3B** plot the model fit (red line) and standard error of the model fit (red shading), along with the group mean (blue line) and SEM (blue shading) of the experimental data for these models. (Model fits for all models are provided in the **Supplementary materials**). Despite falling within the cAIC threshold of the best model D1, the PI and PID models produce traces that poorly match the overall shape of the data trace, casting doubt as to whether they are effectively capturing the physiological mechanisms responsible for the subjects' productions. This is particularly clear from the Study 2 fits in **Figure 3**. The four-parameter model D5 appears to best capture the overall shape of the data trace, which shows an initial plateau followed by a second rise approximately 100 ms after the start of the plateau (again more clearly visible in **Figure 3**). As noted in the *Introduction*, a two-component response pattern has been noted in prior pitch perturbation experiments (e.g., Larson, 1998; Hain et al., 2000). Models D1, D2, and D11 capture the overall shape of the data trace reasonably well, but they fail to properly capture the shape of the plateau and second rise.

In sum, the three-parameter model D1 provides the best fit of the group mean traces according to the cAIC criterion while also capturing the overall shape of the experimentally measured response reasonably well. The four-parameter model D5 best captures the overall shape of the data traces amongst the three- and four-parameter models and falls within the cAIC threshold of model D1. The additional parameters of models with more than four parameters appear to provide little additional improvement.

The optimized values of all parameters for all models are provided in the **Supplementary materials**. For the basic DIVA best models (D1 and D2), the parameter values were very similar between models for a given dataset. For Study 1, the mean values (across the two models) were 0.011 for α_A, 0.013 for α_S, 115 ms for τ_A, and 130 ms for τ_S. For Study 2, they were 0.006 for α_A, 0.033 for α_S, 93 ms for τ_A, and 54 ms for τ_S. These parameters had similar values in the generalized DIVA best models (D5–D8 and D11–D13), whereas the additional parameters in the generalized DIVA models were considerably more variable across models.

Fits to individual subjects

The second set of simulations compared the models on their ability to fit individual subject data using parameters optimized for the individual subject rather than the group mean. These simulations gauge how well the models can account for individual differences through subject-specific parameterizations. For each subject, model parameters were optimized to fit the subject's mean trace (averaged across trials). The RMSE values of these fits are provided in the columns labeled "Subject" in **Table 2**, along with the cAIC values resulting from comparing the models' fits to the individual subject mean traces. With the exception of models P, PD, D3, and D4, all models fell within the cAIC threshold of the best model (D11 for both studies).

Cross-validated classification simulations

The columns labeled "Xval Classification" in **Table 2** provide cAIC, overall classification accuracy, and pairwise classification accuracy for each model in each study. The models within the cAIC threshold of the best cAIC value for both studies were models PI, PID, D1, D2, D5–D7, and D11–D13. The highest overall classification accuracies were 40.71% for model D14 in Study 1 (chance level of 5.6%) and 26.32% for model D15 in Study 2 (chance level 5%). Even the worst-performing models had overall accuracies that were well above chance: 23.39% for model P in Study 1 and 18.21% for models PD and D3 in Study 2. Pairwise classification accuracies were also well above chance (50%) for all models, ranging from 84.23% (model D1) to 89.22% (model D15) for Study 1, and from 72.11% (models P and D1) to 75.27% (model D15) in study 2.

Overall, these results indicate that reflexive responses to f_o perturbations are largely individual-specific, and a number of models perform nearly equivalently on the cross-validated classification tasks. For comparison, we also calculated cross-validated classification accuracy when we used the mean of the training trials for classification rather than one of the models. This resulted in overall and pairwise accuracies of 38.14 and 89.25%, respectively, for Study 1 and 25.89 and 75.38% for Study 2. These are similar to values obtained for the best-performing models in **Table 2**.

The cross-validation training iterations also provide information regarding the stability of model parameters across the 10 iterations for a given subject. In other words, do the 10 iterations yield approximately the same values for a given parameter (as would be expected if the parameter has a reliable physiological basis) or do they vary substantially across iterations (indicative of a model whose parameters do not have a reliable physiological interpretation)? To assess this, we calculated ICC for each parameter in each model for each data set. The mean parameter values and ICC values from the 10 cross-validation iterations are provide in **Table 3** (Study 1) and **Table 4** (Study 2). Boldface type indicates the model with the highest ICC value per parameter. Dark blue shading indicates ICC values greater than 0.75 (corresponding to good reliability), and light blue shading indicates ICC values between 0.5 and 0.75 (moderate reliability).

Generally speaking, parameter stability was higher for the PID-based models and models D1–D10 compared to models D11–D15. In particular, all PID models and models D1–D9 had highly reliable values for the auditory feedback control

TABLE 3 Study 1 mean values and ICC of optimized parameters in cross-validation simulations.

	α_P/α_A		α_S		$\alpha_D/\alpha_{Av}/\alpha_{As}$		$\alpha_I/\alpha_{Sv}/\alpha_{Ss}$		τ_A		τ_S		τ_{Av}/τ_{As}		τ_{Sv}/τ_{Ss}	
Model	**Mean**	**ICC**	**Mean**	**ICC**	**Mean**	**ICC**	**Mean**	**ICC**	**Mean**	**ICC**	**Mean**	**ICC**	**Mean**	**ICC**	**Mean**	**ICC**
P	0.005	**0.956**							0.044	0.647						
PI	0.009	0.939					−4.9E-05	0.826	0.109	0.660						
PD	0.010	0.924			0.923	0.592			0.109	0.601						
PID	0.013	0.918			0.551	0.414	−1.0E-04	**0.842**	0.135	0.594						
D1	0.012	0.851	0.016	0.705					0.128	0.586						
D2	0.010	0.870	0.013	0.657					0.111	0.556	0.159	**0.470**				
D3	0.010	0.924			0.923	0.592			0.109	0.601						
D4	0.008	0.941			0.788	0.688			0.101	**0.690**			0.274	**0.582**		
D5	0.015	0.852	0.018	**0.772**	0.393	0.328			0.143	0.536						
D6	0.012	0.903	0.014	0.747	0.568	0.335			0.128	0.568	0.226	0.337				
D7	0.011	0.851	0.009	0.650	0.731	0.662			0.118	0.534	0.203	0.369	0.162	0.455		
D8	0.014	0.827	0.011	0.699	0.789	0.468	0.170	0.317	0.123	0.502			0.218	0.469		
D9	0.012	0.823	0.008	0.576	0.648	0.565	0.403	0.412	0.124	0.500	0.152	0.419	0.240	0.350		
D10	0.012	0.836	0.008	0.619	0.734	0.644	0.444	0.296	0.123	0.488	0.143	0.318	0.201	0.257	0.122	**0.268**
D11	0.013	0.662			−0.010	0.584			0.120	0.613			0.296	0.510		
D12	0.018	0.545	0.069	0.657	0.013	**0.694**			0.110	0.397	0.080	0.235	0.307	0.358		
D13	0.033	0.592	0.123	0.762	0.020	0.686	0.573	0.206	0.120	0.354			0.300	0.396		
D14	0.022	0.448	0.070	0.611	0.014	0.575	0.618	0.472	0.119	0.391	0.078	0.226	0.319	0.309		
D15	0.015	0.666	0.035	0.555	0.011	0.607	0.697	0.440	0.114	0.380	0.052	0.277	0.287	0.301	0.185	0.264

Only one parameter listed per column is optimized in a given model. For example, the PID models have an α_P parameter whereas the DIVA models have an α_A parameter. See Table 1 for a complete list of parameters included in each model.
ICC, intraclass correlation coefficient.
Boldface type indicates the model with the highest ICC value per parameter.
Light blue shading indicates ICC values are between 0.5 and 0.75 (moderate reliability).
Dark blue shading indicates ICC values are > 0.75 (good-excellent reliability).

10.3389/fnhum.2022.929687

TABLE 4 Study 2 mean values and ICC of optimized parameters in cross-validation simulations.

	α_P/α_A		α_S		$\alpha_D/\alpha_{Av}/\alpha_{As}$		$\alpha_I/\alpha_{Sv}/\alpha_{Ss}$		τ_A		τ_S		τ_{Av}/τ_{As}		τ_{Sv}/τ_{Ss}	
Model	**Mean**	**ICC**	**Mean**	**ICC**	**Mean**	**ICC**	**Mean**	**ICC**	**Mean**	**ICC**	**Mean**	**ICC**	**Mean**	**ICC**	**Mean**	**ICC**
P	0.001	0.969							0.031	0.761						
PI	0.003	**0.972**					−2.6E-05	**0.955**	0.062	0.810						
PD	0.003	0.969			1.019	0.597			0.055	0.793						
PID	0.005	0.954			0.811	0.678	−3.4E-05	0.947	0.091	**0.823**						
D1	0.009	0.881	0.098	0.725					0.110	0.746						
D2	0.005	0.896	0.035	0.543					0.088	0.746	0.109	0.744				
D3	0.003	0.969			1.019	0.597			0.055	0.793						
D4	0.002	0.969			0.972	0.689			0.057	0.800			0.359	**0.747**		
D5	0.010	0.912	0.078	0.780	0.439	0.633			0.122	0.726						
D6	0.006	0.935	0.037	0.623	0.622	0.612			0.101	0.755	0.144	**0.758**				
D7	0.006	0.927	0.034	0.792	0.663	**0.762**			0.098	0.746	0.133	0.552	0.175	0.601		
D8	0.008	0.910	0.038	0.651	0.719	0.641	0.005	0.380	0.110	0.734			0.193	0.537		
D9	0.006	0.942	0.027	0.785	0.524	0.704	0.403	0.480	0.099	0.751	0.136	0.456	0.259	0.374		
D10	0.006	0.946	0.027	**0.797**	0.539	0.646	0.520	0.483	0.098	0.718	0.114	0.441	0.265	0.293	0.189	0.368
D11	0.004	0.702			−0.004	0.702			0.089	0.583			0.228	0.647		
D12	0.005	0.666	0.053	0.700	0.002	0.430			0.097	0.741	0.072	0.353	0.322	0.562		
D13	0.008	0.642	0.083	0.607	0.004	0.536	0.226	0.423	0.103	0.686			0.306	0.502		
D14	0.006	0.646	0.052	0.536	0.002	0.504	0.510	0.303	0.097	0.663	0.075	0.317	0.333	0.489		
D15	0.006	0.726	0.046	0.547	0.002	0.584	0.607	0.660	0.100	0.767	0.054	0.405	0.324	0.543	0.221	**0.480**

Only one parameter listed per column is optimized in a given model. For example, the PID models have an αP parameter whereas the DIVA models have an αA parameter. See Table 1 for a complete list of parameters included in each model.
ICC, intraclass correlation coefficient.
Boldface type indicates the model with the highest ICC value per parameter.
Light blue shading indicates ICC values are between 0.5 and 0.75 (moderate reliability).
Dark blue shading indicates ICC values are > 0.75 (good-excellent reliability).

10.3389/fnhum.2022.929687

gain parameter (α_P in PID models and α_A in DIVA-based models) and moderately to highly reliable values for the auditory feedback control delay parameter in both studies. The somatosensory feedback control gain parameter was also moderately to highly reliable in all DIVA-based models (D1–D15) in both studies, and the parameter α_I was highly reliable in the PI and PID models in both studies.

Discussion

The primary goal of this study is the identification of a model that captures population responses in auditory perturbation experiments, and perhaps more importantly characterizes individual differences in a stable manner with parameters that relate to underlying motor control capabilities. The latter capability is particularly important if the model is to be used to characterize individuals with communication disorders for the purpose of providing individualized treatments that capitalize on the individual's strengths and weaknesses. For this approach to bear fruit, it is important that the behavioral responses exhibited by experimental subjects are reasonably stable and differ between individuals; if not, then no model will be capable of achieving our goal. A key finding from the current study (independent of any modeling) is that reflexive responses to f_o perturbations are largely individual-specific, providing optimism that such responses may reveal key insights into the individual's speech motor control processes. Although all subjects were healthy adults with no communication disorders (and therefore likely to have somewhat similar speech motor systems, in contrast to individuals with a speech disorder), the cross-validation classification analyses indicate that the mean of 10 reflexive responses from an individual is enough to distinguish that individual from another neurotypical individual with approximately 90% accuracy in Study 1 (see *pair* column in *Study 1 Xval Classification* section of **Table 2**) and 75% in Study 21 (see *pair* column in *Study 2 Xval Classification* section of **Table 2**). This highlights a rather remarkable property of the PSR independent of any modeling: an individual's pitch shift response is akin to a "fingerprint" that largely distinguishes them from other individuals (though not to the degree of an actual fingerprint). We expect that individuals with speech motor disorders will show much greater variability than our current healthy sample and therefore may be easier to distinguish based on their reflexive responses; verification of this expectation is an important topic for future research.

The three-parameter model D1 provided fits to group mean data with the lowest cAIC values of any model for both Study 1 and Study 2. Furthermore, this model was within the cAIC threshold of the lowest cAIC for individual subject fits and cross-validation simulations for both studies. Other models that fell into the best model category (i.e., those within the cAIC threshold of the lowest cAIC value) for all simulations were PI, D2, D5, D6, D7, D11, D12, and D13. Model D1 also had amongst the most stable parameters across cross-validation iterations as measured by ICC (see **Table 3**), and its pairwise classification scores were within 1–2% of the best-performing model.

Concerning the three-parameter model PI, although this model performed well according to the cAIC, parameter stability, and cross-validated classification criteria, the overall shape of the responses of the PI model differed considerably from the shape of the subject responses (compare the fits of models PI and D1 in **Figures 2B**, **3B**). The anomalous response shape for the PI model is the result of the fact that the optimized values for the parameter α_I, which determines how much the corrective response increases as error accumulates, were negative (see **Tables 3**, **4**), indicating that the correction actually *decreased* with accumulating error. This is contrary to the theoretical motivation for this term (which is to increase the correction if the error keeps accumulating) and results in the "inverted U" shape of the PI model responses in **Figures 2**, **3** that is not found in the data traces nor in model D1. It is also worth noting that fixing α_I at 0 so it will not go negative reduces the PI and PID models to the poorly performing P and PD models, respectively.

For these reasons, we conclude that the best 3-parameter model for characterizing reflexive responses to f_o perturbations is D1 (EQ5), which has free parameters α_A, τ_A, and α_S. These parameters have straightforward interpretations: α_A (which corresponds to the parameter B in a state-space formulation—see *Basic DIVA/SS equation* in the "*Materials and Methods*" section) is the gain of the auditory feedback controller's response to a perceived error, τ_A is the delay of this response, and α_S is the gain of the "resistance" to this correction. Within the DIVA model, this latter parameter corresponds to the gain of the somatosensory feedback controller, which is attempting to keep f_o (as detected through somatic sensation, which is not perturbed in the current experiment) at the target level. α_S is related to the parameter A in a state-space formulation (specifically, $A = -\alpha_S$); this parameter similarly acts to resist changes due to perceived auditory error, though it is not typically specifically associated with somatosensory feedback control. Model D1 is also equivalent to a low-pass filter/leaky integrator model, as proposed by Larson et al. (2000).

A more general interpretation of α_S, which is consistent with both the DIVA and state-space formulations is that it reflects the influence of non-auditory-based motor subsystems on the overall motor output. This can include both feedforward control mechanisms and somatosensory feedback control mechanisms. Indeed, the estimate of the somatosensory state in DIVA is envisioned as a combination of an efference copy of the motor command (which provides a predictive estimate of

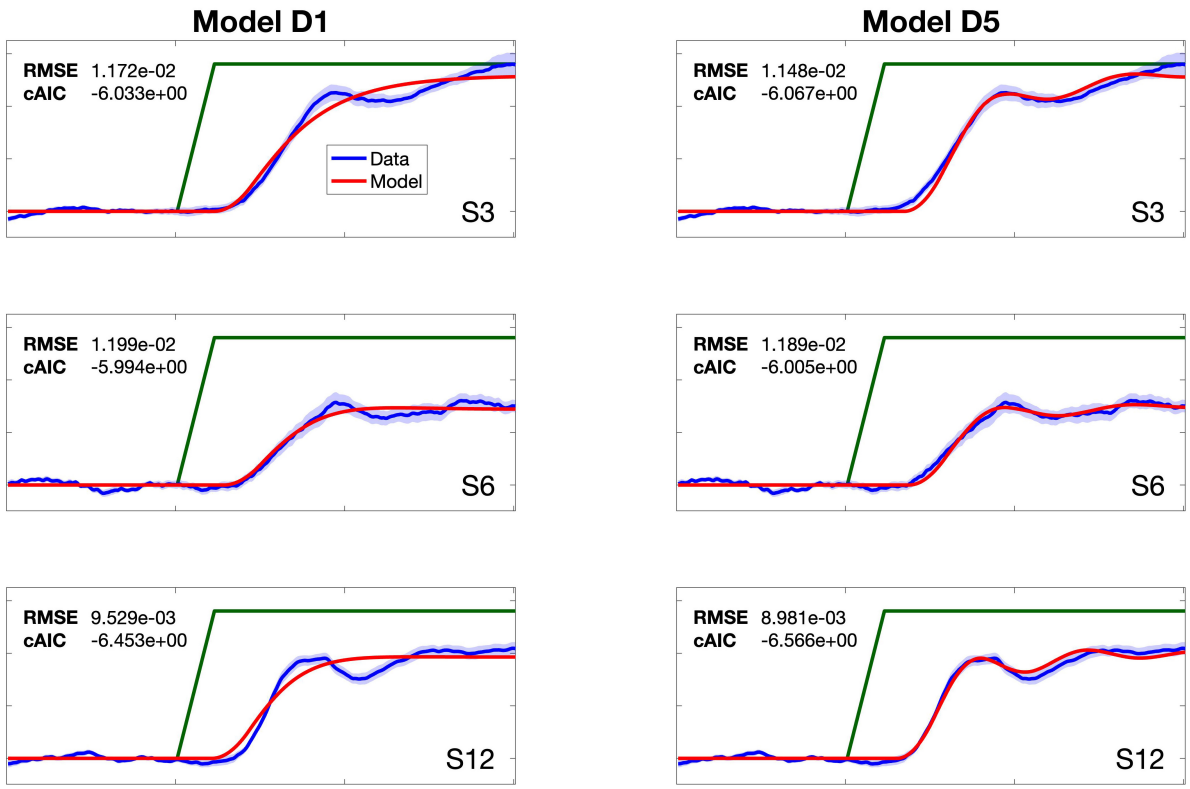


FIGURE 4
Comparison of the fits for models D1 and D5 for three subjects (S3, S6, S12) from Study 1 who showed multi-component response profiles. Subjects' mean data traces and standard error of the mean are shown with blue line and shading, respectively. Model fits shown with red line.

somatosensory state) and incoming somatosensory information (see for example **Figure 1** in Guenther et al., 1998). The use of a predictive estimate of the sensory state within a sensory feedback control architecture (see also Houde and Nagarajan, 2011) is, in essence, a form of feedforward control since it does not depend on sensory feedback for generating control signals.

Amongst the four-parameter models (PID, D2, D4, D5, and D11), models D5 and D11 were within the cAIC threshold of the lowest cAIC for all simulations for both studies (shaded cells in **Table 2**), and both of these models exhibited relatively high parameter stability (**Table 3**). Of these two models, D5 produced fits that better captured the overall shapes of the response profiles (**Figures 2**, **3**). Although the cAIC values for models D5 and D11 were in no cases significantly better than the 3-parameter model D1, it is noteworthy that models D5 and D11 (as well as most of the models with five or more parameters) are better capable of accounting for multi-component response profiles. This is illustrated in **Figure 4**, which compares the fits of models D1 and D5 to individual subjects from Study 1 who exhibited multi-component responses. Multi-component responses have also been reported in several prior PSR studies (Burnett et al., 1997, 1998; Larson, 1998; Larson et al., 2000; Hain et al., 2000), and it appears that the second response component is under more conscious control than the earlier "automatic" component; for example, the second component is much more influenced by instructions provided to subjects regarding whether they should attempt to oppose or follow the perturbation direction (Hain et al., 2000). The 4th parameter in model D5 is an auditory velocity error gain, α_{AV}. This term has the effect of resisting any perceived changes in pitch (beyond the abrupt change at perturbation onset, which is ignored by the model), in keeping with the fact that subjects are attempting to maintain a constant pitch, as they were instructed to do in the studies modeled here.

Despite D5 better capturing multicomponent responses, D5 was not superior to D1 according to the cAIC criterion in any of the simulations; in other words, the reduction in RMSE afforded by the 4th parameter in D5 was offset by the AIC penalty term for increasing the number of model parameters by 1. This suggests that the secondary responses, which are better characterized by D5, are quite variable compared to the primary response, which is captured well by both D1 and D5. While the later components could be more influenced by cognitive variables such as attention level and conscious intent

(Burnett et al., 1997, 1998; Hain et al., 2000), modeling the contribution of those processes was beyond the scope of the current study. Additional parameters beyond 4 provide little additional improvement.

It is reasonable to wonder what is gained from characterizing and individual's reflexive responses to f_o perturbations with a parameterized model, given that the average of a set of training traces provides classification results that are on par with the best model characterizations. The key difference is that *a model whose parameters correspond to physiological motor control processes provides a quantitative assessment of an individual's motor speech capabilities.* For example, a past pitch perturbation study involving individuals with Parkinson's disease indicated greater compensation than age-matched controls (Liu et al., 2012). By itself, this observation is of limited value for characterizing the motor control processes of an individual with Parkinson's disease since a larger response might indicate enhanced auditory feedback control or, alternatively, degraded somatosensory feedback control. In contrast, the optimal fit of model D1 to the subject's response traces provides values of α_A and α_S that best capture the subject's response. These values can be compared to normative values to separately assess the integrity of the auditory and somatosensory feedback control subsystems. If, for example, an individual with Parkinson's disease has an abnormally low α_S with normal α_A, a clinician may favor approaches that leverage intact auditory feedback control capabilities to overcome deficient somatosensory feedback control capabilities. In contrast, the parameters in the PID models are interpreted relative to error correction (and whether that correction is proportional to the error, or an integral or derivative of the error). This interpretation does not convey information about the mechanisms driving the correction and may limit how that information could be used in a therapeutic context. Although much work remains to be done to verify the veracity of the D1 model's characterization, such an approach holds the promise of informing personalized therapeutic interventions, much like other reflexes such as the pupillary light reflex have proven useful for characterizing the integrity of the nervous system in cases of neurological impairment.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

EK and FG conceptualized and designed the study. FG and AD developed the computational models. EK, AN-C, and RF developed the software. EH and DS collected and processed the data. EK, AN-C, and FG wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the final submitted version.

Funding

This research was supported by NIH grants: R01 DC002852 (FG, PI), R01 DC016270 (FG and C. Stepp, PIs), F31 DC016197 (EH, PI), and R21 DC017563 (AD, PI).

Acknowledgments

We are grateful to Cara Stepp for kindly sharing data with us for this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnhum.2022.929687/full#supplementary-material

References

Abur, D., Subaciute, A., Daliri, A., Lester-Smith, R. A., Lupiani, A. A., Cilento, D., et al. (2021a). Feedback and feedforward auditory-motor processes for voice and articulation in Parkinson's disease. *J. Speech Lang. Hear. Res.* 64, 4682–4694. doi: 10.1044/2021_JSLHR-21-00153

Abur, D., Subaciute, A., Kapsner-Smith, M., Segina, R. K., Tracy, L. F., Noordzij, J. P., et al. (2021b). Impaired auditory discrimination and auditory-motor integration in hyperfunctional voice disorders. *Sci. Rep.* 11:13123. doi: 10.1038/s41598-021-92250-8

Ballard, K. J., Halaki, M., Sowman, P. F., Kha, A., Daliri, A., Robin, D., et al. (2018). An investigation of compensation and adaptation to auditory perturbations in individuals with acquired apraxia of speech. *Front. Hum. Neurosci.* 12:510. doi: 10.3389/fnhum.2018.00510

Bär, K.-J., Boettger, M. K., Schulz, S., Harzendorf, C., Agelink, M. W., Yeragani, V. K., et al. (2008). The interaction between pupil function and cardiovascular regulation in patients with acute schizophrenia. *Clin. Neurophysiol.* 119, 2209–2213. doi: 10.1016/j.clinph.2008.06.012

Behroozmand, R., Bonilha, L., Rorden, C., Hickok, G., and Fridriksson, J. (2022). Neural correlates of impaired vocal feedback control in post-stroke aphasia. *Neuroimage* 250:118938. doi: 10.1016/j.neuroimage.2022.118938

Behroozmand, R., Ibrahim, N., Korzyukov, O., Robin, D. A., and Larson, C. R. (2014). Left-hemisphere activation is associated with enhanced vocal pitch error detection in musicians with absolute pitch. *Brain Cogn.* 84, 97–108. doi: 10.1016/j.bandc.2013.11.007

Behroozmand, R., Phillip, L., Johari, K., Bonilha, L., Rorden, C., Hickok, G., et al. (2018). Sensorimotor impairment of speech auditory feedback processing in aphasia. *Neuroimage* 165, 102–111. doi: 10.1016/j.neuroimage.2017.10.014

Boersma, P., and Weenink, D. (2018). *Praat: Doing phonetics by computer [Computer software]*. Available online at: http://www.praat.org/

Burnett, T. A., Freedland, M. B., Larson, C. R., and Hain, T. C. (1998). Voice F0 responses to manipulations in pitch feedback. *J. Acoust. Soc. Am.* 103, 3153–3161. doi: 10.1121/1.423073

Burnett, T. A., Senner, J. E., and Larson, C. R. (1997). Voice F0 responses to pitch-shifted auditory feedback: A preliminary study. *J. Voice* 11, 202–211. doi: 10.1016/S0892-1997(97)80079-3

Burnham, K., and Anderson, D. (2002). *Model selection and multimodel Inference: A practical information-theoretic approach*, 2nd Edn. New York, NY: Springer.

Cai, S., Boucek, M., Ghosh, S. S., Guenther, F. H., and Perkell, J. S. (2008). "A system for online dynamic perturbation of formant frequencies and results from perturbation of the Mandarin triphthong /iau/," in *Proceedings of the 8th intl. seminar on speech production* (Strasbourg), 65–68.

Chen, S. H., Liu, H., Xu, Y., and Larson, C. R. (2007). Voice F0 responses to pitch-shifted voice feedback during English speech. *J. Acoust. Soc. Am.* 121, 1157–1163.

Daliri, A. (2021). A computational model for estimating the speech motor system's sensitivity to auditory prediction errors. *J. Speech Lang. Hear. Res.* 64, 1841–1854. doi: 10.1044/2021_JSLHR-20-00484

Daliri, A., and Dittman, J. (2019). Successful auditory motor adaptation requires task-relevant auditory errors. *J. Neurophysiol.* 122, 552–562. doi: 10.1152/jn.00662.2018

Daliri, A., Chao, S.-C., and Fitzgerald, L. C. (2020). Compensatory responses to formant perturbations proportionally decrease as perturbations increase. *J. Speech Lang. Hear. Res.* 63, 3392–3407. doi: 10.1044/2020_JSLHR-19-00422

Demopoulos, C., Kothare, H., Mizuiri, D., Henderson-Sabes, J., Fregeau, B., Tjernagel, J., et al. (2018). Abnormal speech motor control in individuals with 16p11.2 deletions. *Sci. Rep.* 8:1274. doi: 10.1038/s41598-018-19751-x

Elman, J. L. (1981). Effects of frequency-shifted feedback on the pitch of vocal productions. *J. Acoust. Soc. Am.* 70, 45–50. doi: 10.1121/1.386580

Fan, X., Miles, J. H., Takahashi, N., and Yao, G. (2009). Abnormal transient pupillary light reflex in individuals with autism spectrum disorders. *J. Autism Dev. Disord.* 39, 1499–1508. doi: 10.1007/s10803-009-0767-7

Franken, M. K., Acheson, D. J., McQueen, J. M., Hagoort, P., and Eisner, F. (2018). Opposing and following responses in sensorimotor speech control: Why responses go both ways. *Psychon. Bull. Rev.* 25, 1458–1467. doi: 10.3758/s13423-018-1494-x

Galea, J. M., Mallia, E., Rothwell, J., and Diedrichsen, J. (2015). The dissociable effects of punishment and reward on motor learning. *Nat. Neurosci.* 18, 597–602. doi: 10.1038/nn.3956

Guenther, F. H. (2016). *Neural control of speech*. Cambridge, MA: MIT Press.

Guenther, F. H., Hampson, M., and Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychol. Rev.* 105, 611–633. doi: 10.1037/0033-295X.105.4.611-633

Hain, T., Burnett, T., Kiran, S., Larson, C., Singh, S., and Kenney, M. (2000). Instructing subjects to make a voluntary response reveals the presence of two components to the audio-vocal reflex. *Exp. Brain Res.* 130, 133–141. doi: 10.1007/s002219900237

Hall, C. A., and Chilcott, R. P. (2018). Eyeing up the future of the pupillary light reflex in neurodiagnostics. *Diagnostics* 8:19. doi: 10.3390/diagnostics8010019

Hantzsch, L., Parrell, B., and Niziolek, C. A. (2022). A single exposure to altered auditory feedback causes observable sensorimotor adaptation in speech. *Elife* 11:e73694.

Heller Murray, E. S., and Stepp, C. E. (2020). Relationships between vocal pitch perception and production: A developmental perspective. *Sci. Rep.* 10, 1–10. doi: 10.1038/s41598-020-60756-2

Heller Murray, E. S., Lupiani, A. A., Kolin, K. R., Segina, R. K., and Stepp, C. E. (2019). Pitch shifting with the commercially available eventide eclipse: Intended and unintended changes to the speech signal. *J. Speech Lang. Hear. Res.* 62, 2270–2279. doi: 10.1044/2019_JSLHR-S-18-0408

Houde, J. F., and Nagarajan, S. S. (2011). Speech production as state feedback control. *Front. Hum. Neurosci.* 5:82. doi: 10.3389/fnhum.2011.00082

Houde, J. F., Gill, J., Agnew, Z., Kothare, H., Hickok, G., Parrell, B., et al. (2019). Abnormally increased vocal responses to pitch feedback perturbations in patients with cerebellar degeneration. *J. Acoust. Soc. Am.* 145, EL372–EL378. doi: 10.1121/1.5100910

Huberdeau, D. M., Krakauer, J. W., and Haith, A. M. (2015). Dual-process decomposition in human sensorimotor adaptation. *Curr. Opin. Neurobiol.* 33, 71–77. doi: 10.1016/j.conb.2015.03.003

Kearney, E., and Guenther, F. H. (2019). Articulating: The neural mechanisms of speech production. *Lang. Cogn. Neurosci.* 34, 1214–1229. doi: 10.1080/23273798.2019.1589541

Kearney, E., Nieto-Castañón, A., Weeratunge, H., Falsini, R., Daliri, A., Abur, D., et al. (2020). A simple 3-parameter model for examining adaptation in speech and voice production. *Front. Psychol.* 10:2995. doi: 10.3389/fpsyg.2019.02995

Kiran, S., and Larson, C. R. (2001). Effect of duration of pitch-shifted feedback on vocal responses in patients with Parkinson's disease. *J. Speech Lang. Hear. Res.* 44, 975–987. doi: 10.1044/1092-4388(2001/076)

Koo, T. K., and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* 15, 155–163. doi: 10.1016/j.jcm.2016.02.012

Larson, C. R. (1998). Cross-modality influences in speech motor control: The use of pitch shifting for the study of F0 control. *Clin. Commun. Disord.* 31, 489–503. doi: 10.1016/S0021-9924(98)00021-5

Larson, C. R., Burnett, T. A., Kiran, S., and Hain, T. C. (2000). Effects of pitch-shift velocity on voice F0 responses. *J. Acoust. Soc. Am.* 107, 559–564. doi: 10.1121/1.428323

Li, W., Zhuang, J., Guo, Z., Jones, J. A., Xu, Z., and Liu, H. (2019). Cerebellar contribution to auditory feedback control of speech production: Evidence from patients with spinocerebellar ataxia. *Hum. Brain Mapp.* 40, 4748–4758. doi: 10.1002/hbm.24734

Liu, H., Wang, E. Q., Metman, L. V., and Larson, C. R. (2012). Vocal responses to perturbations in voice auditory feedback in individuals with Parkinson's disease. *PLoS One* 7:e33629. doi: 10.1371/journal.pone.0033629

Loucks, T., Chon, H., and Han, W. (2012). Audiovocal integration in adults who stutter. *Int. J. Lang. Commun. Disord.* 47, 451–456. doi: 10.1111/j.1460-6984.2011.00111.x

Master, C. L., Podolak, O. E., Ciuffreda, K. J., Metzger, K. B., Joshi, N. R., McDonald, C. C., et al. (2020). Utility of pupillary light reflex metrics as a physiologic biomarker for adolescent sport-related concussion. *JAMA Ophthalmol.* 138, 1135–1141. doi: 10.1001/jamaophthalmol.2020.3466

Minorsky, N. (1922). Directional stability of automatically steered bodies. *J. Am. Soc. Nav. Eng.* 34, 280–309. doi: 10.1111/j.1559-3584.1922.tb04958.x

Natke, U., and Kalveram, T. (2001). Effects of frequency-shifted auditory feedback on fundamental frequency of long stressed and unstressed syllables. *J. Speech Lang. Hear. Res.* 44, 577–584. doi: 10.1044/1092-4388(2001/045)

Pamplona, V. F. (2008). *Photorealistic models for oupil light reflex and iridal pattern deformation*. Available online at: https:

//lume.ufrgs.br/bitstream/handle/10183/15309/000677246.pdf;jsessionid=0CBDC969CA82E2EC4E26AD99A72AD927?sequence=1 (accessed April 12, 2022).

Parkinson, A., Behroozmand, R., Ibrahim, N., Korzyukov, O., Larson, C., and Robin, D. (2014). Effective connectivity associated with auditory error detection in musicians with absolute pitch. *Front. Neurosci.* 8:46. doi: 10.3389/fnins.2014

Ranasinghe, K. G., Gill, J. S., Kothare, H., Beagle, A. J., Mizuiri, D., Honma, S. M., et al. (2017). Abnormal vocal behavior predicts executive and memory deficits in Alzheimer's disease. *Neurobiol. Aging* 52, 71–80. doi: 10.1016/j.neurobiolaging.2016.12.020

Rubin, L. S. (1980). Pupillometric studies of alcoholism. *Int. J. Neurosci.* 11, 301–308. doi: 10.3109/00207458009147594

Russo, N., Larson, C. R., and Kraus, N. (2008). Audio-vocal system regulation in children with autism spectrum disorders. *Exp. Brain Res.* 188, 111–124. doi: 10.1007/s00221-008-1348-2

Sares, A. G., Deroche, M. L. D., Ohashi, H., Shiller, D. M., and Gracco, V. L. (2020). Neural correlates of vocal pitch compensation in individuals who stutter. *Front. Hum. Neurosci.* 14:18. doi: 10.3389/fnhum.2020.00018

Sares, A. G., Deroche, M. L. D., Shiller, D. M., and Gracco, V. L. (2018). Timing variability of sensorimotor integration during vocalization in individuals who stutter. *Sci. Rep.* 8:16340. doi: 10.1038/s41598-018-34517-1

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin* 2, 110–114.

Scheerer, N. E., Jacobson, D. S., and Jones, J. A. (2016). Sensorimotor learning in children and adults: Exposure to frequency-altered auditory feedback during speech production. *Neuroscience* 314, 106–115. doi: 10.1016/j.neuroscience.2015.11.037

Scheerer, N. E., Liu, H., and Jones, J. A. (2013). The developmental trajectory of vocal and event-related potential responses to frequency-altered auditory feedback. *Eur. J. Neurosci.* 38, 3189–3200. doi: 10.1111/ejn.12301

Smith, D. J., Stepp, C. E., Guenther, F. H., and Kearney, E. (2020). Contributions of auditory and somatosensory feedback to vocal motor control. *J. Speech Lang. Hear. Res.* 63, 2039–2053. doi: 10.1044/2020_JSLHR-19-00296

Smith, M. A., Ghazizadeh, A., and Shadmehr, R. (2006). Interacting adaptive processes with different timescales underlie short-term motor learning. *PLoS Biol.* 4:e179. doi: 10.1371/journal.pbio.0040179

Stergiou, V., Fotiou, D., Tsiptsios, D., Haidich, B., Nakou, M., Giantselidis, C., et al. (2009). Pupillometric findings in patients with Parkinson's disease and cognitive disorder. *Int. J. Psychophysiol.* 72, 97–101. doi: 10.1016/j.ijpsycho.2008.10.010

Tales, A., Troscianko, T., Lush, D., Haworth, J., Wilcock, G. K., and Butler, S. R. (2001). The pupillary light reflex in aging and Alzheimer's disease. *Aging (Milan)* 13, 473–478.

Thompson, H. S. (2003). The vitality of the pupil: A history of the clinical use of the pupil as an indicator of visual potential. *J. Neuroophthalmol.* 23, 213–224. doi: 10.1097/00041327-200309000-00007

Thoroughman, K. A., and Shadmehr, R. (2000). Learning of action through adaptive combination of motor primitives. *Nature* 407, 742–747. doi: 10.1038/35037588

Xu, Y., Larson, C. R., Bauer, J. J., and Hain, T. C. (2004). Compensation for pitch-shifted auditory feedback during the production of mandarin tone sequences. *J. Acoust. Soc. Am.* 116, 1168–1178. doi: 10.1121/1.1763952

Zarate, J. M., Wood, S., and Zatorre, R. J. (2010). Neural networks involved in voluntary and involuntary vocal pitch regulation in experienced singers. *Neuropsychologia* 48, 607–618. doi: 10.1016/j.neuropsychologia.2009.10.025

frontiers | Frontiers in Human Neuroscience

TYPE Hypothesis and Theory
PUBLISHED 15 February 2023
DOI 10.3389/fnhum.2023.893785

Check for updates

OPEN ACCESS

EDITED BY
John Houde,
University of California, San Francisco, United States

REVIEWED BY
Pascal Perrier,
UMR5216 Grenoble Images Parole Signal Automatique (GIPSA-lab), France
Connor Mayer,
University of California, Irvine, United States

*CORRESPONDENCE
Melissa A. Redford
✉ redford@uoregon.edu

SPECIALTY SECTION
This article was submitted to Speech and Language, a section of the journal Frontiers in Human Neuroscience

RECEIVED 10 March 2022
ACCEPTED 26 January 2023
PUBLISHED 15 February 2023

CITATION
Davis M and Redford MA (2023) Learning and change in a dual lexicon model of speech production. *Front. Hum. Neurosci.* 17:893785. doi: 10.3389/fnhum.2023.893785

COPYRIGHT
© 2023 Davis and Redford. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Learning and change in a dual lexicon model of speech production

Maya Davis and Melissa A. Redford*

Department of Linguistics, University of Oregon, Eugene, OR, United States

Speech motor processes and phonological forms influence one another because speech and language are acquired and used together. This hypothesis underpins the Computational Core (CC) model, which provides a framework for understanding the limitations of perceptually-driven changes to production. The model assumes a lexicon of motor and perceptual wordforms linked to concepts and whole-word production based on these forms. Motor wordforms are built up with speech practice. Perceptual wordforms encode ambient language patterns in detail. Speech production is the integration of the two forms. Integration results in an output trajectory through perceptual-motor space that guides articulation. Assuming successful communication of the intended concept, the output trajectory is incorporated into the existing motor wordform for that concept. Novel word production exploits existing motor wordforms to define a perceptually-acceptable path through motor space that is further modified by the perceptual wordform during integration. Simulation results show that, by preserving a distinction between motor and perceptual wordforms in the lexicon, the CC model can account for practice-based changes in the production of known words and for the effect of expressive vocabulary size on production accuracy of novel words.

KEYWORDS

computational model, development, exemplar theory, schema theory, speech motor plan

Introduction

How do we produce an unfamiliar word that we have just heard? One answer is that we hear and encode the word as a sequence of phonemes; when the sequence is activated for production, the phonetic aspect is filled in, syllable structure is imposed, and the corresponding motor programs are selected and executed (Levelt, 1989; Levelt et al., 1999; Guenther, 2016). But, if our production of the unfamiliar word is inaccurate, how exactly do we improve on it over time? The Computational Core (CC) model presented in this paper was built to address this question and others that arise from the developmental problem of learning and change in production— learning and change that occurs across the lifespan.

One approach to the problem of learning and change in production is to assume both perceptual representations linked to phonemes and online control over execution (e.g., Houde and Nagarajan, 2011; Parrell et al., 2019). Under these assumptions, predictive control can be used to adjust a planned articulation that will miss the acoustic goal linked to a phoneme (Niziolek et al., 2013). But what if the unfamiliar word that a speaker attempts makes use of familiar phonemes linked to unfamiliar sounds arranged according to an unfamiliar timing pattern? The standard approach to this problem, encountered in adult second language learning, is to assume perceptual learning at the level of the acoustic categories that define speech motor goals (Flege, 1995; Samuel and Kraljic, 2009; Holt and Lotto, 2010; Flege and Bohn, 2021). Such learning could induce change in production based on online control. Yet, studies on second language acquisition indicate that accurate perceptual learning does not result in production accuracy (Nagle and Baese-Berk, 2022), especially if the newly learned acoustic category cannot be mapped onto a speaker's prior production experience (Nielsen, 2011; Nagle, 2018). Despite learning, changes in production accuracy are constrained.

Also, even if an unfamiliar sound can be attained based on perceptual learning, how is an unfamiliar timing pattern achieved? Native-like production of relative timing patterns within a word are acquired early by first language speakers, but not nearly as easily—if ever—by adult second language speakers (e.g., Redford and Oh, 2017). The question of how relative timing patterns are acquired is especially difficult to address within a framework where word production and perception are mediated by phonemes. An alternative approach is to assume that learning is instead mediated by wordform representations. For example, the detailed acoustic-perceptual wordform representations of exemplar-based theories (Johnson, 1997, 2006; Pierrehumbert, 2002; Smith and Hawkins, 2012) necessarily include time-varying information about acoustic goals that could be referenced during execution. Predictive control could be used to adjust planned articulations accordingly, which would result in changes to production. But, if accurate production of unfamiliar words with unfamiliar sounds and timing patterns can be attained simply with reference to whole-word perceptual representations, then why is the correlation between perception and production in second language acquisition so far from perfect? Put another way: What constrains production during learning? Relatedly, why does production accuracy, measured against perceptual input, appear to plateau in adult second language speakers?

The typical explanation for constrained production accuracy in second language speech is that unfamiliar words are not directly read off from perceptual representations; rather, they are filtered through a speaker's phonology (Major, 1998, 2001). In exemplar-based theories, the phonology is language-specific knowledge about phonemes, phonotactics, and other suprasegmental patterns abstracted from across the perceptual wordforms of the lexicon (Bybee, 2002; Pierrehumbert, 2003). When these abstractions are stored ("labeled") separately from the lexicon, an exemplar-based model of production makes assumptions similar to phoneme-driven models of production (see, e.g., Pierrehumbert, 2001; Wedel, 2006); that is, it assumes acoustic goals linked to phonemes and so it assumes phoneme-guided production. Given that time-varying information must also be learned and implemented by the motor system to effect change in production, this type of model is unsatisfactory. The CC model presents a word-based alternative to the phoneme-driven model of production. The goal of the model is to account for perceptually-driven learning and change in production and for the constraints on said change.

The CC model addresses learning and change from a developmental perspective. This perspective is adopted because (a) the problem of learning and change is especially acute in early language development, and (b) the adult's production system emerges from the child's and so should be derived from it. The latter reason constitutes a working hypothesis that has led us to propose a developmentally sensitive theory of speech production (Redford, 2015, 2019)—a framework for understanding the evolution of speech production across the lifespan. The CC model details an important piece of the theory: the idea that speech motor processes and phonological forms influence one another because speech and language are acquired together. The model instantiation of this idea captures language-specific limits on perceptually-driven motor learning and change in production.

Background to the CC model

The CC model assumes a dual lexicon. More specifically, it assumes a lexicon comprised of separate perceptual and motor wordforms that are jointly linked to shared concepts. The CC model also assumes whole-word production. These assumptions are motivated by our developmental perspective. Both extend specific ideas from child phonology to provide the basis for a developmentally sensitive account of adult production.

The shapes of children's first words deviate markedly from adult wordforms. Work in child phonology shows that these deviations are idiosyncratic. For example, one child will say [bɑbɑ] for *bottle* (Velleman, 1998; cited in Velleman and Vihman, 2002, p. 20) while another says [bɑdi] (Vihman, 2014, p. 80) and a third says [papm:] (Jaeger, 1997; Vihman and Croft, 2007, p. 702). The idiosyncratic productions of single words are associated with child-specific systematicities across multiple words. For example, the 18-month-old who says [pɑpm:] for "bottle" replaces voiced stops with voiceless ones in "baby" and "byebye," rendering these as [peipi] and [(pə)pa:i], respectively; she also produces word-final nasals in other words where they are not required (e.g., [kʌkŋ] for "cracker" and [takŋ] for "doggie"; see Table 9 in Vihman and Croft, 2007, p. 702). In general, children's deviations from adult-like wordforms are interpreted to suggest strong motor constraints on first word production (Menn, 1983; Nittrouer et al., 1989; McCune and Vihman, 2001; Davis et al., 2002). Ferguson and Farwell (1975) proposed that individual children overcome these constraints by applying their favored sound patterns to best approximate whole word targets, resulting in systematic patterns of individual difference in production. McCune and Vihman (2001) went further to specify that a child's favored patterns are selected from among their vocal motor schemes that are established with vocal-motor practice during the pre-speech period. Redford (2015) combined this idea with the ideas of generalized motor programs from schema theory (see Schmidt, 1975, 2003) and gestural scores from Articulatory Phonology (Browman and Goldstein, 1986, 1992) to propose that, even beyond the first word period, the child continues to rely on established motor representations to guide production and that this reliance continues on through adulthood.

In Redford (2015), the motor representations that guide production were defined as temporally-structured memories built up from motor traces associated with the successful communication of concepts. They are first established when communication of a new concept is first attempted. Of course, this first attempt requires that the child also have stored a perceptual representation of the wordform that denotes a concept. This representation serves as the goal for production. Its presence in the lexicon allows for developmental change in the direction of the adult form (Redford, 2019). But, with a hypothesis of whole-word production, comes the problem of how to explain the emergence of segment-like control over speech articulation. Davis and Redford (2019) proposed the Core model to address this problem. In brief, Core demonstrated that segment-like control could emerge under the assumption of whole-word production with practice-based structuring of the perceptual-motor map. This specific solution to the problem entailed formalizing a number of concepts that are also central to the CC model. Figure 1 itemizes and illustrates these concepts for quick reference. More complete descriptions of the concepts follow.

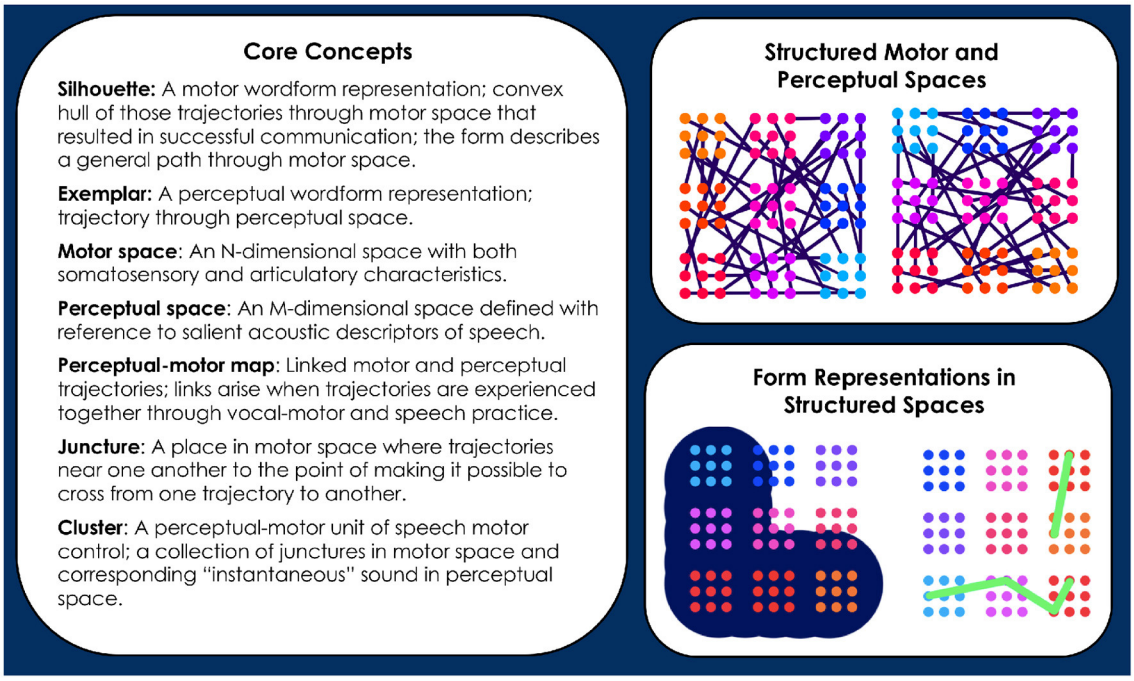


FIGURE 1
Informal definitions of Core concepts are provided (see text for detail). The illustrations to the right of the definitions depict several of the concepts. The top right panel depicts 2-dimensional motor **(left)** and perceptual **(right)** spaces that have already been structured by the trajectory crossings that occur with vocal-motor exploration and speech practice. Junctures are represented as dots, clusters as groups of identically colored dots. Each cluster of a particular color in motor space corresponds to one of the same color in perceptual space. Links between the motor and perceptual spaces are assumed but not shown. The bottom right panel depicts a silhouette **(left)** and an exemplar **(right)** in relation to the motor and perceptual spaces, respectively. The depiction of the silhouette highlights the idea that it describes a broad path through motor space. The depiction of an exemplar highlights its status as a specific trajectory through perceptual space. The distinct layouts of clusters in the simplified motor and perceptual spaces illustrates that these spaces have different topologies.

Core concepts

The CC model assumes that motor wordforms are established with reference to perceptual wordforms and that, once established, the motor and perceptual forms are integrated during production (Redford, 2019). We first formalized this hypothesis in the Core model (Davis and Redford, 2019). In so doing, we defined a lexicon of perceptual and motor wordforms with respect to a *perceptual space* and a *motor space*.

The perceptual space is the set of all possible instantaneous sounds, along with a distance metric and subsequent topology. The motor space is the set of all possible articulatory configurations, along with a distance metric and subsequent topology. The perceptual and motor spaces are grounded in the acoustic and articulatory dimensions of speech. This grounding is assumed but not defined in the CC model. In Davis and Redford (2019) the dimensions were as follows. A point in perceptual space was represented by coordinates measuring sound periodicity, Bark-transformed formant values, the spectral center of gravity, the width of the spectral peak, and the time derivatives of the formant and other spectral measures, as well as the time derivative of amplitude. A point in motor space was represented by coordinates measuring glottal width, the cross-sectional areas of 8 regions of the vocal tract from lips to larynx, the time derivatives of each of the cross-sectional areas, velum height, the time derivative of velum height, and the direction and force of the opening/closing movement of the jaw. Euclidean distance metrics were used to calculate the relationship between points in these spaces.

The perceptual wordform, defined with respect to perceptual space, is called an *exemplar*. The label indicates our embrace of exemplar-based accounts of phonology, sociolinguistic knowledge, and perceptual learning. None of these topics are explicitly addressed here. Instead, the exemplar is merely a precise whole-word perceptual representation. It is a function that takes a moment in time as an input and gives as an output a point in perceptual space. Such a function describes a trajectory through perceptual space; it is called an exemplar only when linked to a concept.

The motor wordform, defined with respect to motor space, is called a *silhouette*. It is a temporally-structured memory of the movements needed to achieve a wordform that communicates a concept. It is built up over time whenever its concept is successfully communicated. It is most analogous to the idea of a generalized motor program (GMP) for skilled action (Schmidt, 1975, 2003), except that it is a more specific representation than the GMP. Unlike a GMP, a silhouette is effector-dependent: it is defined along dimensions determined by possible movements of the speech articulators.

In first-word production, exemplars are purely exogenous representations. Silhouettes are endogenous representations that begin to emerge when the infant first successfully communicates a concept C by targeting the exemplar, e_C. The silhouette for the concept, SIL_C, is a function that takes a point in time as an input, and

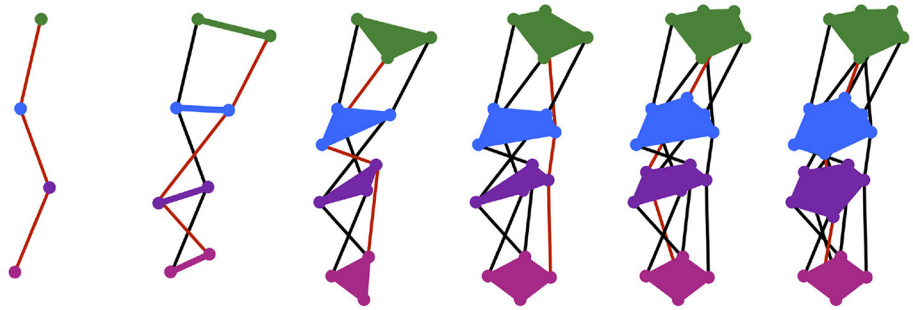


FIGURE 2
Each panel shows the silhouette expanded from the previous panel to include an additional motor trajectory, whose path is shown in red. The regions of the silhouette at four time steps are drawn—the region at the first time is shown in green, at the second time in blue, at the third time in purple, and at the fourth time in pink—but theoretically infinitely many regions exist along the whole length of the silhouette. Silhouette expansion is a continual process. Motor trajectory traces are added whenever communication succeeds.

gives as an output a region in motor space that describes a general vocal tract configuration to be targeted by the motor system at that time. As with the exemplar, the subscript C denotes the silhouette's link to the concept C. Each time C is successfully communicated, SIL_C expands to include a trace of the motor trajectory, m, that was executed. More specifically, for each time t, the region $\text{SIL}_C(t)$ expands the smallest amount possible such that (1) the new region also includes $m(t)$ (as well as the old region) and (2) the new region is convex. In the CC model, new and old regions are also weighted over time with the addition of new traces representing successful communication of C, which effectively skews the silhouette in the direction of the most frequently used motor trajectories. An illustration of motor silhouette expansion is shown in Figure 2. Silhouette weighting is not shown; it is instead described at length later in this paper.

First word production is the effective communication of a novel concept C that has been learned along with e_C from the ambient language. The infant first achieves communication of C through a matching and selection process that leverages motor trajectories established through babbling and other vocal-motor exploration. Because the trajectories in motor space are self-produced, they are automatically linked to perceptual trajectories in perceptual space. The linked motor and perceptual trajectories make up the *perceptual-motor map* that is exploited during the matching and selection process used to attempt a new word. This process computes the distance in perceptual space between an exemplar and the perceptual aspect of established motor trajectories through motor space. The computation allows for the combination of multiple established trajectories, one after another in time, to best approximate the intended exemplar. Along the way, the matching and selection process structures the perceptual-motor map by creating *junctures*, which are motor points at which the speaker shifts from one established trajectory to another nearby one.

Even during the initial stages of vocal-motor exploration, very specific regions of motor space are passed over multiple times in a variety of trajectories (e.g., the [ɑ] region in babbled utterances "bɑbɑ" and "dɑdɑ"). In Davis and Redford (2019), we proposed that frequently traversed regions in motor space become populated with junctures through the matching and selection process during the first word stage of development. The specific suggestion was that children create junctures when they combine chunks of previously experienced perceptually-linked motor trajectories in their first word attempts. For example, a child will first link the perceptual and motor spaces of speech during the pre-linguistic period, including with trajectories such as "bɑbɑ" and "dɑdɑ" produced during the babbling phase. When this child first attempts the word "bottle" they may seek to match its perceptual form by leveraging the "bɑbɑ" or "dɑdɑ" trajectory. They may even combine these trajectories to produce "bɑdɑ" by following the (motor) path for "bɑbɑ" and then transitioning to the path for "dɑdɑ" where the two trajectories (nearly) meet in the [ɑ] region of motor space. If the resulting "bɑdɑ" trajectory contributes to communicative success (e.g., receiving the requested bottle), then the motor trace of the "bɑdɑ" trajectory is stored with a link to the concept "bottle." This trace provides the first outline for the silhouette associated with that concept (see Figure 2).

As junctures proliferate with vocal-motor practice and vocabulary expansion, they are grouped together based on their proximity to one another in motor space. These groupings are *clusters*. A cluster designates a specific region in motor space that is crossed over and over again while achieving similar sounds within various words. Over developmental time, clusters begin to serve as perceptual-motor units of control. They can be targeted quasi-independently because they designate regions within motor space that many trajectories go through, allowing the speaker to target the region from many other locations within the space. At a higher level of abstraction, clusters represent turning points in motor trajectories. These turning points can be conceived of as linguistically-significant vocal tract constrictions—something similar to "gestures" in Articulatory Phonology (Browman and Goldstein, 1986, 1992), albeit with context-dependent timing that is defined by the trajectory leading into and out of the turning point. In perceptual space, clusters represent a quasi-static acoustic goal associated with a particular articulatory configuration—such as the sound that we might associate with a segment (e.g., [ɑ]) or with a critical feature (e.g., the silence of stop closure). Although it is possible to associate clusters with gestural or featural descriptions of the phonology, we stress that they are simply units of speech motor control. Clusters only exist at the level of the perceptual-motor map. They do not necessarily create meaning contrasts. They emerge from and remain embedded in a well-defined perceptual-motor context.

Having introduced the Core concepts of perceptual and motor spaces, exemplars, silhouettes, the perceptual-motor map, junctures, and clusters, we are ready to describe the CC model. This model picks

up after the first-word stage where the mathematical Core model leaves off.

Architecture of the CC model

In Davis and Redford (2019), we modeled the first-word stage of spoken language development and its structuring effects on the perceptual-motor map. In this paper, we model word production at a later stage in development; a stage when the perceptual-motor map has already been structured with speech practice and so is already discretized into clusters. This new focus entails making explicit the relationship between wordform representations and the perceptual-motor map. This relationship is critical to the perceptual-motor integration of wordforms that is at the heart of speech production in the theory.

The silhouette and exemplar activate clusters in motor and perceptual space, respectively. In the CC model, sequential information is preserved by the silhouette with the time-varying activation of clusters in motor space.[1] By contrast, the exemplar activates all its clusters at the same time in perceptual space. The time-varying activation of clusters in motor space is consistent with the ecological–dynamic hypothesis that phonological representations incorporate time-varying (i.e., dynamic) information (Fowler, 1980; Browman and Goldstein, 1986, 1992). The simultaneous activation of clusters in perceptual space is consistent with the structural hypothesis that paradigmatic relations are more important than syntagmatic ones when acoustic-auditory categories serve as speech motor goals (Diehl and Lindblom, 2004; Flemming, 2004). Very importantly, the different activation patterns ensure unique motor and perceptual contributions to wordform integration. The silhouette-driven activation pattern highlights context-dependent constraints on articulation. The exemplar-driven activation pattern highlights the goal of attaining (more) context-independent sounds in articulation. The different activation patterns and their specific consequences are inspired by Lindblom's (1990) H&H theory of production. Lindblom proposes that speakers have two modes of production, a hypo mode and a hyper mode, that serve as ends of a speaking style continuum. The hypo mode results in highly coarticulated speech. The hyper mode results in more context-independent attainment of acoustic goals. The CC model reflects these extreme modes in its different activation patterns of motor and perceptual space.[2]

The silhouette and exemplar are integrated with cluster activation. More specifically, the activation pattern across clusters in motor space and the activation pattern across clusters in perceptual space are combined and used to determine a trajectory through the perceptual-motor map that guides speech movement. Look-ahead and look-back windows specify the extent to which information about the combined activation pattern in the future and/or past is incorporated into the current activation pattern. At any given time, the integration process thus results in the differential activation of multiple clusters. As clusters represent perceptual-motor units that are both spatial targets and perceptual goals, the simultaneous activation of several of these at once means that articulation represents a compromise between competing targets/goals.

Overall, the CC model claim is one of real-time speech motor planning and execution. Speech motor control is not modeled but the planning process remains compatible with current models (e.g., Houde and Nagarajan, 2011; Guenther, 2016; Parrell et al., 2019). In what follows, the production process from cluster activation to perceptual-motor integration to the computation of the (perceptual-)motor output trajectory is formally described. We would point those interested in further detail to the source code, which is available on GitHub (https://github.com/mayaekd/core).

Cluster activation

Let C be a word-sized concept. The speech plan for C is the activation pattern of clusters in the perceptual-motor map that results from the selection of the silhouette that corresponds to C, SIL_C, and an exemplar, e_C, chosen from among the set of exemplars associated with C. The perceptual-motor map itself contains many clusters: $\text{CLUSTER}_1, \text{CLUSTER}_2, \ldots, \text{CLUSTER}_n$. Each of these is made up of some number of junctures; assume CLUSTER_i is made up of $\text{JUNCTURE}_{i,1}, \text{JUNCTURE}_{i,2}, \ldots, \text{JUNCTURE}_{i,m_i}$. The silhouette, SIL_C, activates clusters in motor space while the exemplar, e_C, activates clusters in perceptual space. For the reasons explained in the preceding section, the activation of clusters in motor space varies across time; the activation of clusters in perceptual space is simultaneous. The details of the activation patterns are as follows.

Activation in motor space

First, the silhouette activates the region in motor space corresponding to the first step on the time interval. At the next time step, it activates the next corresponding region. At the one after that, the next region is activated, and so on until the path through motor space associated with the entire silhouette has been traversed.

When a region in motor space is activated, the activation immediately spreads across junctures that are inside that region or within a certain distance of that region. Juncture activation spreads evenly within the bounds of each cluster. This means that clusters are activated as units within motor space. Clusters that are further away from the region that is highlighted by a silhouette at a particular time step will be less activated than those that are closer to the region or are in the region itself, as depicted in Figure 3. More precisely, the motor activation at time t of CLUSTER_i is defined to be the average of the motor activation of every juncture in that cluster:

$$\text{MOTORACTIVATION}_t(\text{CLUSTER}_i) = \frac{1}{m_i}\sum_{j=1}^{m_i}\text{MOTORACTIVATION}_t(\text{JUNCTURE}_{i,j})$$

1 Time is modeled discretely for computational reasons, but the concept is one of a continuous unfolding process (see Davis and Redford, 2019).

2 Style-shifting is not addressed in this paper, but can be modeled within CC as the greater weighting of either the motor or perceptual activation pattern during integration. A reviewer points out that style could also be modeled in other ways within the model, including by the selection of specific formal or casual exemplars of words or by changing the size of the look-back and look-ahead windows of integration. This is also true. The main point here is that the distinct motor and perceptual activation patterns in the CC model are meant to incorporate the tension between "ease" and "distinctiveness" that is at the heart of Lindblom's H&H theory of production.

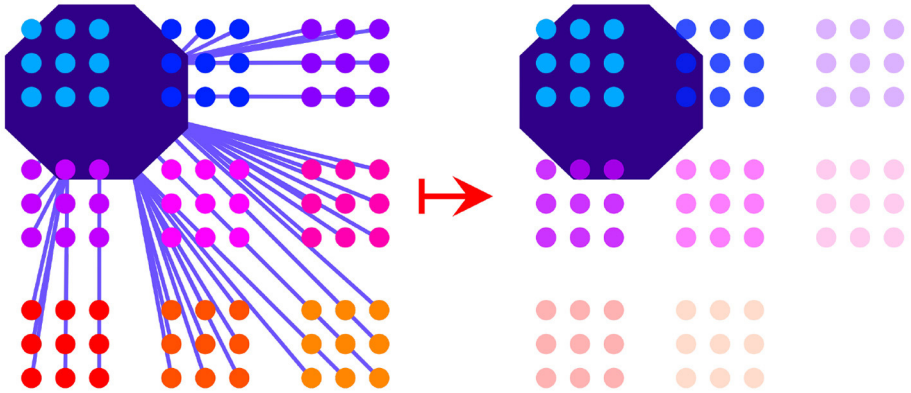


FIGURE 3
The activation process in motor space is shown. The region of the silhouette at a particular time step activates junctures that are overlapping with the region or less than a certain distance away from it (distances shown by purple lines, **left**). Activation spreads evenly within a cluster. Activation levels are determined by the distances of the junctures to the silhouette region. Activation strength of clusters is depicted by the relative transparency-opacity of the clusters **(right)**.

Where the motor activation of $\textsc{Juncture}_{i,j}$ is defined to be the highest when $\textsc{Juncture}_{i,j}$ is contained in $\textsc{sil}_C(t)$ and to fall off linearly as the distance between $\textsc{Juncture}_{i,j}$ and $\textsc{sil}_C(t)$ increases, bottoming out at zero:

$$\textsc{MotorActivation}_t(\textsc{Juncture}_{i,j}) = \textsc{HighestActivationMotor} - \left(\textsc{DropoffSlopeMotor} \times \textsc{Distance}(\textsc{sil}_C(t), \textsc{Juncture}_{i,j})\right)$$

We generally set

$$\textsc{HighestActivationMotor} = 1$$

and

$$\textsc{DropOffSlopeMotor} = 0.1.$$

Although we refer here to the motor activations of the junctures, note that this should be thought of as an initial theoretical state of the cluster that is quickly changed once the activation spreads within a cluster.

Activation in perceptual space

Although the exemplar is also a function on a time interval, its set of points activate nearby junctures in perceptual space all at once when the exemplar is selected. Similar to juncture activation in motor space, activation spreads outwards from points along the exemplar trajectory; activation also decreases in strength with distance from the exemplar trajectory, and the activation is averaged across the points in the exemplar. Again, activation spreads so that all junctures within a particular cluster receive the same activation. For an exemplar consisting of points $p_1, \ldots, p_r$, and a cluster $\textsc{Cluster}_i$ consisting of junctures $\{\textsc{Juncture}_{i,1}, \ldots, \textsc{Juncture}_{i,m_i}\}$, we can write

$$[\textsc{ExemplarActivation}(\textsc{Cluster}_i) = \frac{1}{m_i}\sum_{j=1}^{m_i} \textsc{ExemplarActivation}(\textsc{Juncture}_{i,j})$$

where

$$\textsc{ExemplarActivation}(\textsc{Juncture}_{i,j}) = \frac{1}{r}\sum_{k=1}^{r} (\textsc{HighestActivationPerceptual} - (\textsc{DropoffSlopePerceptual} \times \textsc{Distance}(p_k, \textsc{Juncture}_{i,j}))).$$

Like in the motor case, we generally set

$$\textsc{HighestActivationPerceptual} = 1$$

and

$$\textsc{DropOffSlopePerceptual} = 0.1$$

Perceptual-motor integration

The silhouette and exemplar are integrated as follows to produce speech output. First, the combined activation pattern across the motor and perceptual spaces is computed. This pattern consists of activation that varies by time and by cluster, and is determined by the following equation for the activation at time t of cluster $\textsc{Cluster}_i$:

$$\textsc{Activation}_t(\textsc{Cluster}_i) = \left(\textsc{MotorActivation}_t(\textsc{Cluster}_i) \times \textsc{ExemplarActivation}(\textsc{Cluster}_i)\right)^{\frac{1}{2}}$$

We take the geometric mean (multiplicative mean) of the two activations rather than the arithmetic mean (additive mean) in order to determine the combined activation of a cluster in a way that ensures the correct sequencing of articulatory movements. The geometric mean functions as an AND gate rather than as an OR gate to activation—if the activation of a cluster in either motor or perceptual space is zero, then the combined activation of that cluster is zero. Multiple clusters may compete to influence articulation, but

competing clusters should all be within some limited distance of the region specified by the silhouette at that moment in time. If they are not, they should not influence articulation at all. Although the same constraint applies to both spaces, the constraint from motor space is more important. By ensuring that zero activation of a cluster in motor space cannot be overridden by some activation of the cluster in perceptual space, we are ensuring that activation from parts of the exemplar trajectory not relevant to the current time do not have an overwhelming influence on the output trajectory at that time.

The activation values of the cluster vary over time. When activation is computed for a specific time t, this yields a set of values $a_i(t)$, for $i = 1, \ldots, n$, where $a_i(t)$ is the activation of CLUSTER_i. The CC model assumes that the motor system works out a compromise among the various clusters. In the model, the estimated outcome of this compromise at time t is computed as the weighted average of cluster locations in motor space, with the weights being the activations of the clusters at time t. That is, the estimated motor coordinate list, $\text{ESTMOTOR}(t)$, is defined as:

$$\text{ESTMOTOR}(t) = \frac{\sum_{i=1}^{n} a_i(t) \times \text{MOTORCENTER}(\text{CLUSTER}_i)}{\sum_{i=1}^{n} a_i(t)},$$

Where $\text{MOTORCENTER}(\text{CLUSTER}_i)$ is the motoric center of CLUSTER_i, which could be defined multiple ways, but which we choose to define as the average of all the junctures' motor locations.

When computed for each time step determined by the silhouette, the result of integration is an output trajectory through motor space that reflects the influences from perceptual space due to the exemplar. Figure 4 provides an example of the integration process over 11 time steps ($t = 11$). The combined motor and perceptual activation pattern is shown in motor space, where relative activation is depicted by the relative opacity of the clusters. The trajectory (whose direction is light green to light blue) moves through motor space over time, mainly within the path described by the silhouette. This silhouette path is shown by the region in motor space (the royal blue octagon) that is highlighted at each time step. The full output trajectory for the selected silhouette–exemplar pair is shown at time step 11 in motor space. It is also shown in perceptual space along with the exemplar trajectory. It is represented as a discontinuous trajectory in perceptual space to illustrate that this space has a different topology than motor space and because true discontinuities exist in perceptual space but never in motor space.

Finally, a reminder that not every path through motor space is physically possible because the dimensions of this space are not (usually) independent of one another (e.g., the cross-sectional areas of 8 regions of the vocal tract from lips to larynx and the time derivatives of each of these cross-sectional areas). That said, the CC model assumes a perceptual-motor map that has been structured by experience. Under this assumption, there are a high number of paths that exist between clusters. The path that the motor system chooses to follow is estimated based on the linear combination of cluster weighting. The output trajectory that results could be predicted internally or it could be the trace of movement that has happened. Either way, the output trajectory is a result of cluster activations that are commands to the motor system; it is not itself a control structure.

Learning and change in production

In the Core/CC model framework, an activated exemplar represents the perceptual goal of speech production. The jointly-activated silhouette constrains goal achievement by biasing movement toward familiar paths through motor space. In first and second language acquisition, these familiar paths are likely to diverge very substantially from the perceptual goal. Over time, path divergence narrows and production accuracy improves. This happens in one of two ways: (1) *via* change in the structure of the perceptual-motor map; (2) *via* change in the shape of existing silhouettes. The Core model addressed the former type of learning; the CC model captures the latter.

Practice-driven change

Recall that silhouettes are only established after the perceptual-motor map is at least partially structured through prelinguistic speech practice. First word production is based on the perceptual matching and selection process that was described under the Core Concepts section. This process gives rise to the first silhouettes. Once enough silhouettes have been established, speech production is fast and automatic because it is largely driven by silhouette–exemplar pairs that are activated when concepts are selected for communication. The repository of concepts with associated silhouette–exemplar pairs is the expressive vocabulary. It is about half the size of the speaker's overall vocabulary (Brysbaert et al., 2016). The other half is the receptive-only vocabulary. It includes only concept-associated exemplars that the speaker may choose to target at some point.

Production that is guided by the expressive vocabulary will entrench structure at the level of the perceptual-motor map because it constrains production to established motor paths. Accordingly, it will also slow the rate at which speech production patterns change. Some deviation from established paths is possible with the expansion of a silhouette due to random noise.[3] But, in general, the perceptual-motor integration of wordforms greatly reduces the exploration of new regions in motor space. Also, it is only with a return to a matching and selection process that new junctures and clusters can be generated (see Core Concepts). This means that practice-based changes to speech are initially more likely to occur at the level of wordform representation than at the level of the perceptual-motor map once an expressive vocabulary of a certain size is established. In the CC model, changes to the wordform occurs because practice results in silhouettes with weighted regions. These weighted regions encode frequency information and shift the silhouette in the direction of frequently used output trajectories

3 Recall that the silhouette incorporates motor traces of words that were successfully communicated. This allows for the influence of the periphery (i.e., articulation) on representation. The periphery introduces noise into the representation in any number of ways, including by virtue of poorly established "functional synergies" (see, e.g., Smith and Zelaznik, 2004).

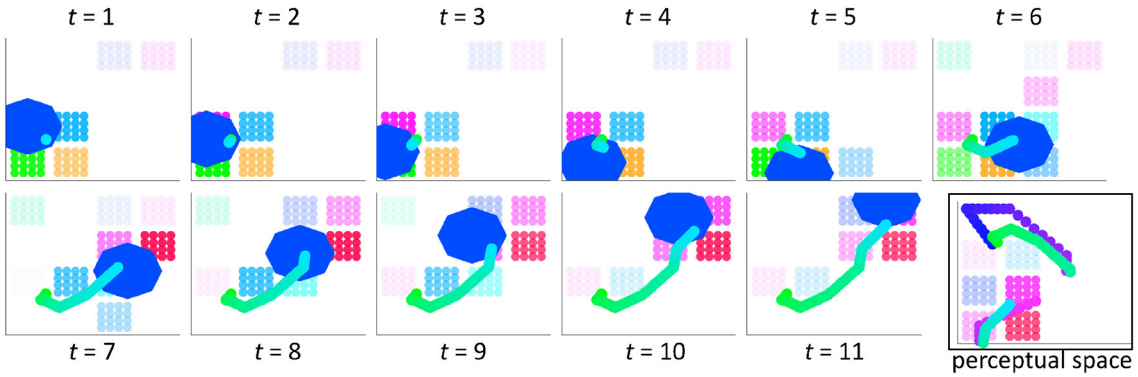


FIGURE 4
The perceptual-motor integration of wordforms results in an output trajectory through the perceptual and motor spaces, which are linked *via* clusters. Activation strength due to the silhouette and exemplar is depicted by the relative transparency/opacity of the clusters. These clusters are shown here in motor space for a silhouette that is 11 time steps in length. The region of the silhouette (blue octagon) is shown at each time step in this space. The resulting trajectory incorporates directional information (line shading from green to cyan blue). The trajectory is also shown in perceptual space, with the exemplar trajectory (dots shading from blue to pink). It is discontinuous in perceptual space because the topology of this space is different from that of motor space.

that meet with communicative success. The details of the weighting algorithm are as follows.

Weighted silhouettes

Recall that the silhouette highlights time-varying regions of motor space. The highlighted region is computed as the convex hull of the points associated with previously experienced trajectories (see Davis and Redford, 2019; Sections 2.5.2, 2.5.3). In the CC model, the convex hull is partitioned into simplices (n-dimensional "triangles"), each of which are assigned a weight. This means that, at each time, the highlighted region in motor space, returned by the function that is the silhouette, is a weighted homogenous simplicial complex. More specifically, let $\text{SIL}_{C,n}$ be the silhouette for concept C at a particular time in development, denoted by n. Assume the current silhouette is T (relative) time units long, and let k be a sufficiently large number. Then $\text{SIL}_{C,n}$ is defined to be a function with domain $[0, T]$ that takes an input of a particular time and gives an output of the weighted region corresponding to that time in the form of a weighted simiplicial complex. That is, $\text{SIL}_{C,n}(t) = (R_1, \ldots, R_k, v_1, \ldots, v_k)$, where each R_i is a simplex, and v_i is the weight of that simplex, and the following are satisfied:

1. $\bigcup_{i=1}^{k} \overline{R_i}$ is a homogenous simplicial complex, where $\overline{R_i}$ is the simplicial complex consisting of R_i and all of its faces; and
2. The union of the simplices, $\bigcup_{i=1}^{k} R_i$, is convex.

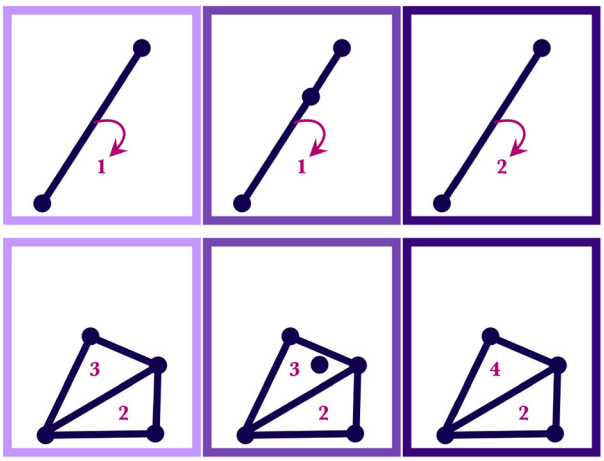


FIGURE 5
Both the upper and lower diagrams show how regions of a silhouette are reweighted as motor traces are absorbed by the wordform. In a given row, the leftmost panel shows the initial weighted region, $SIL_{C,n}(t)$; the middle panel shows the point, $M(t)$, that will be added; the rightmost panel shows the resulting region, $SIL_{C,n+1}(t)$, with new weights. The numbers indicate the weights of the simplices. The upper diagram shows a simplicial 1-complex and the lower diagram shows a simplicial 2-complex.

As before, the silhouette is built recursively by expanding it over time to include motor trajectories that have been successfully used to communicate a selected concept (see Figure 2). But now that the regions specified by a silhouette are weighted, new motor trajectories will either add weight to the regions that it passes through (see Case 1) or it will affect the overall shape of the silhouette (see Case 2). The two cases are briefly described here.

Assume the speaker uses $\text{SIL}_{C,n}$ to successfully communicate C using the motor trajectory M. Then the next iteration of the silhouette, $\text{SIL}_{C,n+1}$, will be defined at time t in the following way:

Case 1. If $M(t)$ is a point that is already in one of the simplices in $\text{SIL}_{C,n}(t)$, then $\text{SIL}_{C,n+1}(t)$ is the same as $\text{SIL}_{C,n}(t)$ except with the weight of the simplex (subregion) containing $M(t)$ increased by one. Similarly, if $M(t)$ is contained in multiple simplices—that is, if it lies on a shared boundary—then $\text{SIL}_{C,n+1}(t)$ is the same as $\text{SIL}_{C,n}(t)$ but with all the simplices containing $M(t)$ having their weight increased by one. This case is illustrated in Figure 5.

Case 2. On the other hand, if $M(t)$ is totally outside $\text{SIL}_{C,n}(t)$, then $\text{SIL}_{C,n+1}(t)$ is created by adding a minimal number of simplices to $\text{SIL}_{C,n}(t)$ to create a homogenous simplicial complex in which $M(t)$ is now contained, with the weights of the new simplices being 1. Examples of this case are illustrated in Figure 6.

The integration of a weighted silhouette, SIL_C, and an exemplar, e_C, will be similar to the integration described in the previous section

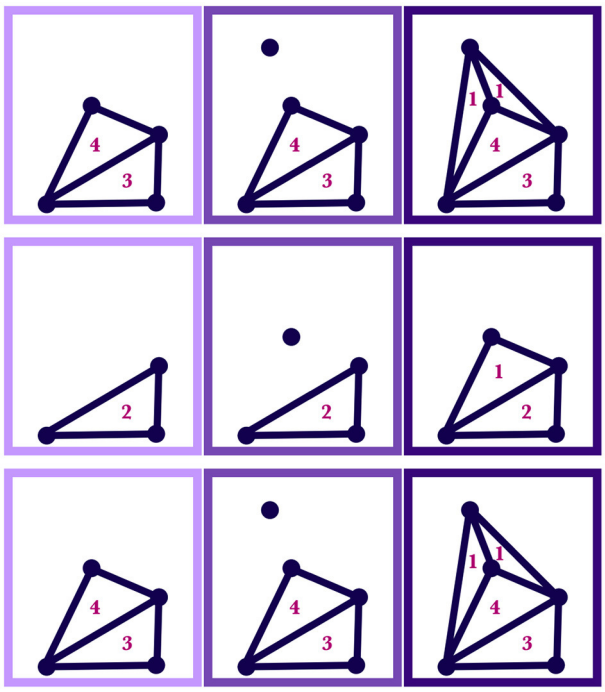


FIGURE 6
Each row of panels shows how a silhouette region, $SIL_{C,n}(t)$, changes with the inclusion of an additional point, $M(t)$. In a given row, the leftmost panel shows the initial region, $SIL_{C,n}(t)$, with the numbers indicating the weights; the middle panel shows the point to be added, $M(t)$; the rightmost panel shows the new resulting shape and weighting of the region, $SIL_{C,n+1}(t)$.

but must take into account the weighting. The only thing that changes is how we compute the motor activation of a juncture. Suppose $\text{SIL}_C(t) = (R_1, \ldots, R_k, v_1, \ldots, v_k)$. Then we define the weighted motor activation of $\text{JUNCTURE}_{i,j}$ to be the weighted average of the activations that come from each region:

$$\begin{aligned}\text{MOTORACTIVATION}_t(\text{JUNCTURE}_{i,j}) = \frac{1}{\sum_{s=1}^{k} v_s} \times \sum_{s=1}^{k} v_s \\ \times (\text{HIGHESTACTIVATIONMOTOR} \\ - (\text{DROPOFFSLOPEMOTOR} \\ \times (\text{DISTANCE}(R_s, \text{JUNCTURE}_{i,j}))))\end{aligned}$$

The effect of practice on accuracy

To examine the effect of practice on learning and change in the model, we can use the silhouette at iteration n to produce an output trajectory that is absorbed as a motor trace into the silhouette; the new silhouette is then used for production at iteration $n + 1$. When we do this repeatedly (= practice), learning occurs with changes to the silhouette. Figure 7 shows what this change looks like, step-by-step, in a 3-dimensional space. The space represents the topology of clusters in both motor and perceptual space since these were identical in the simulation to facilitate the visualization of silhouette movement toward the exemplar in perceptual-motor space.

Imagine that the z-axis in Figure 7 represents a close–open vocal tract dimension in motor space and the aperiodic–periodic sound dimension in perceptual space, which do roughly correspond to one another. This would mean that activation of clusters near the $x - y$ plane would result in consonantal-like articulations and that activation of clusters that are further above the $x - y$ plane would result in vowel-like articulations. The silhouette, exemplar, and output paths in Figure 7 all travel from clusters near the $x - y$ plane toward those furthest from this plane and then back again—a path that describes a CVC-shaped word. The upper-left panel shows a starting silhouette (blue triangular shapes) that might be an early representation of this word in that it is both far away from the exemplar trajectory (blue to pink dots) and is itself built up from only a few motor trajectories. With each of the 6 iterations of practice shown, the silhouette's path expands and changes shape: its weight gets distributed more toward the exemplar.

Practice-based changes to the silhouette mean that, with time, the output trajectory will draw nearer to those clusters that are especially activated by the exemplar. This effect of practice is more easily visualized in 2-dimensional space than in 3-dimensional space. Figure 8 therefore displays the results of a simulation in 2D space where, similar to Figure 7, clusters are separated to model vowel- vs. consonant-like articulations and the motor and perceptual spaces have identical layouts. With this in mind, the exemplar trajectory shown in purple in the figure again describes a CVC trajectory. The silhouette in blue highlights a path that diverges from this trajectory. The output trajectory, which is linearly interpolated in red, is shown as a dotted line after the first time the exemplar and silhouette are integrated; it is shown as a dashed line after 50 iterations of the simulation and as a solid line after 200 iterations. Overall, the figure illustrates the expansion of the output trajectory in the direction of the larger exemplar trajectory with changes to the silhouette resulting from speech practice.

Intriguingly, the simulation result shown in Figure 8 indicates a period of relatively rapid change in production followed by a longer period of very marginal change. This unanticipated result is qualitatively similar to well-described patterns of early gains followed by plateaus in the motor learning literature (Adams, 1987; Newell et al., 2001). It also suggests that unsupervised speech practice is unlikely to drive substantial changes to production after a certain point. This is probably a good thing. After all, the persistent effect of "accent" in highly-proficient second language speakers would be hard to account for in the model if sheer practice were sufficient for a speaker to match exogenously-derived exemplars. Still, the result also suggests that other mechanisms besides practice are needed to describe the steep and relatively prolonged increase in speech production accuracy that is observed during the first 3 years of childhood. One possibility, not modeled here, is that feedback from listeners shapes learning— especially in children's speech when utterances are too short to present much in the way of context for the listener. This possibility is already an assumption of the overarching theory. Recall, that motor traces are only absorbed into the silhouette if communication is successful (Redford, 2019). Another possibility is that the production process can be perturbed to facilitate learning in such a way that merits, say, a return to the (slow) matching and selection process. If the speaker returns to the process of finding best perceptual matches between established motor trajectories and novel exemplars, new junctures may be created where different established trajectories near each other in motor space. The creation of new junctures may change the shape of existing clusters or establish new ones, thus changing the overall the structure of the perceptual-motor

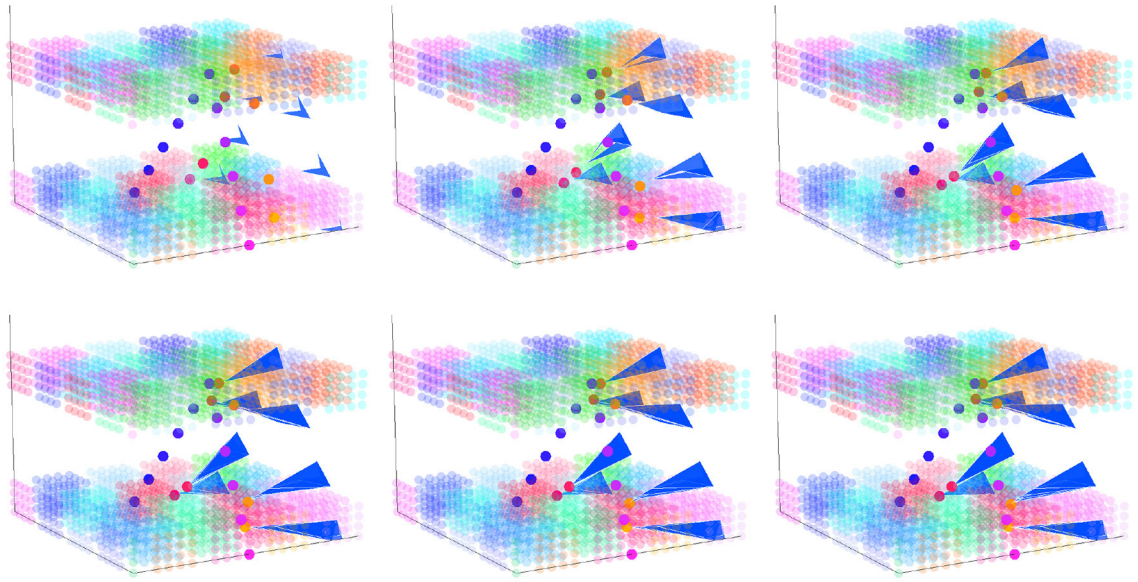


FIGURE 7
Silhouette change over time with each iteration of practice in a 3D perceptual-motor space where the perceptual and motor spaces have identical layouts. The upper-left panel shows the starting silhouette (blue triangular shapes), the exemplar trajectory (blue to pink dots), and the output trajectory (red to orange dots). Reading from right-to-left and then top-to-bottom, the figure illustrates how the silhouette changes in shape as it incorporates the output trajectory from each prior production.

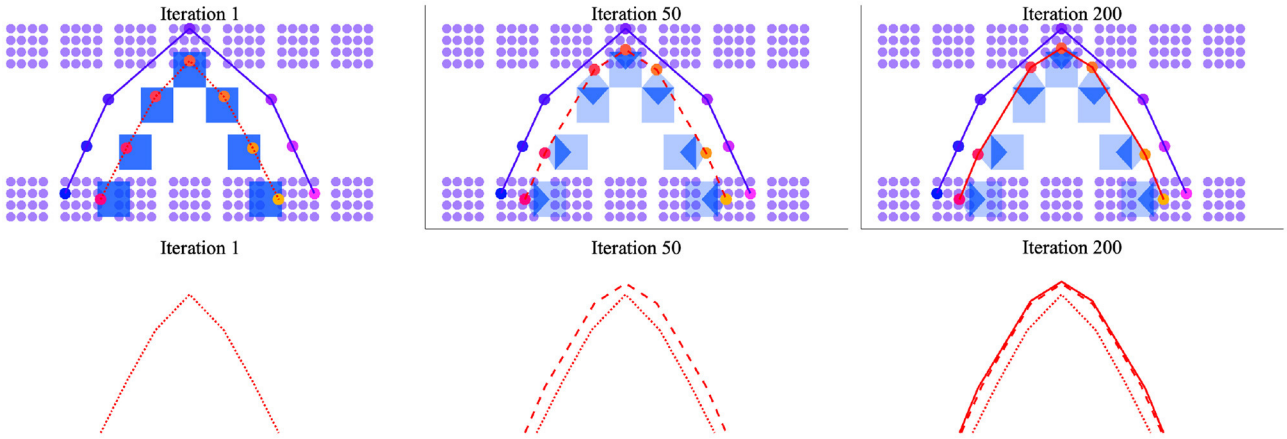


FIGURE 8
Change in output trajectory over time with iterations of practice in a 2D perceptual-motor space where the perceptual and motor spaces have identical layouts. **(Upper)** Silhouettes are shown as blue squares **(left)** or blue polygons of varying-opacity **(center and right)** to indicate weighting; the exemplar trajectory is traced in purple; the output trajectory in red. **(Lower)** The output trajectory is depicted after 1 iteration (dotted line), 50 iterations (dashed line), and 200 iterations (solid line). Reading from **(left-to-right)**, the output trajectory is shown to change shape to better approximate the exemplar trajectory over time.

map in the direction of new ambient language input. Alternatively, the speaker may focus on the acoustic-perceptual shape of the word resulting in the up-weighting of contributions from the exemplar to overall cluster activation patterns during the integration process with consequences for the shape of the output trajectory. The theory allows for all of these alternatives.

Novel word production

Although it is necessary to account for changes to known word production in a developmentally sensitive theory of production, it is not sufficient. This is especially true under the assumption of whole-word production as this assumption begets the problem of novel word production. Since we hypothesize that the default production strategy is silhouette–exemplar integration once an expressive vocabulary is established, the CC model adopts a silhouette-based approach to novel word production. Although the approach is motivated by the model architecture, it also allows us to capture an empirical finding from the literature on nonword repetition: the effect of vocabulary size on production accuracy in children's speech and in adult second language speech.

Not surprisingly, older children repeat nonwords more accurately than younger children and adults with more exposure to a second

language repeat nonwords in the target language more accurately than those with less exposure. But accuracy also varies independently from age and experience with vocabulary size: children with smaller vocabularies repeat nonwords less accurately than children with larger vocabularies (e.g., Metsala, 1999; Verhagen et al., 2022); college-aged adults with smaller second language vocabularies produce less native-like renditions of nonwords than those with larger vocabularies (Bundgaard-Nielsen et al., 2012). Importantly, it is a child's expressive vocabulary size that correlates with production accuracy; not their overall vocabulary size (Edwards et al., 2004; Munson et al., 2005). In addition to vocabulary size, the production accuracy of novel words, or nonwords, varies with properties of the given nonword, including its "wordlikeness" and the relative frequency of its phonological patterning (e.g., Edwards et al., 2004; Guion et al., 2004; Munson et al., 2005; Redford and Oh, 2016). In brief, nonwords that obey the phonotactics of the (target) language and/or contain high frequency phonotactic patterns are repeated more accurately than those that are less "wordlike" with respect to phonotactics and/or contain less frequent patterns. The latter findings suggest that nonword production relies on existing wordform representations (Edwards et al., 2004; Guion et al., 2004; Redford and Oh, 2016).[4] The CC model implements this hypothesis. When there is no silhouette for a given word, the speaker leverages the silhouettes that do exist to generate an archi-silhouette, or an A-silhouette, to provide the time-varying information needed to guide production. The A-silhouette is built by pulling together silhouettes from the nearest phonological neighbors of the targeted novel word form. In the psycholinguistic literature, phonological neighbors are wordforms that differ from one another by one phoneme (Luce and Pisoni, 1998). In the CC model, they are based on similarity in perceptual space, which is defined using the distance metric on that space. The algorithm for building an A-silhouette is described next.

Building an A-silhouette

Recall that the CC model has a function that measures distances between points in perceptual space. Let d_{PERC} be a function that measures the distance between perceptual trajectories (see Davis and Redford, 2019). The function operates by (1) aligning trajectories in perceptual space so their endpoints line up, using linear interpolation if necessary to fill in points, so that every point in one trajectory corresponds to one in the other, (2) finding the distances between corresponding points, and then (3) taking the average of these distances.

Now, suppose the speaker is attempting a new word W with exemplar E. Let k be a parameter with a fixed value representing the number of similar words from which to build an A-silhouette for W. For each word w_i ($i = 1, 2, 3, \dots$) in the expressive lexicon, let e_i be its corresponding exemplar and let SIL_i be its corresponding silhouette. Assume that the expressive words are already ordered by perceptual closeness to W; that is, $d_{\text{PERC}}(w_1, W) \leq d_{\text{PERC}}(w_2, W) \leq d_{\text{PERC}}(w_3, W) \leq \dots$ Then $w_1, w_2, \dots, w_k$ are the k perceptually closest words to W in the expressive lexicon, and their silhouettes, $\text{SIL}_1, \text{SIL}_2, \dots, \text{SIL}_k$, are chosen to build the A-silhouette.

We assume that the chosen silhouettes have already been modified so that they are aligned with each other in time. The A-silhouette is a silhouette ASIL such that at each time t, ASIL is defined as a combination of $\text{SIL}_i(t)$ for $i = 1, 2, \dots, k$. More specifically, fix t and let $\text{SIL}_i(t) = (R_{i,1}, R_{i,2}, \dots, R_{i,n_i}, v_{i,1}, v_{i,2}, \dots, v_{i,n_i})$ where $R_{i,1}, R_{i,2}, \dots, R_{i,n_i}$ are the n_i subregions making up $\text{SIL}_i(t)$ and $v_{i,1}, v_{i,2}, \dots, v_{i,n_i}$ are their respective weights. The weights are scaled so that the maximum weight at time t is the same for each silhouette. That is, let $\text{MAXWEIGHT}_i = \max(v_{i,j})_{j=1,2,\dots,n_i}$, meaning MAXWEIGHT_i is the maximum weight of the regions in the ith silhouette (at time t). Then we use $v'_{i,j}$ to denote the scaled version of $v_{i,j}$, and we define $v'_{i,j} = \dfrac{v_{i,j} \times \max(\text{MAXWEIGHT}_i)_{i=1,2,\dots,k}}{\text{MAXWEIGHT}_i}$. That is, for each region, we take the original weight, multiply it by the maximum weight of all the regions in all the silhouettes, and then divide that by the maximum weight of the regions in that silhouette. Finally, the regions from all the silhouettes at time t are combined using the new weights. The combination process is demonstrated first with an example. The general process is given afterwards.

Suppose we have 3 aligned silhouettes, $\text{SIL}_1, \text{SIL}_2, \text{SIL}_3$, and suppose that at time 2, each silhouette consists of two regions, $R_{1,1}$ and $R_{1,2}$; $R_{2,1}$ and $R_{2,2}$; and $R_{3,1}$ and $R_{3,2}$, respectively, where they overlap as shown in Figure 9. Suppose these regions have respective weights $v_{1,1} = 3$ and $v_{1,2} = 4$; $v_{2,1} = 5$ and $v_{2,2} = 8$; and $v_{3,1} = 2$ and $v_{3,2} = 1$. That is,

$\text{SIL}_1(2) = (R_{1,1}, R_{1,2}, 3, 4)$ where $R_{1,1}$ and $R_{1,2}$ are the pink triangles

$\text{SIL}_2(2) = (R_{2,1}, R_{2,2}, 5, 8)$ where $R_{2,1}$ and $R_{2,2}$ are the purple triangles

$\text{SIL}_3(2) = (R_{3,1}, R_{3,2}, 2, 1)$ where $R_{3,1}$ and $R_{3,2}$ are the blue triangles

Then scaling the weights as described above yields a maximum weight of 8 for each region; that is,

$$v'_{1,1} = 6, \quad v'_{1,2} = 8, \quad v'_{2,1} = 5, \quad v'_{2,2} = 8, \quad v'_{3,1} = 8, \quad v'_{3,2} = 4.$$

Then we will define the combination of these regions, $\text{ASIL}(2)$, to be the weighted region shown in red. That is, $\text{ASIL}(2) = (T_1, T_2, T_3, T_4, T_5, T_6, T_7, T_8, T_9, T_{10}, T_{11}, T_{12}, T_{13}, T_{14}, T_{15}, 8, 4, 1, 10, 20, 10, 9, 5, 6, 16, 16, 8, 1, 8, 1)$, where T_i are the red triangles shown in Figure 9.

Returning to the general case where the selected silhouettes are $\text{SIL}_1, \text{SIL}_2, \dots, \text{SIL}_k$, we define $\text{ASIL}(t) = (T_1, T_2, \dots, T_n, v_1, v_2, \dots, v_n)$ where $T_1, T_2, \dots, T_n$ is a triangulation of the convex hull of all the regions making up all the $\text{SIL}_i(t)$. For each i, the weight v_i of the region T_i is defined as follows: either (1) v_i is equal to the sum of the weights of all the original regions that T_i lies inside, or (2) $v_i = 1$ if it lies in none of the original regions but is still part of the convex hull.

That is, $\text{ASIL}(t) = (T_1, T_2, \dots, T_n, v_1, v_2, \dots, v_n)$ such that

1. $T_1 \cup T_2 \cup \dots \cup T_n = \text{CONVHULL}(R_{1,1}, R_{1,2}, \dots, R_{1,n_1}, R_{2,1}, R_{2,2}, \dots, R_{2,n_2}, \dots, R_{k,1}, R_{k,2}, \dots, R_{k,n_k})$
2. Each T_i is a simplex (an "n-dimensional triangle")
3. The regions do not overlap each other more than at a boundary: $\text{interior}(T_i) \cap \text{interior}(T_j) = \varnothing$ for all $1 \leq i < j \leq n$
4. For every set $A = \{R_{i_1,j_1}, \dots, R_{i_m,j_m}\}$, either $\bigcap_{a \in A} a = \varnothing$ or $\bigcap_{a \in A} a = T_{k_1} \cup T_{k_2} \cup \dots \cup T_{k_s}$ for some $k_1, k_2, \dots, k_s \in \{1, 2, \dots, n\}$

4 For a substantially different interpretation of these findings see Gathercole (2006).

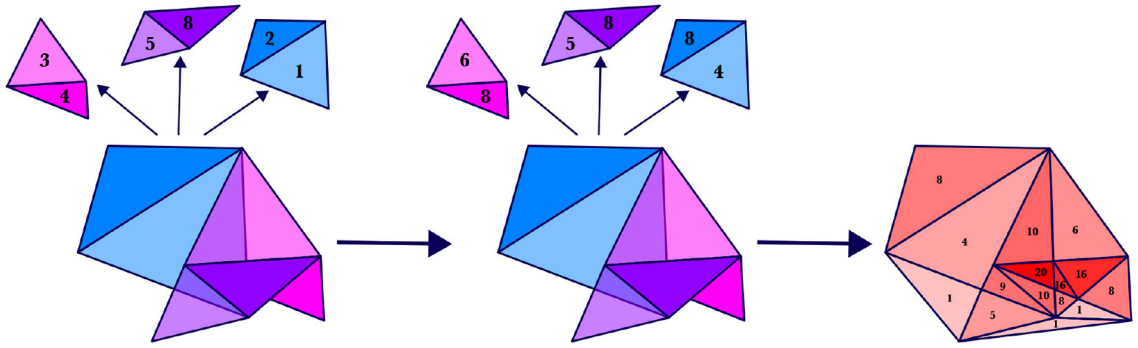


FIGURE 9
(Left) Regions corresponding to three motor silhouettes at a particular time; each silhouette at this time has a two subregions, with weights labeled. **(Center)** The regions with scaled weights. **(Right)** The weighted combination of the regions.

5. $$v_\ell = \begin{cases} \sum_{R_{i,j} \text{ containing } T_\ell} v'_{i,j} & \text{if at least one } R_{i,j} \text{ contains } T_\ell, \text{ i.e. if this sum is nonzero} \\ 1 & \text{otherwise} \end{cases}$$

The effect of vocabulary size on accuracy

According to the process outlined above, exemplars of words that belong only to the receptive vocabulary are attempted by combining the silhouettes of perceptually similar words that belong to the expressive vocabulary. But how good is this combined form? To what extent will it allow for a path through motor space that overlaps with the clusters activated by the novel exemplar in perceptual space? In this section, we demonstrate that the answer to these questions depends on the size of the expressive vocabulary. More specifically, we show that the goodness of the A-silhouette depends on the goodness of the perceptual matches to the novel wordform. The goodness of the perceptual matches in turn depends on the size of the speaker's expressive vocabulary, V, in relation to the larger vocabulary, L.

The larger vocabulary, L, is a theoretic construct that represents the set of words in a language over which the phonology is defined. The size of L depends on what exactly it represents. L could represent the size of a dictionary vocabulary or the size of an adult's overall vocabulary (10,000 words to 200,000 words) or the expressive vocabulary only, that is, half of the overall vocabulary size (Brysbaert et al., 2016). Alternatively, L could represent the total number of words required for normal every-day communication. We estimate that number here as 2500 words. This number is based on Nation and Waring's (1997) synthesis of research findings on the relationship between vocabulary size and second language acquisition for pedagogical purposes. Nation and Waring suggest that "a vocabulary size of 2,000–3,000 words provides a very good basis for language use." This suggestion is based on the vocabulary size needed to achieve over 90% coverage of English texts aimed at young adult readers (e.g., 2,600 words result in 96% text coverage and a density of 1 unknown word occurring every 25 words). Insofar as young adults are perfectly good speakers of their native language, a vocabulary of roughly 2500 wordforms should adequately cover the phonological space of a language. It therefore provides a good basis for L.

Given that the words in L describe the phonological space for a particular language, it is clear that a subset V of L may fail to do so. And, if it fails to do so, then the A-silhouettes that are built up from wordforms in V are unlikely to reliably provide accurate information regarding the best path to take through motor space in order to approximate an exemplar that represents a novel word target. In particular, suppose W is the novel word, and suppose the A-silhouette is going to be built from the k words in V that are perceptually closest to W. What is the probability that these k words from V are actually some of the closest words to W in all of L? To make it more concrete, let $k = 3$ and let "best" be a synonym for "perceptually closest to W." We can ask:

- What is the probability that the 3 best words in L are contained in V (and thus are also the 3 best words in V)?
- What is the probability that 3 of the 4 best words in L are contained in V?
- What is the probability that 3 of the 5 best words in L are contained in V?

More generally:

- What is the probability that 3 of the $3 + r$ best words in L are contained in V?

And even more generally:

- What is the probability that k of the $k + r$ best words in L are contained in V?

Naturally, this probability increases as the size of V increases. In particular, if n is the number of words in L and m is the number of words in V, the probability that k of the $k + r$ best words in L are contained in V, i.e. that the k best words in V are a subset of the $k + r$ best words in all of L, is:

$$\sum_{i=0}^{r} \frac{(k+r)!}{(k+i)!(r-i)!} \times \frac{(n-k-r)!}{(m-k-i)!(n-r-m+i)!} \times \frac{m!(n-m)!}{n!}$$

(assuming $k \leq m$ and $k + r \leq n$). This is illustrated in Figure 10 for an L of size 2,500, and various values of k and $p(= k + r)$. As the

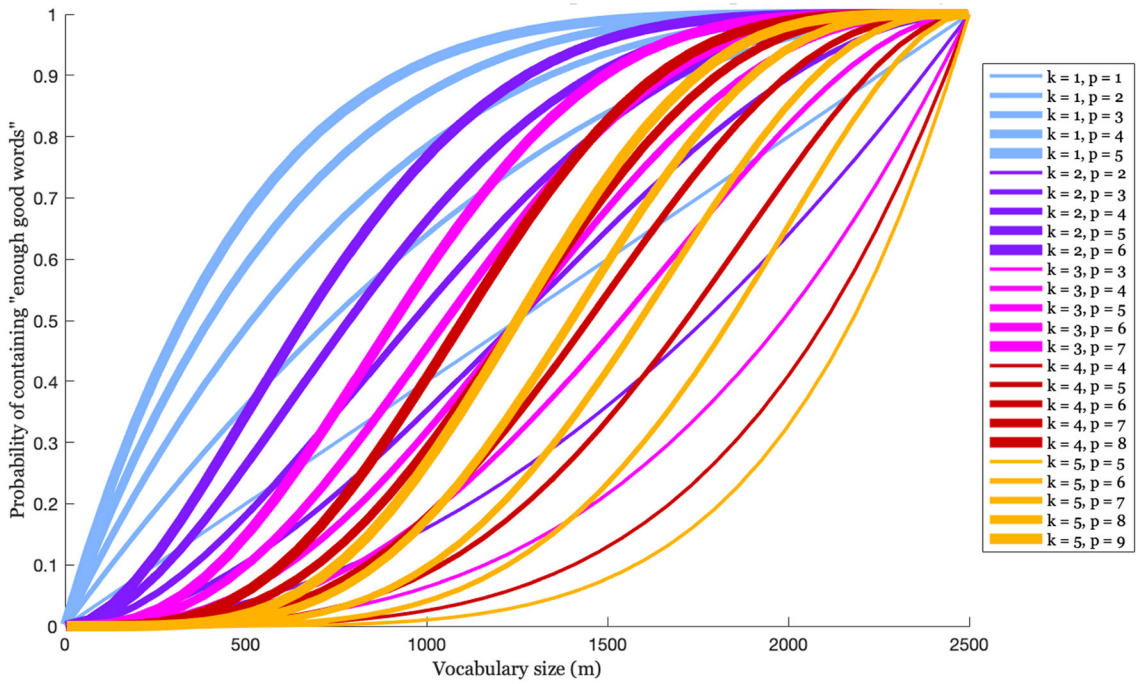


FIGURE 10
The probability that an expressive vocabulary of size *m*, drawn randomly from a larger vocabulary size of 2,500 words, contains at least *k* of the *p* words in *L* that are the closest perceptual matches to an exemplar that represents some novel word.

size of V increases (as we move right on the x-axis) the probability that the speaker's expressive vocabulary includes enough of the larger vocabulary's perceptually closest words to W also increases. This increase differs somewhat depending on the value of k, which, recall, is the number of words that are chosen to create the A-silhouette, and the value of $p(= k + r)$, which is the number of words in L that are perceptually "close enough" to the novel word that any subset of k of those words could be used to create a very good A-silhouette for guiding production. The data here suggest that if V is 500—which is approximately the size of a typically-developing 3-year-old's expressive vocabulary (Shipley and McAfee, 2019)[5]—then it has a good chance (about 70%) of containing at least 1 of the 5 best words in L, but a poor chance (about 15%) of containing at least 3 of the 7 best words. This observation begs the question of how many closest perceptual wordforms are needed to generate an A-silhouette that will yield a good approximation of the novel word target. The data in the figure suggests that if in general any 3 of the closest 6 words to a goal word will yield a good A-silhouette, then good A-silhouettes can be reliably generated when V is 70% of L, or 1,750 words.

The predicted effect of an A-silhouette that is built up from a subset of "close enough" silhouettes is an output trajectory that approximates the exemplar of the novel word that is being attempted. Less good A-silhouettes result in less accurate output trajectories. To test this prediction, and so the effect of vocabulary size on the production accuracy, we simulated novel CVCV word production given different expressive vocabulary sizes and an all-CVCV language of 1,296 words. The language was built up from paths through a 2D motor space and a 2D perceptual space. The spaces had 6 clusters deemed consonantal articulations and 6 clusters deemed vocalic articulations. These groups of 6 were separated from one another in the y direction in motor space. The transformation from motor space to perceptual space was one that maintained this consonant-vowel separation, but shuffled the clusters in the x direction to render different topologies for the two spaces.[6] The 1,296 wordforms were all the possible paths going from center-of-cluster to center-of-cluster in a CVCV-like pattern ($1,296 = 6$ consonants $\times$ 6 vowels $\times$ 6 consonants $\times$ 6 vowels). The silhouettes consisted of 7 uniformly-weighted square regions, with regions 1, 3, 5, 7 centered on the appropriate CVCV clusters, and regions 2, 4, 6 falling evenly between them. The exemplar paired with a silhouette was built by taking the motor trajectory going through the center of the silhouette and finding the corresponding perceptual trajectory based on the transformation between the spaces.

5 This assumes an expressive vocabulary that is half the size of the overall vocabulary, which Shipley and McAfee (2019) place at about 1,000 words for a typically-developing 3-year-old.

6 Specifically, the clusters were 4×4 squares of 16 junctures, with the horizontal distance between two adjacent junctures within a cluster being 1 and the horizontal distance between two adjacent clusters being 2. The vertical distance between adjacent junctures within a cluster was 1 and the vertical distance between the bottom row of clusters and the top row of clusters was 15. Let us designate the bottom-row clusters as "consonants" and the top as "vowels." The transformation between motor and perceptual space can then be described as follows: If in motor space, the consonants from left to right were $C_1, C_2, C_3, C_4, C_5, C_6$, then in perceptual space they were $C_3, C_4, C_1, C_2, C_5, C_6$; if in motor space, the vowels from left to right were $V_1, V_2, V_3, V_4, V_5, V_6$, then in perceptual space they were $V_3, V_4, V_5, V_6, V_1, V_2$.

In the simulation, the novel word was an exemplar randomly selected from the language. The initial expressive vocabulary consisted of 5 silhouette–exemplar pairs randomly selected from the 1,296-word language (minus the novel word). An A-silhouette was built from the 3 words in the expressive vocabulary that were perceptually closest to the novel word. An output trajectory was computed based on the integration of the A-silhouette and the novel word exemplar. The distance in perceptual space between the output trajectory and the novel word exemplar trajectory was calculated to measure the accuracy of the output trajectory. The initial vocabulary was then increased to 10 words by adding an additional 5 random CVCV words to the expressive vocabulary. A new A-silhouette was made, again using the 3 closest words, an output trajectory computed, and the distance in space from the exemplar calculated. The expressive vocabulary was next increased to 20, then 40, and so on for a range of sizes up to 1,200. For each vocabulary size, the output trajectory based on A-silhouette–exemplar integration was found and the distance from the novel word exemplar calculated.

The entire simulation was run 20 times with different randomly-selected novel words and expressive vocabularies. Figure 11 shows the mean distance between output and exemplar trajectory as a function of vocabulary size for the 20 runs. The data indicate increasing production accuracy with increasing vocabulary size. The increase is steeper early on and more gradual later on. The pattern qualitatively matches the very robust increases in production accuracy seen during the earliest stages of speech acquisition followed by slower gains but continuing improvement.

Summary and conclusion

The CC model captures the observation that speech develops with language use to address the problem of learning and change in production. The child's first words represent both a first attempt at speech and a first attempt to communicate using language. Control over speech action evolves in this communicative context with speech practice. And we engage in a whole lot of practice. The estimate from voice recordings of college-aged adults is that we speak about 16,000 words a day (Mehl et al., 2007). This kind of practice must have implications for speech production. In our theory it does.

The theory assumes a dual lexicon and whole-word speech production. The motor wordforms (silhouettes) in the lexicon are endogenous representations built up with speech practice. The perceptual wordforms (exemplars) are exogenous representations that reflect ambient language patterns. Speech production is the integration of these forms in the perceptual-motor map. The perceptual-motor map is discretized with vocal-motor practice, including speech practice, into language-specific clusters that represent units of speech motor control. The perceptual aspect of these units can be related to sound categories or to perceptual features; the motor aspect to vocal tract constrictions similar in some respects to the "gestures" of Articulatory Phonology except that do not necessarily code meaning contrast. They are units that represent both acoustic-auditory goals and spatial targets for the speech motor system.

When a word is selected for output from the expressive vocabulary, its silhouette and exemplar activate clusters in motor and perceptual space. The silhouette contributes time-varying information about movement through motor space within a window of activation that allows contextual effects to emerge (i.e., syntagmatic relations). The exemplar provides static information about the acoustic-auditory goals to be achieved for successful communication (i.e., paradigmatic relations). Perceptual-motor integration of the forms results in an output trajectory that traces speech movement due to the integration process. If the speech movement described by an output trajectory results in successful communication, then its trace is absorbed into the silhouette for the concept intended and communicated. By this mechanism, the silhouette for a word is shifted in the direction of the exemplar(s) of a word. This is the practice-based mechanism for motor learning and change in the model. Simulation results suggest that practice has a large initial effect on production accuracy, and that this effect plateaus relatively quickly, or is, at least reduced to only a very marginal effect over time. Overall, the pattern recalls the power law function of motor learning (see Newell et al., 2001).

Learning and change in the model also occurs with novel word production. In a system where silhouette–exemplar integration is the dominant mode of production, the accurate rendition of a novel word requires a silhouette-like form to achieve the targeted exemplar. The new silhouette, an A-silhouette, is created by combining existing silhouettes, which are selected based on the closeness of their perceptual counterparts to the novel-word exemplar. The algorithm for combining existing silhouettes to generate an A-silhouette relies on the model-internal fact that the expressive lexicon is structured according to the perceptual and motor spaces within which the dual wordforms reside. The receptive-only lexicon is also structured by the perceptual space within which single wordforms reside alongside their dual wordform neighbors. Although merely a logical consequence of the CC model architecture, the phonetically-structured lexicon of our theory parallels the well-established psycholinguistic hypothesis of a phonologically organized lexicon (Pisoni et al., 1985; Luce and Pisoni, 1998).

The integration of an A-silhouette and an exemplar associated with a novel word results in an output trajectory. The extent to which this output trajectory is similar to the exemplar varies naturally with vocabulary size. Smaller vocabularies do not regularly allow for the same quality of perceptual matches as larger vocabularies and so the A-silhouettes that are created based on a small vocabulary result in poorer production accuracy than those created based on larger vocabularies. This implication of the model is consistent with the effect of vocabulary size on nonword repetition accuracy in children's speech and in adult second language speech.

Why core?

The CC model provides an intellectual framework within which to understand developmental changes in speech production. For this reason, it also provides a framework for understanding the emergence of individual differences in speech production, including differences due to developmental disorder. The model perspective is that these differences are the result of developmental trajectories that are themselves defined by iterative processes that may compound over time the effects of small differences in initial parameter settings.

No existing linguistic or psycholinguistic theory of speech production that we know of has been advanced with the particular

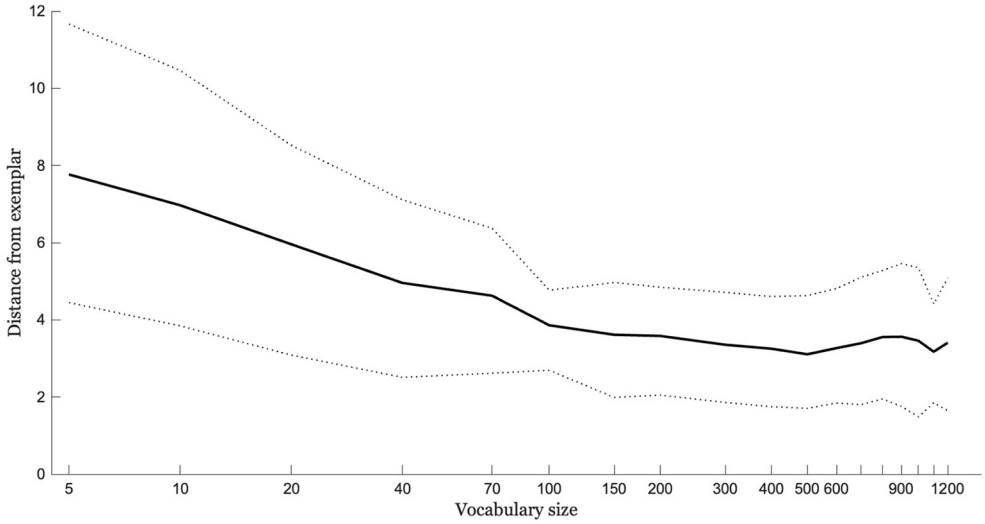


FIGURE 11
The average distance of the output trajectory from a novel word exemplar given integration based on an A-silhouette built from the 3 perceptually-closest silhouettes to the novel word exemplar. Distances are shown as a function of vocabulary size (log-transformed). Dotted lines show ±1 standard deviation around the mean for the simulation, which was run 20 times.

aim of explaining change in a manner that naturally gives rise to different outcomes. To the best of our knowledge, every instantiated theory that handles adult spoken language production assumes (more or less) current descriptions of the adult speech behavior as its starting and ending point. They are teleological in this way. For this reason, individual differences are often treated as specific deviations from normativity rather than as the product of differing initial conditions and constraints on development. The teleological frame is, in part, the legacy of Saussure and his emphasis on the synchronic over the diachronic. It is, in part, the legacy of Chomsky and his emphasis on what is universal and so what might be innate. Collective knowledge about speech and language has grown enormously under these legacies. Our goal is to reframe some of this existing knowledge within an emergentist framework to better understand individual differences and to encourage new avenues of empirical research.

Future directions

The Core/CC model framework emphasizes the role of variability in learning and change. Recall that speakers can only target previously experienced paths through motor space, even when attempting a new perceptual goal (sound or word). Under this hypothesis, noise in the periphery due to immature motor control provides an important learning benefit, not least of which is better and more thorough exploration of the motor space than would otherwise be possible; and it is through exploration that junctures proliferate in the perceptual-motor map in the first place. Clusters, the units of speech motor control, are created from these junctures. Clusters allow speakers to achieve language-specific acoustic-auditory goals. The proliferation of junctures in motor space is a prerequisite for doing so. The highly variable speech movements of children's speech compared to adults' speech may therefore be what allows them to acquire native-like speech sound articulation in a second language—something that adult learners are purportedly unable to do. The prediction is then for an increase in perceived accentedness in speech with age of acquisition, but one that tracks more specifically with age-related changes in the variability of speech movements. Age-of-acquisition effects are, of course, well-described in studies of second language speech—in fact, the age of 5–7 years has been suggested as a cut-off for nativelike acquisition of a second language speech category (e.g., Guion, 2003)–but the explanation for why this might be is elusive. Our prediction suggests that the cut-off is causally tied to the rapid leveling off of articulatory variability during developmnet (see, e.g., Smith and Zelaznik, 2004). Also, note that, just as children's speech continues to exhibit greater variability than adult speech until age 12–14 years, so too the age-of-acquisition effect on second language speech is graded—there is not an abrupt cut-off in native-like attainment of a second language at age 5 or 7 years across all individuals. Future research on second language acquisition could investigate the extent to which greater variability in the realization of sounds at one stage in development predicts more accurate (= target-like) attainment of these sounds at a later stage.

The Core/CC model framework also predicts a relatively abrupt transition from a period of exceptionally high variability in the production of novel words to a period of relative stability in word production that corresponds to a change in strategy from the matching and selection of existing motor trajectories to create best perceptual approximations of novel exemplar trajectories to a strategy based on an expressive vocabulary and so on the integration of perceptual and motor wordforms. Consistent with this, Vihman (2014) describes a shift in word production around 2 years of age that she attributes to a shift away from a strategy of schema-based production and toward template-based word production. Our A-silhouettes might be considered templates in that they are not word-specific, but rather an amalgam of similar sounding words. Vihman (2014) also notes that a

schema-based and templatic-based production strategy may co-exist for some time during development, and that some children never really exihibit a phase that can truly be described as templatic. The CC model suggests that the path toward understanding these individual differences is through more careful study of the relationship between expressive vocabulary and phonological development during the young preschool years. This study should include not just the size of the expressive vocabulary, but also its detail regarding its phonological structure in perceptual and motor spaces.

In the CC model, the extent to which A-silhouettes allow for matching exogenous wordform representations varies with the size and structure of the expressive vocabulary. As already noted, this pattern is consistent with the effect of vocabulary size on nonword repetition accuracy in children's speech. But a detailed consideration of this relationship leads us now to wonder about an inflection point in development when production is no longer driven by the integration of the specific perceptual and motor wordforms that are stored together in an expressive vocabulary. Rather, it could be driven by the integration of perceptual wordforms and A-silhouettes. What this might mean is a question for future research. But, to give that research some structure, let us consider the problem in a little more detail.

Under the simplifying assumption that an expressive vocabulary is some random subset of the words in a language, it is clear that an A-silhouette will provide as good guidance as a more specific motor wordform once the expressive vocabulary reaches a certain size. The question then becomes: What is that certain size? This depends in part on the number of words needed to adequately describe the language. In our simulations, the language vocabulary was 2,500 words. This number of words was chosen on the grounds that between 2,000 and 3,000 words is adequate for everyday communication in English. We presume that this means that a specific set of 2,500 words adequately describes the phonology of English. But the number 2,500 was also chosen with young children's speech patterns in mind. In particular, 30% of 2,500 words is 750 words, which is a good approximation of a 3-year-old's expressive vocabulary size. And, since we know that 3-year-old speech is different from adult speech, it was convenient to consider the potential shape of A-silhouettes in this context. But the reader will have also noted that 2,500 words falls well short of the average expressive vocabulary size of a typical adult. In fact, the lower bound estimate of an average adults' expressive vocabulary size is 10,000 words; and, 30% of 10,000 words is even larger than our language vocabulary estimate. Given this, by the logic of our own model, 10,000 distinct silhouettes are clearly not required to produce 10,000 words. This observation suggests several paths for future research, including a version of the prior suggestion: more careful studies of the structure and size of developing expressive vocabularies are needed to better understand the relationship between the accuracy with which a novel word can be produced and the size of the expressive vocabulary.

Finally, the developmental perspective adopted here motivates our view that perceptual experience and motor practice interact and build on each other through time; together, they provide the foundation for an individualized account of spoken language patterns. The Core/CC model framework assumes the evolution of speech perception and of perceptual wordform representations, but addresses only the effects of motor practice on change. This limitation argues for future research that has as its aim to understand, in precise terms, how much of developmental change in the sound patterns of speech is due to perception and how much is due to production. It will also be important to determine how exactly to tell the difference between the two. The Core model framework suggests, consistent with much other theory, that perceptually-driven changes should be in the direction of increasing contrasts, and that motor-driven changes are in the timing domain. But timing differences also give rise to contrast. This is, in fact, the foundational insight on which Articulatory Phonology was built (i.e., language-specific gestural coordination). So, again, under the now well-articulated assumption of a dual lexicon, future research will need to detail the separate and interacting contributions from perceptual learning and speech motor learning to understand the emergence and evolution of individualized speech patterns.

Data availability statement

Publicly available data for this study can be found at: https://github.com/mayaekd/core.

Author contributions

The research reported here was fully collaborative. Both authors contributed to the writing and approved the submitted version of the manuscript.

Funding

This research was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) under grant R01HD087452 (PI: MR).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

The content is solely the authors' responsibility and does not necessarily reflect the views of NICHD.

References

Adams, J. A. (1987). Historical review and appraisal of research on the learning, retention, and transfer of human motor skills. *Psychol. Bull.* 101, 41–74. doi: 10.1037/0033-2909.101.1.41

Browman, C. P., and Goldstein, L. (1992). Articulatory phonology: an overview. *Phonetica* 49, 155–180. doi: 10.1159/000261913

Browman, C. P., and Goldstein, L. M. (1986). Towards an articulatory phonology. *Phonology* 3, 219–252. doi: 10.1017/S0952675700000658

Brysbaert, M., Stevens, M., Mandera, P., and Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Front. Psychol.* 7, 1116. doi: 10.3389/fpsyg.2016.01116

Bundgaard-Nielsen, R. L., Best, C. T., Kroos, C., and Tyler, M. D. (2012). Second language learners' vocabulary expansion is associated with improved second language vowel intelligibility. *Appl. Psycholinguist.* 33, 643–664. doi: 10.1017/S0142716411000518

Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Lang. Var. Change* 14, 261–290. doi: 10.1017/S0954394502143018

Davis, B. L., MacNeilage, P. F., and Matyear, C. L. (2002). Acquisition of serial complexity in speech production: a comparison of phonetic and phonological approaches to first word production. *Phonetica* 59, 75–107. doi: 10.1159/000066065

Davis, M., and Redford, M. A. (2019). The emergence of discrete perceptual-motor units in a production model that assumes holistic phonological representations. *Front. Psychol.* 10, 2121. doi: 10.3389/fpsyg.2019.02121

Diehl, R. L., and Lindblom, B. (2004). "Explaining the structure of feature and phoneme inventories: the role of auditory distinctiveness," in *Speech Processing in the Auditory System*, eds S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. R. Fay (New York, NY: Springer), 101–162.

Edwards, J., Beckman, M. E., and Munson, B. (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *J. Speech Lang. Hear. Res.* 47, 421–436. doi: 10.1044/1092-4388(2004/034)

Ferguson, C. A., and Farwell, C. B. (1975). Words and sounds in early language acquisition. *Language* 15, 419–439 doi: 10.2307/412864

Flege, J. E. (1995). "Second language speech learning: theory, findings, and problems," in *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*, ed W. Strange (Timonium, MD: York Press), 233–277.

Flege, J. E., and Bohn, O. S. (2021). "The revised speech learning model (SLM-r)," in *Second Language Speech Learning: Theoretical and Empirical Progress*, ed R. Wayland (Cambridge: Cambridge University Press), 3–83.

Flemming, E. (2004). "Contrast and perceptual distinctiveness," in *Phonetically Based Phonology*, eds B. Hayes, R. Kirchner, and D. Steriade (Cambridge: Cambridge University Press), 232–276.

Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing. *J. Phon.* 8, 113–133. doi: 10.1016/S0095-4470(19)31446-9

Gathercole, S. E. (2006). Nonword repetition and word learning: the nature of the relationship. *Appl. Psycholinguist.* 27, 513–543. doi: 10.1017/S0142716406060383

Guenther, F. H. (2016). *Neural Control of Speech*. Cambridge, MA: MIT Press.

Guion, S. G. (2003). The vowel systems of quichua-spanish bilinguals. *Phonetica* 60, 98–128. doi: 10.1159/000071449

Guion, S. G., Harada, T., and Clark, J. J. (2004). Early and late Spanish-English bilinguals' acquisition of English word stress patterns. *Biling Lang Cogn* 7, 207–226. doi: 10.1017/S1366728904001592

Holt, L. L., and Lotto, A. J. (2010). Speech perception as categorization. *Attent. Percept. Psychophys.* 72, 1218–1227. doi: 10.3758/APP.72.5.1218

Houde, J. F., and Nagarajan, S. S. (2011). Speech production as state feedback control. *Front. Hum. Neurosci.* 5, 82. doi: 10.3389/fnhum.2011.00082

Jaeger, J. J. (1997). How to say 'Grandma': The problem of developing phonological representations. *First Lang.* 17, 1–29. doi: 10.1177/014272379701705101

Johnson, K. (1997). The auditory/perceptual basis for speech segmentation. *Work. Papers Linguist.* 50, 101–113.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: the emergence of social identity and phonology. *J. Phon.* 34, 485–499. doi: 10.1016/j.wocn.2005.08.004

Levelt, W. J., Roelofs, A., and Meyer, A. S. (1999). A theory of lexical access in speech production. *Behav. Brain Sci.* 22, 1–38. doi: 10.1017/S0140525X99001776

Levelt, W. J. M. (1989). *Speaking: From Intention to Articulation*. Cambridge, MA: The MIT Press.

Lindblom, B. (1990). "Explaining phonetic variation: a sketch of the HandH theory," in *Speech Production and Speech Modelling*, eds W. J. Hardcastle and A. Marchal (Dordrecht: Springer), 403–439.

Luce, P. A., and Pisoni, D. B. (1998). Recognizing spoken words: the neighborhood activation model. *Ear. Hear.* 19, 1. doi: 10.1097/00003446-199802000-00001

Major, R. C. (1998). Interlanguage phonetics and phonology: an introduction. *Stud. Second Lang. Acquisit.* 20, 131–137. doi: 10.1017/S0272263198002010

Major, R. C. (2001). *Foreign Accent: The Ontogeny and Phylogeny of Second Language Phonology*. London: Routledge.

McCune, L., and Vihman, M. M. (2001). Early phonetic and lexical development: a productivity approach. *J. Speech Lang. Hear. Res.* 44, 670–684. doi: 10.1044/1092-4388(2001/054)

Mehl, M. R., Vazire, S., Ramírez-Esparza, N., Slatcher, R. B., and Pennebaker, J. W. (2007). Are women really more talkative than men? *Science* 317, 82–82. doi: 10.1126/science.1139940

Menn, L. (1983). "Development of articulatory, phonetic, and phonological capabilities," in *Language Production, Vol. 2*, ed B. Butterworth (London: Academic Press), 3–50.

Metsala, J. L. (1999). Young children's phonological awareness and nonword repetition as a function of vocabulary development. *J. Educ. Psychol.* 91, 3. doi: 10.1037/0022-0663.91.1.3

Munson, B., Kurtz, B. A., and Windsor, J. (2005). The influence of vocabulary size, phonotactic probability, and wordlikeness on nonword repetitions of children with and without specific language impairment. *J. Speech Lang. Hear. Res.* 48, 1033–1047. doi: 10.1044/1092-4388(2005/072)

Nagle, C. L. (2018). Examining the temporal structure of the perception-production link in second language acquisition: a longitudinal study. *Lang. Learn.* 68, 234–270. doi: 10.1111/lang.12275

Nagle, C. L., and Baese-Berk, M. M. (2022). Advancing the state of the art in L2 speech perception-production research: revisiting theoretical assumptions and methodological practices. *Stud. Second Lang. Acquisit.* 44, 580–605. doi: 10.1017/S0272263121000371

Nation, P., and Waring, R. (1997). Vocabulary size, text coverage and word lists. *Vocabulary* 14, 6–19.

Newell, K. M., Liu, Y. T., and Mayer-Kress, G. (2001). Time scales in motor learning and development. *Psychol. Rev.* 108, 57–82. doi: 10.1037/0033-295X.108.1.57

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *J. Phon.* 39, 132–142. doi: 10.1016/j.wocn.2010.12.007

Nittrouer, S., Studdert-Kennedy, M., and McGowan, R. S. (1989). The emergence of phonetic segments: evidence from the spectral structure of fricative-vowel syllables spoken by children and adults. *J. Speech Lang. Hear. Res.* 32, 120–132. doi: 10.1044/jshr.3201.120

Niziolek, C. A., Nagarajan, S. S., and Houde, J. F. (2013). What does motor efference copy represent? Evidence from speech production. *J. Neurosci.* 33, 16110–16116. doi: 10.1523/JNEUROSCI.2137-13.2013

Parrell, B., Lammert, A. C., Ciccarelli, G., and Quatieri, T. F. (2019). Current models of speech motor control: a control-theoretic overview of architectures and properties. *J. Acoust. Soc. Am.* 145, 1456–1481. doi: 10.1121/1.5092807

Pierrehumbert, J. (2001). "Exemplar dynamics: word frequency, lenition and contrast," in *Frequency and the Emergence of Linguistic Structure*, eds J. L. Bybee and P. J. Hopper (Amsterdam: John Benjamins Publishing Company), 137–157.

Pierrehumbert, J. (2002). Word-specific phonetics. *Lab. Phonol.* 7, 101–140. doi: 10.1515/9783110197105.101

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Lang. Speech* 46, 115–154. doi: 10.1177/00238309030460020501

Pisoni, D. B., Nusbaum, H. C., Luce, P. A., and Slowiaczek, L. M. (1985). Speech perception, word recognition and the structure of the lexicon. *Speech Commun.* 4, 75–95. doi: 10.1016/0167-6393(85)90037-8

Redford, M. A. (2015). Unifying speech and language in a developmentally sensitive model of production. *J. Phon.* 53, 141–152. doi: 10.1016/j.wocn.2015.06.006

Redford, M. A. (2019). Speech production from a developmental perspective. *J. Speech Lang. Hear. Res.* 62, 2946–2962. doi: 10.1044/2019_JSLHR-S-CSMC7-18-0130

Redford, M. A., and Oh, G. (2017). The representation and execution of articulatory timing in first and second language acquisition. *J. Phon.* 63, 127–138. doi: 10.1016/j.wocn.2017.01.004

Redford, M. A., and Oh, G. E. (2016). Children's abstraction and generalization of English lexical stress patterns. *J. Child Lang.* 43, 338–365. doi: 10.1017/S0305000915000215

Samuel, A. G., and Kraljic, T. (2009). Perceptual learning for speech. *Attent. Percept. Psychophys.* 71, 1207–1218. doi: 10.3758/APP.71.6.1207

Schmidt, R. A. (1975). A schema theory of discrete motor skill learning. *Psychol. Rev.* 82, 225. doi: 10.1037/h0076770

Schmidt, R. A. (2003). Motor schema theory after 27 years: reflections and implications for a new theory. *Res. Q. Exerc. Sport*. 74, 366–375. doi: 10.1080/02701367.2003.10609106

Shipley, K. G., and McAfee, J. G. (2019). *Assessment in Speech-Language Pathology: A Resource Manual*. San Diego, CA: Plural Publishing.

Smith, A., and Zelaznik, H. N. (2004). Development of functional synergies for speech motor coordination in childhood and adolescence. *Dev. Psychobiol*. 45, 22–33. doi: 10.1002/dev.20009

Smith, R., and Hawkins, S. (2012). Production and perception of speaker-specific phonetic detail at word boundaries. *J. Phon*. 40, 213–233. doi: 10.1016/j.wocn.2011.11.003

Velleman, S. (1998). *Making Phonology Functional: What do I do First*? Boston, MA: Butterworth-Heinemann.

Velleman, S. L., and Vihman, M. M. (2002). Whole-word phonology and templates: trap, bootstrap, or some of each? *Lang. Speech Hear. Serv. Sch*. 33, 9–23. doi: 10.1044/0161-1461(2002/002)

Verhagen, J., Van Stiphout, M., and Elma, B. L. O. M. (2022). Determinants of early lexical acquisition: effects of word-and child-level factors on Dutch children's acquisition of words. *J. Child Lang*. 49, 1–21. doi: 10.1017/S0305000921000635

Vihman, M. M. (2014). *Phonological Development: The First Two Years, 2nd Edn*. Malden, MA: Wiley-Blackwell.

Vihman, M. M., and Croft, W. (2007). Phonological development: toward a 'radical' templatic phonology. *Linguistics* 45, 683–725. doi: 10.1515/LING.2007.021

Wedel, A. B. (2006). Exemplar models, evolution and language change. *Linguist. Rev*. 23, 247–274. doi: 10.1515/TLR.2006.010

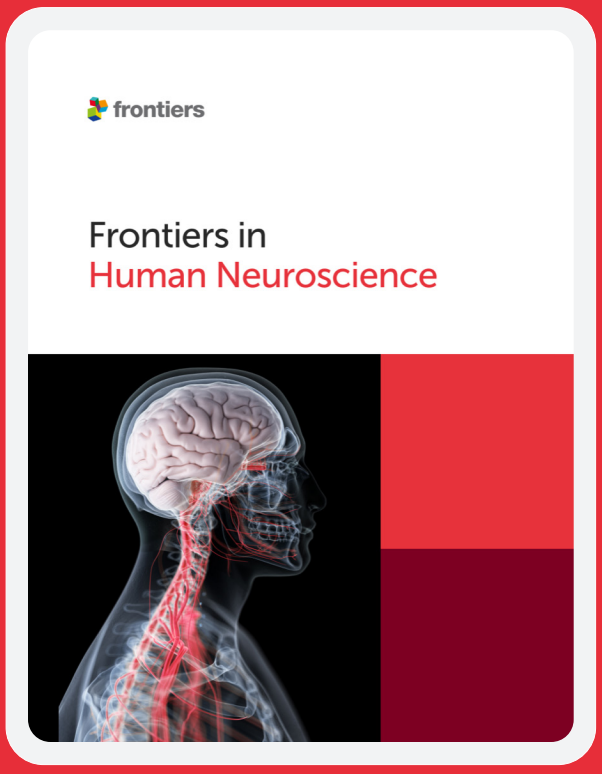
Frontiers in
Human Neuroscience

Bridges neuroscience and psychology to understand the human brain

The second most-cited journal in the field of psychology, that bridges research in psychology and neuroscience to advance our understanding of the human brain in both healthy and diseased states.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact