frontiers | Frontiers in Human Neuroscience

TYPE Perspective
PUBLISHED 08 January 2024
DOI 10.3389/fnhum.2023.1331129

OPEN ACCESS

EDITED BY
Karl David Neergaard,
Tecnologico de Monterrey, Spain

REVIEWED BY
Mikio Kubota,
Washington University in St. Louis,
United States

*CORRESPONDENCE
Josh Dorsi
✉ Jdorsi@pennstatehealth.psu.edu

RECEIVED 31 October 2023
ACCEPTED 11 December 2023
PUBLISHED 08 January 2024

CITATION
Dorsi J, Lacey S and Sathian K (2024) Multisensory and lexical information in speech perception.
Front. Hum. Neurosci. 17:1331129.
doi: 10.3389/fnhum.2023.1331129

Multisensory and lexical information in speech perception

Josh Dorsi[1]*, Simon Lacey[1,2,3] and K. Sathian[1,2,3]

[1]Department of Neurology, Penn State College of Medicine, Hershey, PA, United States, [2]Department of Neural and Behavioral Sciences, Penn State College of Medicine, Hershey, PA, United States, [3]Department of Psychology, Penn State Colleges of Medicine and Liberal Arts, Hershey, PA, United States

Both multisensory and lexical information are known to influence the perception of speech. However, an open question remains: is either source more fundamental to perceiving speech? In this perspective, we review the literature and argue that multisensory information plays a more fundamental role in speech perception than lexical information. Three sets of findings support this conclusion: first, reaction times and electroencephalographic signal latencies indicate that the effects of multisensory information on speech processing seem to occur earlier than the effects of lexical information. Second, non-auditory sensory input influences the perception of features that differentiate phonetic categories; thus, multisensory information determines what lexical information is ultimately processed. Finally, there is evidence that multisensory information helps form some lexical information as part of a phenomenon known as sound symbolism. These findings support a framework of speech perception that, while acknowledging the influential roles of both multisensory and lexical information, holds that multisensory information is more fundamental to the process.

KEYWORDS

multisensory, lexical, speech perception, sound symbolism, audiovisual

Introduction

Both lexical and multisensory information are known to support the perception of speech. For example, it is easier to identify speech comprising words rather than non-words (Hirsh et al., 1954) and speech conveyed through multiple rather than single sensory channels (i.e., audiovisual speech vs. auditory-only speech; Sumby and Pollack, 1954). In natural settings, speech typically involves real words (providing lexical information) spoken by talkers we can both see and hear (providing multisensory information). Thus, understanding how multisensory and lexical information are processed in relation to each other is important for a comprehensive understanding of the speech mechanism. Although work jointly testing lexical and multisensory information in speech perception is limited, some inferences can be made from such studies, as well as from work examining multisensory and lexical effects separately. Based on a review of this literature, we argue in this perspective that multisensory information plays a more fundamental role in speech perception than lexical information.

The McGurk effect is one of the most prominent demonstrations that lexical and multisensory information interact during speech perception. It refers to the finding that for some audiovisually incongruent speech, listeners hear the visual, not the auditory, signal (i.e., auditory /ba/ + visual /da/ is heard as /da/; McGurk and MacDonald, 1976). While originally studied at the level of syllables, the effect is influenced by the lexical composition of the incongruent speech, with visually consistent percepts being more common for auditory-non-word + visual-word (e.g., auditory /bɛf/ + visual /dɛf/) compared to auditory-word + visual-non-word (e.g., auditory /bænd/ + visual /dænd/; Brancazio, 2004). Such findings demonstrate the interactive effects of lexical and multisensory information in speech perception but do not allow us to distinguish whether (a) lexical information changes audiovisual integration or (b) lexical information influences some post-integration process.

Understanding the relative roles of lexical and multisensory processing in speech perception is vital for a fuller understanding of language comprehension. There is a lack of consensus on this issue, with some researchers favoring a more fundamental role of multisensory information (Rosenblum, 2008), whereas others assume the primacy of lexical information (Ostrand et al., 2016). Here, we review evidence that multisensory processing precedes lexical processing, affects the perception of pre-lexical speech units, and influences the formation of lexical representations. These lines of evidence support the contention that multisensory information is more fundamental to speech perception than lexical information.

What is the relative timing of lexical and multisensory processing?

Information processed earlier can influence information processed later. Both lexical and multisensory processing can occur rapidly (e.g., Baart and Samuel, 2015). This section examines studies that include multisensory and lexical information within the same paradigm to evaluate whether one might be handled earlier.

Baart and Samuel (2015) measured event-related potentials (ERPs) in response to three-syllable words and pseudowords presented auditorily, visually, or audiovisually. ERPs relative to the onset of the third syllable, where the word-determining information occurred, showed significant effects of multisensory information at earlier time windows (0–50 ms) than for lexical information (100–150 ms). An analogous second experiment also found that multisensory effects (50–100 ms) were earlier than lexical effects (150–200 ms). The multisensory effects do not seem to be driven by the early availability of visual information, as they were associated with frontal electrodes. These data have notably high subject-level variability (see Baart, 2016); however, they resemble results obtained in another multisensory-lexical ERP study (Basirat et al., 2018) investigating the word repetition effect (the finding of facilitated processing for repeated words). Basirat et al. (2018) measured ERPs starting at the onset of initially presented and repeated auditory-only and audiovisual words. They found that the earliest effect of lexical information (word repetition) was in the 170–280 ms window; in contrast, modality had a main effect in the 0–80 ms window, suggesting that multisensory processing preceded lexical processing (Basirat et al., 2018).

Ostrand et al. (2016) investigated the relative timing of lexical and multisensory processing by testing whether semantic priming, a lexical process, is sensitive to multisensory integration. Auditory-only target words (e.g., /wɜrm/) were categorized faster when they followed audiovisual-incongruent prime words with an auditory component semantically related to the target word (i.e., auditory /bet/ + visual /det/); semantic priming was consistent with the auditory channel of incongruent primes. These incongruent primes could be integrated such that participants "heard" either the visual or the auditory word (Brancazio, 2004). Dorsi et al. (2023) replicated the semantic priming paradigm of Ostrand et al. (2016) and included a McGurk effect assessment for the primes. This study found that priming to the auditory words corresponded to how likely the incongruent stimulus was to be heard as the auditory word. Likewise, primes frequently heard as the visual word were associated with priming consistent with the visual word (Dorsi et al., 2023). This suggests that multisensory integration precedes lexical processing because semantic priming appears contingent on the multisensory interactions determining the incongruent word's perception. However, alternative explanations also exist, such as the possibility of a lexical contribution to the perception of the incongruent word. Using ERPs in a semantic priming paradigm might be helpful to confirm that multisensory perception, indexed by the McGurk effect, precedes the availability of lexical information.

Baart and Samuel (2015) and Basirat et al. (2018) measured multisensory and lexical effects in a time-sensitive way. Both studies were interested in the P2, an ERP whose latency and amplitude are modulated approximately 200 ms after relevant lexical or multisensory information appears in the speech signal (see Baart and Samuel, 2015 for a discussion). The P2 is assumed to be associated with early lexical processes (Basirat et al., 2018), and indeed, both studies found lexical effects in the P2 time window. However, both studies also converge in showing effects of multisensory information in earlier windows (e.g., 0–80 ms) than those for lexical information. The results of Dorsi et al. (2023) are consistent with this conclusion since lexical processing apparently depends on the outcome of multisensory integration, although this question should be more thoroughly tested to exclude alternative possibilities.

Does multisensory information influence what lexical information is processed?

While lexical processing might begin with pre-lexical speech units, there is evidence that multisensory information shapes the perception of even the most basic pre-lexical information. For example, visual speech affects the perception of pre-phonetic auditory information such as voice-onset-time (VOT), the time from acoustic onset to the sudden increase in acoustic energy (Green and Miller, 1985) that distinguishes phonemes such as /b/ from /p/ (/b/ = shorter VOT). Despite its link to the acoustic signal, VOT is perceived as being shorter when accompanied by fast vs. slow visual speech (Green and Miller, 1985), suggesting that multisensory interactions influence a pre-lexical feature that presumably is involved with initial lexical processing.

Multisensory input also influences the perception of the speech signal's more basic acoustic parameters (e.g., Plass et al., 2020). For example, the visible shape of the mouth opening improves the perception of degraded auditory speech through its influence on the perception of spectro-temporal properties of the acoustic signal (e.g., formants; Plass et al., 2020). Likewise, the correlation between visible changes in the area of the mouth opening and the auditory speech envelope corresponds to audiovisual improvement of speech-in-noise perception (Grant and Seitz, 2000). Activity in auditory cortical areas is known to correlate with the auditory speech envelope (e.g., Abrams et al., 2008); the addition of visual speech (Crosse et al., 2015) or even vibrotactile speech (Riecke et al., 2019) improves this cortical tracking. Likewise, audiovisual speech influences auditory cortical activity (Okada et al., 2013). Moreover, visual speech influences auditory speech-associated activity in the brainstem (Musacchia et al., 2006) and the cochlea (Namasivayam et al., 2015). While these latter effects may result from feedback from cortical locations, they demonstrate how multisensory input influences the neural fate of even the most basic auditory information. Thus, the perception and neural handling of basic speech information, even with lexical feedback (e.g., Marian et al., 2018; Li et al., 2020), is not free from multisensory influences.

Is lexical information formed independent of multisensory processing?

It takes months of experience before lexical information becomes useful to listeners (e.g., Jusczyk et al., 1994). Multisensory effects on speech perception likely occur while lexical representations are being formed in childhood (Walton and Bower, 1993). A set of findings related to sound symbolism is consistent with this notion. Sound symbolism is the association between the sound of a word and its meaning. While sound symbolism is still poorly understood, we review evidence here suggesting that it may be inherent to language processing, involve multisensory processing, and support language acquisition. These points suggest the intriguing possibility that multisensory information is involved in forming some lexical information.

Sound symbolism may be inherent to language. The sound-symbolic associations of pseudowords (e.g., /buba/ sounds rounded, /kiki/ sounds pointed: Ramachandran and Hubbard, 2001) generalize to phonetic-to-meaning correspondences in real words (Sidhu et al., 2021). This correspondence is common across the world's languages (Blasi et al., 2016). Sound symbolism also seems to be related to the neural basis of language. In a recent functional magnetic resonance imaging (fMRI) study, a multivoxel pattern analysis (MVPA) indicated that activity in language-associated areas such as the left supramarginal gyrus and Broca's area in the left inferior frontal gyrus could distinguish rounded/pointed stimuli more accurately for sound symbolically matched pseudoword-shape pairs (e.g., /molo/ + rounded shape) than for mismatched pairs (e.g., /molo/ + pointed shape) (Barany et al., 2023).

There is also evidence that sound symbolism may involve multisensory processing. For example, sound symbolically matched audiovisual pseudoword-shape pairs produce more activation in auditory areas than unmatched pairs (Barany et al., 2023). In visual areas, the activation difference between mismatched and matched pseudoword-shape pairs correlates with behavioral measures of implicit pseudoword-shape associations (Peiffer-Smadja and Cohen, 2019). Likewise, MVPA indicates that activity in early visual areas more accurately distinguishes rounded and pointed stimuli that are part of sound symbolically matched, as opposed to mismatched, pseudoword-shape pairs (Barany et al., 2023). While the exact nature of the neural computations underlying sound symbolism are still not understood, these findings indicate that sound symbolism may, at least partly, involve multisensory processing.

Moreover, sound symbolism may support language acquisition. The sound symbolism bootstrapping hypothesis proposes that sound-symbolic associations facilitate initial word learning (Imai and Kita, 2014). Indeed, words rated as sounding like their meaning are overrepresented in the earliest words learned by children (Perry et al., 2017). Infants are sensitive to sound-symbolic correspondences; four-month-olds prefer sound symbolically matched to mismatched speech-shape pairs (Ozturk et al., 2013), and 14-month-olds are better at learning sound symbolically matched than mismatched labels for novel shapes (Imai et al., 2015). Adults are better at learning sound symbolically congruent than incongruent pseudoword-shape mappings (Revill et al., 2018), are more accurate in learning the correct than incorrect meanings of sound-symbolic foreign language words (Lockwood et al., 2016), and are better than chance at choosing the correct meaning of sound-symbolic foreign word pairs (Revill et al., 2014). The role of multisensory interactions in sound symbolism suggests that, at least for some words, multisensory processes influence the formation of lexical representations.

Discussion and conclusion

In this perspective, we reviewed literature suggesting that multisensory information is more fundamental to speech perception than lexical information. Three sets of observations support our argument: there may be earlier processing of multisensory information; the basic units of lexical representations are sensitive to multisensory information; and, through sound symbolism, some lexical representations may be formed with multisensory inputs. Each of these ideas requires further testing. Such testing could include methods with high temporal resolution to simultaneously measure the timing of multisensory processes in relation to the recovery of pre-lexical information (e.g., phonetic features or spectro-temporal acoustic parameters) and subsequent lexical processes. Experiments that more directly test the role of multisensory interactions in sound symbolism and examine the lexical effects of sound symbolism will also be useful. While work remains to be done, we conclude that multisensory information is likely more fundamental to speech perception than lexical information. There are clinical implications of this view. For example, while cochlear implant recipients demonstrate reduced multisensory integration, multisensory information reliably supports word recovery in this population (Stevenson et al., 2017). Recent work has also found word perception improvements when cochlear implant recipients wore a device that transduced

the acoustic speech signal into vibratory stimulation in real time (Fletcher et al., 2019). Similarly, people with aphasia show improved lexical processing in challenging listening conditions when speech is presented in a multisensory context (Krason et al., 2023). These observations demonstrate the importance of considering the relative impacts of multisensory and lexical information on speech processing, as the present perspective has discussed.

Data availability statement

The original contributions presented in this study are included in this article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

JD: Conceptualization, Writing – original draft, Writing – review & editing. SL: Writing – review & editing. KS: Writing – review & editing.

Funding

The authors declare financial support was received for the research, authorship, and/or publication of this article. Institutional start-up funds awarded to KS from Penn State College of Medicine supported this work.

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

References

Abrams, D. A., Nicol, T., Zecker, S., and Kraus, N. (2008). Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech. *J. Neurosci.* 28, 3958–3965. doi: 10.1523/JNEUROSCI.0187-08.2008

Baart, M. (2016). Quantifying lip-read-induced suppression and facilitation of the auditory N1 and P2 reveals peak enhancements and delays. *Psychophysiology* 53, 1295–1306. doi: 10.1111/psyp.12683

Baart, M., and Samuel, A. G. (2015). Turning a blind eye to the lexicon: ERPs show no cross-talk between lip-read and lexical context during speech sound processing. *J. Mem. Lang.* 85, 42–59. doi: 10.1016/j.jml.2015.06.00

Barany, D. A., Lacey, S., Matthews, K. L., Nygaard, L. C., and Sathian, K. (2023). Neural basis of sound-symbolic pseudoword-shape correspondences. *Neuropsychologia* 188:108657. doi: 10.1016/j.neuropsychologia.2023.108657

Basirat, A., Brunellière, A., and Hartsuiker, R. (2018). The role of audiovisual speech in the early stages of lexical processing as revealed by the ERP word repetition effect. *Lang. Learn.* 68, 80–101. doi: 10.1111/lang.12265

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., and Christiansen, M. H. (2016). Sound-meaning association biases evidenced across thousands of languages. *Proc. Natl Acad. Sci. U.S.A.* 113, 10818–10823. doi: 10.1073/pnas.1605782113

Brancazio, L. (2004). Lexical influences in audiovisual speech perception. *J. Exp. Psychol. Hum. Percept. Perform.* 30, 445–463. doi: 10.1037/0096-1523.30.3.445

Crosse, M. J., Butler, J. S., and Lalor, E. C. (2015). Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *J. Neurosci.* 35, 14195–14204. doi: 10.1523/JNEUROSCI.1829-15.2015

Dorsi, J., Ostrand, R., and Rosenblum, L. D. (2023). Semantic priming from McGurk words: Priming depends on perception. *Attent. Percept. Psychophys.* 85, 1219–1237. doi: 10.3758/s13414-023-02689-2

Fletcher, M. D., Hadeedi, A., Goehring, T., and Mills, S. R. (2019). Electro-haptic enhancement of speech-in-noise performance in cochlear implant users. *Sci. Rep.* 9:11428. doi: 10.1038/s41598-019-47718-z

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. doi: 10.1121/1.1288668

Green, K. P., and Miller, J. L. (1985). On the role of visual rate information in phonetic perception. *Percept. Psychophys.* 38, 269–276.

Hirsh, I. J., Reynolds, E. G., and Joseph, M. (1954). Intelligibility of different speech materials. *J. Acoust. Soc. Am.* 26, 530–538. doi: 10.1121/1.1907370

Imai, M., and Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philos. Trans. R. Soc. B Biol. Sci.* 369, 20130298. doi: 10.1098/rstb.2013.0298

Imai, M., Miyazaki, M., Yeung, H. H., Hidaka, S., Kantartzis, K., Okada, H., et al. (2015). Sound symbolism facilitates word learning in 14-month-olds. *PLoS One* 10:e0116494. doi: 10.1371/journal.pone.0116494

Jusczyk, P. W., Luce, P. A., and Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *J. Mem. Lang.* 33, 630–645. doi: 10.1006/jmla.1994.1030

Krason, A., Vigliocco, G., Mailend, M. L., Stoll, H., Varley, R., and Buxbaum, L. J. (2023). Benefit of visual speech information for word comprehension in post-stroke aphasia. *Cortex* 165, 86–100. doi: 10.1016/j.cortex.2023.04.011

Li, X., Zhang, X., and Gong, Q. (2020). Evidence of both brainstem and auditory cortex involvement in categorical perception for Chinese lexical tones. *Neuroreport* 31, 359–364. doi: 10.1097/WNR.0000000000001414

Lockwood, G., Hagoort, P., and Dingemanse, M. (2016). How iconicity helps people learn new words: Neural correlates and individual differences in sound-symbolic bootstrapping. *Collabra* 2:7.

Marian, V., Lam, T. Q., Hayakawa, S., and Dhar, S. (2018). Top-down cognitive and linguistic influences on the suppression of spontaneous otoacoustic emissions. *Front. Neurosci.* 12:378. doi: 10.3389/fnins.2018.00378

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748.

Musacchia, G., Sams, M., Nicol, T., and Kraus, N. (2006). Seeing speech affects acoustic information processing in the human brainstem. *Exp. Brain Res.* 168, 1–10. doi: 10.1007/s00221-005-0071-5

Namasivayam, A. K., Yiu, W., and Wong, S. (2015). Visual speech gestures modulate efferent auditory system. *J. Integr. Neurosci.* 14, 73–83. doi: 10.1142/S0219635215500016

Okada, K., Venezia, J. H., Matchin, W., Saberi, K., and Hickok, G. (2013). An fMRI study of audiovisual speech perception reveals multisensory interactions in auditory cortex. *PLoS One* 8:e68959. doi: 10.1371/journal.pone.0068959

Ostrand, R., Blumstein, S. E., Ferreira, V. S., and Morgan, J. L. (2016). What you see isn't always what you get: Auditory word signals trump consciously perceived words in lexical access. *Cognition* 151, 96–107. doi: 10.1016/j.cognition.2016.02.019

Ozturk, O., Krehm, M., and Vouloumanos, A. (2013). Sound symbolism in infancy: Evidence for sound-shape cross-modal correspondences in 4-month-olds. *J. Exp. Child Psychol.* 114, 173–186. doi: 10.1016/j.jecp.2012.05.004

Peiffer-Smadja, N., and Cohen, L. (2019). The cerebral bases of the bouba-kiki effect. *Neuroimage* 186, 679–689. doi: 10.1016/j.neuroimage.2018.11.033

Perry, L. K., Perlman, M., Winter, B., Massaro, D. W., and Lupyan, G. (2017). Iconicity in the speech of children and adults. *Dev. Sci.* 21:e12572. doi: 10.1111/desc.12572

Plass, J., Brang, D., Suzuki, S., and Grabowecky, M. (2020). Vision perceptually restores auditory spectral dynamics in speech. *Proc. Natl Acad. Sci. U.S.A.* 117, 16920–16927. doi: 10.1073/pnas.2002887117

Ramachandran, V. S., and Hubbard, E. M. (2001). Synaesthesia - A window into perception, thought and language. *J. Conscious. Stud.* 8, 3–34.

Revill, K. P., Namy, L. L., and Nygaard, L. C. (2018). Eye movements reveal persistent sensitivity to sound symbolism during word learning. *J. Exp. Psychol. Learn. Mem. Cogn.* 44, 680–698. doi: 10.1037/xlm0000476

Revill, K. P., Namy, L. L., DeFife, L. C., and Nygaard, L. C. (2014). Cross-linguistic sound symbolism and crossmodal correspondence: Evidence from fMRI and DTI. *Brain Lang.* 128, 18–24. doi: 10.1016/j.bandl.2013.11.002

Riecke, L., Snipes, S., van Bree, S., Kaas, A., and Hausfeld, L. (2019). Audio-tactile enhancement of cortical speech-envelope tracking. *Neuroimage* 202:116134. doi: 10.1016/j.neuroimage.2019.116134

Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Curr. Direct. Psychol. Sci.* 17, 405–409. doi: 10.1111/j.1467-8721.2008.00615.x

Sidhu, D. M., Westbury, C., Hollis, G., and Pexman, P. M. (2021). Sound symbolism shapes the English language: The maluma/takete effect in English nouns. *Psychon. Bull. Rev.* 28, 1390–1398. doi: 10.3758/s13423-021-01883-3

Stevenson, R. A., Sheffield, S. W., Butera, I. M., Gifford, R. H., and Wallace, M. T. (2017). Multisensory integration in cochlear implant recipients. *Ear Hear.* 38, 521–538. doi: 10.1097/AUD.0000000000000435

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212–215.

Walton, E., and Bower, T. (1993). Representation in infants of speech. *Infant Behav. Dev.* 16, 233–243.