frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 27 March 2024
DOI 10.3389/fcomm.2024.1287363

Check for updates

OPEN ACCESS

EDITED BY
Luca Campanelli,
University of Alabama, United States

REVIEWED BY
Simon Roessig,
University of York, United Kingdom
Florence Baills,
Universitat de Lleida, Spain

*CORRESPONDENCE
Miguel Jiménez-Bravo
✉ miguel.jimenez.bravo@ucm.es

RECEIVED 01 September 2023
ACCEPTED 04 March 2024
PUBLISHED 27 March 2024

CITATION
Jiménez-Bravo M and Marrero-Aguiar V (2024) Multimodal prosody: gestures and speech in the perception of prominence in Spanish. *Front. Commun.* 9:1287363.
doi: 10.3389/fcomm.2024.1287363

Multimodal prosody: gestures and speech in the perception of prominence in Spanish

Miguel Jiménez-Bravo[1]* and Victoria Marrero-Aguiar[2]

[1]Department of Linguistics and Eastern Studies, Complutense University of Madrid, Madrid, Spain,
[2]Department of General Linguistics, National University of Distance Education (UNED), Madrid, Spain

Multimodal communication cannot be properly understood without analyzing the natural interweaving of speech and gestures as it typically occurs in everyday spoken language, thus moving beyond studies that elicit gestures in the laboratory, most of which are also conducted for English. Therefore, this study addresses the effect of both visual and acoustic cues in the perception of prominence in Castilian Spanish using spontaneous speech from a TV talent-show. Four between-subjects experiments in each modality—audio-only and audiovisual—were conducted online, each including a different combination of manipulated cues: Exp1 (flat F0), Exp2 (flat intensity), and Exp3 (flat F0 + flat intensity), while all cues remained intact in the control experiment Exp0. Additionally, the capability of the different gesture phases to convey prominence was analyzed in their interaction with the acoustic cues. The results showed that, when prominence was perceived in manipulated stimuli, the effect of the visual information depended on the acoustic cues available in the signal and was also reduced when compared to non-manipulated stimuli, pointing to a strong integration of both modalities in prominence perception. In non-manipulated stimuli, all acoustic cues—except for spectral balance—played a role in the perception of prominence; however, when the visual information was added, it reduced the perceptual effect of the acoustic cues, and the main role played by duration was combined with that of the stroke phase of gestures.

KEYWORDS

audiovisual prosody, multimodality, speech perception, acoustic cues, gesture, prominence, Spanish

1 Introduction

Gestures and speech interact in everyday spoken language, and their combination not only helps to express ideas and facilitate comprehension (e.g., Novack and Goldin-Meadow, 2017), but also serves a wide array of pragmatic functions (e.g., Swerts and Krahmer, 2005; Krahmer and Swerts, 2007; Prieto et al., 2011; Kushch and Prieto Vives, 2016). Gestures have traditionally been defined as spontaneous visible body movements—mostly performed with hands, face, and head—accompanying speech and are also referred to as gesticulation (Kendon, 2004). They contribute to communication by adding non-discrete nuances and cueing prominence, much as it has been observed for speech (e.g., Munhall et al., 2004; Foxton et al., 2010). For Kendon (1972), the essential phase of the gesture is the stroke, which corresponds to its most distinct effort and spans over a small interval of time. The apex of gestures, however, which is characterized by the effort peak found within strokes, has been suggested to be the smallest gesture phase (Loehr, 2004). Typically in hand gestures, strokes may be preceded by a preparatory phase and are often followed by a retraction phase. In the optional preparation, the hand is brought to the point where the stroke is initiated, and in the retraction—also called recovery—, the hand is brought back to a resting position, either to its starting point or to any other point from where the next gesture can eventually be initiated. Strokes may optionally also be preceded by a brief hold before they are initiated or can also be followed by a hold in which the hand is maintained in the position at which it arrived (Kita, 1993).

The strong connection between the production of gestures and the production of speech has been corroborated by a large number of studies (e.g., McNeill, 1992; Cavé et al., 1996; McClave, 1998; Krahmer et al., 2002a; Jannedy and Mendoza-Denton, 2005; Prieto et al., 2011; Loehr, 2012). The role played by gestures in combination with—or as part of—language has been referred to as *multimodality*, a term that may also be used differently in the relation of gesture to speech (Sandler, 2022) or that can even be applied in a larger semiotic sense (e.g., Stöckl and Pflaeging, 2022; Cheema et al., 2023).

In this sense, one line of research has focused on the perceptual effects of the visual component of speech. For example, the so-called McGurk effect initially established that visual information in the form of lip movements affects speech perception (McGurk and MacDonald, 1976); later, not only lips, but also the rest of the face was reported to affect the perception of speech (e.g., Pelachaud et al., 1996). Another line of research has studied how the visual cues of prominence systematically increase the production and alter the perception of verbal prominence (e.g., Granström et al., 1999; House et al., 2001; Krahmer and Swerts, 2007; Swerts and Krahmer, 2008; Dohen and Lœvenbruck, 2009; Scarborough et al., 2009; Al Moubayed et al., 2010; Prieto et al., 2011; Kim et al., 2014; Jiménez-Bravo and Marrero-Aguiar, 2020).

The term *prominence* often appears as a synonym of a great variety of other terms such as emphasis, lexical stress, nuclear accent, prosodic focus, pitch accent, intensity peak, etc., depending on the perspective and the research framework under which it is invoked. Terken and Hermes (2000, p. 89) generically state that "a *linguistic entity* is *prosodically* prominent when *it stands out* from *its environment* by virtue of its *prosodic characteristics*", where the place-holders in italics can be replaced with more precise terms depending on the perspective adopted by the researchers (Wagner et al., 2015). Within the phonetic perspective used in this study, prominence is equated with acoustic perceptual salience, so henceforth a word is said to be prosodically prominent when it is acoustically salient within a sentence—typically serving pragmatic functions such as focus and information status marking—by virtue of the interplay of pitch, loudness, and length, which are realized as fundamental frequency (F0), intensity, and duration in the acoustic information of the speech signal produced by the speaker.

However, it is not known how these different cues of prominence relate to one another and what the relative perceptual weight of each one is (e.g., Silipo and Greenberg, 2000; Ortega-Llebaria and Prieto, 2011). Each language gives different roles to each of these prominence cues, so that their relative contribution is language-specific (Leemann et al., 2016). Similarly, the phonology of word and phrasal prominence, rendered by the interplay of these cues, is dictated by the respective grammar of each language (e.g., Vogel et al., 2016). In the case of Spanish, lexical prominence can be produced by a flat pitch contour, together with longer duration and stronger intensity in unaccented stressed syllables; while phrasal prominence—i.e., accented stressed syllables—is cued by longer duration, higher F0, larger F0 excursions, and an increased overall intensity (Ortega-Llebaria, 2006). Furthermore, the preponderant role of duration as a correlate both of lexical stress and phrasal stress in Spanish has also been attested (Vogel et al., 2016). As for the interaction between speech and gestures, there exists a strong temporal coordination between stressed syllables and strokes (McNeill, 1992), and more specifically between F0 peaks and the apex of gestures, i.e., the effort peak within strokes (Kendon, 1972; Loehr, 2004, 2012; Renwick et al., 2004; Jannedy and Mendoza-Denton, 2005; Shattuck-Hufnagel et al., 2007; Esteve-Gibert and Prieto, 2013; Rohrer et al., 2023).

Many studies on the perception of speech prominence have made use of the Rapid Prosody Transcription method, which involves the identification of prosodic prominence during an experimental task by a group naïve listeners (e.g., Cole et al., 2010a,b; Smith and Edmunds, 2013; Luchkina et al., 2015; Hualde et al., 2016). Then, inter-rater agreement is usually computed, and the prominence marks are pooled together over transcribers to obtain a population-wise, probabilistic measure of the prosodic status of each word (e.g., Streefkerk et al., 1997; Swerts, 1997; Cole et al., 2010a, 2014).

Furthermore, prominence perception seems to depend not only on the characteristics of the signal (bottom-up processing), but also on the characteristics and abilities of the receiver (top-down processing), as suggested by the degree of inter-individual variability in age and gender observed in previous experiments (e.g., Strand et al., 2014; Bishop et al., 2020). In this sense, a possible difference between men and women in the context of challenging stimuli has been reported for the audiovisual perception of speech (see Alm and Behne 2015, for a summary; see Jaeger et al., 1998 for neuroanatomical support). Another source of inter-individual differences in the detection of prominence are musical abilities, as is the case for instrumentalists having had continuous musical training—often from an early age—, who showed an alteration of their perception of prosody (e.g., Hutka et al., 2015) due to a transfer of abilities from music to linguistic prosody for perceiving pitch and rhythmic variations (e.g., Thompson et al., 2004). Particularly important is the effect of musical training on the ability to perceive small variations in F0 (e.g., Besson et al., 2007; Patel and Iversen, 2007; Sturgeon et al., 2015). However, not all results have yielded differences between individuals with and without musical training. Niebuhr (2009, p. 107), for example, found his results to "hold in similar ways for both musical and non-musical subjects"; and Madsen's et al. (2017) study points in the same direction right from the title: "Musicians do not benefit from differences in fundamental frequency when listening to speech in competing speech backgrounds".

From a methodological point of view, research on how the information conveyed in the visual modality interacts with prosody has made use of lip-synchronized animated agents (e.g., Granström et al., 1999; House et al., 2001; Krahmer et al., 2002a,b; Al Moubayed and Beskow, 2009; Prieto et al., 2011) or gestures elicited with controlled speech stimuli in experimental settings (e.g., Krahmer and Swerts, 2007; Dohen and Lœvenbruck, 2009; Foxton et al., 2010; Muñoz-Coego et al., 2022) (see Supplementary Table 1).

Nonetheless, not only the external validity of both methods could be improved, but also their potential to study the complex interaction of visual and verbal prosody is to some extent reduced, especially since most studies have focused on the interaction of just one gesture articulator—either hands, head, or eyebrows—with one acoustic correlate of prominence, which is very far from the natural interweaving of gestures performed with hands, head, and eyebrows typically found in everyday spoken language.

Alternatively, rather than relying solely on animated agents and multimodal stimuli elicited in experimental settings, researchers have also used spontaneous speech and spontaneously elicited gestures in order to gain new insights into the interaction between visual and verbal prosody (Swerts and Krahmer, 2010; Ambrazaitis and House, 2017; Jiménez-Bravo and Marrero-Aguiar, 2020; Rohrer et al., 2023). One of the first studies using spontaneous speech analyzed the effects of facial expressions on speech production using samples obtained from two male and two female Dutch TV newsreaders (Swerts and Krahmer, 2010). The recordings used as stimuli in this study contained sentences ranging between 4 and 12 s, which were presented to a group of 35 participants in the auditory modality for binary prominence marking (prominent vs. non-prominent). The results showed that words having a "strong accent" mostly occurred with an accompanying eyebrow movement; however, the mere presence of an eyebrow movement did not imply the presence of a strong accent, more precisely only 47 out of 303 eyebrow movements corresponded to a strong accent. The distribution of head movements followed a similar pattern, and strong accents were especially marked by combinations of eyebrow and head movements. Conversely, single eyebrow or head movements hardly coincided with strong accents.

The audiovisual realization of multimodal prominence was further explored by Ambrazaitis and House (2017) using stimuli from several TV newsreaders. In a study on the the connection between beat gestures—performed with head and/or eyebrows—and pitch accents signaling focal constituents, the researchers found patterns of gesticulation that depended on the news topic as well as on the speaker. Interestingly, the connection of focal accents and head beats—but not eyebrow raises—showed a distinct distribution, so that focal accents were preferably used by newsreaders in the first half of the text, while head beats and the combination of focal accents and head beats occurred during the second half of the text. This was explained on the basis of information structure, since the initial part of a text often presents the theme, defining a common ground, while the second part usually corresponds to the rheme. Thus, head beats seemed to highlight the most important information in a piece of news once it had already been presented in the first half of the text.

Similarly, the connection between prosody and gestures using multimodal stimuli was also addressed by Rohrer et al. (2023), who analyzed a corpus of academic discourses in English (TED Talks). The researchers observed that the patterns of temporal alignment of strokes and apexes within the boundaries of pitch accented syllables was different for both gesture phases, with strokes spanning over pitch accented syllables more frequently than apexes. In addition, they also observed that strokes stably aligned with phrase-initial prenuclear accents over nuclear accents regardless of their relative prominence within the phrase.

Finally, another study on multimodal prominence perception using spontaneous speech materials obtained from a Spanish TV talent show analyzed the role played by gestures—performed with hands, head, and eyebrows—as well as the three acoustic correlates of prominence: F0, intensity, and duration (Jiménez-Bravo and Marrero-Aguiar, 2020). In this study, words from a 30-sentence corpus were rated by listeners in a binary marking task (prominent vs. non-prominent) in two modalities (audio-only and audiovisual) and in three different experimental conditions that involved the reduction of the prominence-lending properties of either F0 or intensity, or of both of them simultaneously. The results showed that phrasal prominence was perceived even in the absence of most of the auditory cues—i.e., when only duration was available to listeners—, but the role of F0 and, particularly, intensity were less determinant. Also, the stroke phase of gestures, rather than the apex, was observed to increase the probability of words to be perceived as prominent. Finally, the analysis confirmed the prevalence of gestures performed simultaneously with more than one articulator, as pointed by Ambrazaitis and House (2017), so gestures were mostly produced by combining different body parts, especially hands and head. Such results speak in favor of using spontaneous speech material—with a caveat on the gestural stiffness typically displayed by TV newsreaders—, since spontaneous speech better reflects the real interplay of gestures and speech in everyday spoken language.

Following this line of reasoning, and differently from previous methodologies, this study relies on spontaneous multimodal speech materials. As already mentioned, previous studies on prominence perception with animated agents have limited themselves to reproduce eyebrow and head movements but have excluded hand movements (e.g., House et al., 2001; Prieto et al., 2011), while other studies have employed experimental stimuli obtained in a very controlled environment in the laboratory (e.g., Krahmer and Swerts, 2007; Dohen and Lœvenbruck, 2009). These methods have made it difficult to address how the different acoustic correlates of prominence relate to one another and also to gestures, and have a low degree of ecological validity.

Thus, in the present study we used spontaneous natural stimuli, which allow us to maintain a certain degree of naturalness in the task despite the manipulation of the acoustic signal, to conduct online four perceptual experiments. In this sense, this study is a follow-up of the study conducted by Jiménez-Bravo and Marrero-Aguiar (2020), applying the same methodology described there, but from which it differs in two aspects. Firstly, four prototypical sentences were chosen based on the prominence marks showing the highest agreement from the initial 30-sentence corpus analyzed in that previous study. Secondly, we also employ here a between-subjects design allowing to conduct independent analyses for both the auditory and the audiovisual cues of prominence for each experiment. This is so because the incomplete within-subjects design employed previously did not allow to compare the same stimuli across different experimental conditions, since participants could only mark a given manipulated stimulus in only one of the experiments and modalities.

The objectives of this study are twofold. Firstly, it aims at analyzing the contribution of the different phases of gestures—whether performed with hands, head, or eyebrows simultaneously or in combination—to the perception of prominence by offering fine-grained details of the interaction between auditory and visual cues in language processing. Secondly, we wish to gain insight into the relative weight and the interactions of the three acoustic cues of phrasal prominence in Castilian Spanish—namely F0, intensity and duration (to which we also added spectral balance)—by means of an experimental design that allows to conduct separate analyses for different combinations of these variables, whose hierarchy and perceptual importance have been the subject of a longstanding debate in Hispanic linguistics (e.g., Contreras, 1964; Navarro Tomás, 1964; Quilis, 1971; Solé, 1984; Enríquez et al., 1989; Llisterri et al., 2003; Ortega-Llebaria and Prieto, 2011).

Following the literature, our first hypothesis is that visual information, when available, in the form of gestures will be used by listeners as a cue of prominence together with the acoustic changes in F0, intensity, or duration. Our second hypothesis is that, when lacking any cues in the visual modality, and with duration as the only cue available to them, listeners will still be able to detect prominence, as suggested by previous results (Jiménez-Bravo and Marrero-Aguiar, 2020). The third hypothesis is that the apex phase of gestures will drive the perception of visual prominence, due to its synchronization with the prosody of the speech signal (Kendon, 1972; Jannedy and Mendoza-Denton, 2005; Loehr, 2012; Esteve-Gibert and Prieto, 2013), even if alternative findings are suggestive of a possible stronger relevance of strokes in perception (Jiménez-Bravo and Marrero-Aguiar, 2020). Finally, considering evidence previously mentioned, we wished to control for both the gender and the level of musical training of participants in their realization of the experimental task.

This paper is organized as follows. Initially, the second section describes the multimodal stimuli used in the four perception experiments as well as the methodology used for the manipulation of the acoustic signal and the annotation of gestures. Later, the third section is divided into two parts: in the first part we assess the marks of prominence in the four experiments and compute inter-rater agreement and the prominence score pooled over participants. In the second part, a set of generalized linear mixed models are estimated in a global analysis for the four experiments combined, from which the model that best account for the data is chosen, to compare the results with those initially obtained for a 30-sentence corpus used in Jiménez-Bravo and Marrero-Aguiar (2020). Later, generalized linear mixed models are fitted separately for each experiment, which enables to conduct independent analyses for both the auditory and the audiovisual cues of prominence. Finally, the fourth and the fifth sections assess the implications of the results and offer an interpretation in the light of previous studies.

2 Materials and methods

2.1 Stimuli

All audiovisual clips used as multimodal stimuli showed a speaker engaged in a spontaneous conversation uttering a sentence without being interrupted while performing a movement produced with hands, and alternatively also accompanied by an eyebrow raise or a head nod. The stimuli, available on the website *youtube.com*, were obtained from audiovisual recordings captured with hidden cameras in the talent show "Operación Triunfo" (1st edition), in which participants could not see the cameras but were aware of being recorded. A previous experiment was conducted with the aim of selecting, from an initial 30-sentence corpus (Jiménez-Bravo and Marrero-Aguiar, 2020), the most appropriate target sentences for this study, which was designed to be conducted online as four independent experiments. Care was taken that stimuli had similar grammatical complexity and similar word frequency. As a result, four prototypical sentences—two uttered by a male speaker and two uttered by a female speaker—were chosen for this study based on the consistent agreement on the prominence ratings showed by participants in the study by Jiménez-Bravo and Marrero-Aguiar (2020). The four target sentences, ranging between 9 and 18 words (M = 13.5, SD = 3.7), made up a total of 54 words available to each participant for marking.

Apart from four target sentences, trial and filler sentences also extracted from the talent show and having similar characteristics as the target sentences were introduced in the experiments, so that a total of 13 sentences were randomly presented to participants in each experiment, who received, firstly, three trial sentences to get familiar with the task, and then six target sentences for prominence marking (two non-manipulated for control purposes, and four manipulated ones); additionally, three filler sentences were also randomly interspersed between the target sentences (see Supplementary Table 2). When filler sentences were presented to participants, they were asked to report on either a visual element in the audiovisual modality or a word in the audio-only modality. The purpose of this was to make participants pay close attention to the images displayed on the screen (or uttered in the audio-only modality). In this way we tried to avoid a behavior previously observed, in which participants tended to close their eyes and concentrate only on the auditory signal when the stimuli were presented in the audiovisual modality, thus neglecting the images displayed on the screen (Krahmer and Swerts, 2007; Jiménez-Bravo and Marrero-Aguiar, 2020). Finally, two listeners trained in phonetics also provided marks of prominence to the non-manipulated stimuli both in the audio-only and in the audiovisual modality, so that their marks served as reference to compare with the participants' marks of prominence.

2.1.1 Acoustic manipulation

Three of the four experiments included manipulations of the speech signal, aiming to test the perceptual weight of the acoustic correlates available to listeners in each of them. The remaining experiment, whose stimuli had not been manipulated and kept the original speech signal, served as control. Consequently, each experiment included a different combination of manipulated cues: in Exp0 (control experiment) all acoustic cues were available to the listeners; in Exp1, the prominence-lending properties of F0 had been reduced, but not those of intensity and duration; in Exp2, the available cues were F0 and duration, but not intensity; in Exp3, only duration kept its prominence-lending properties, since both F0 and intensity had been manipulated. Spectral balance, which was also analyzed as a cue of prominence in all four experiments, was not directly manipulated, but was inevitably altered in all cases in which F0 and intensity had been manipulated.

The prominence-lending properties of F0 and intensity were reduced with Praat (Boersma and Weenink, 2023) in a similar way as described in Jiménez-Bravo and Marrero-Aguiar (2020), i.e., for the whole length of the stimulus, F0 was flattened within a 2-semitone range between its maximum and minimum values—following results on just noticeable differences ('t Hart, 1981; Pamies et al., 2002)—, while intensity was flattened at 69 dB; duration was not manipulated so as to avoid a perceptual mismatch between the sound of the speech signal and the articulation movements performed by the speaker in the image, which would lead to a significant increase in the artificiality of the task. For all experiments, regardless of the specific manipulations conducted, acoustic measures were taken for maximum F0 of lexically stressed vowels—or of the adjacent vowel if a F0 shift occurred—, mean intensity of lexically stressed vowels, and duration of stressed syllables. Next to these, spectral balance was also measured in lexically stressed vowels, regardless of vowel quality, as the difference between the intensity of the first harmonic (H1) minus that of the second harmonic (H2) (e.g., Campbell and Beckman, 1997) and was included in the statistical analyses. Table 1 summarizes the actual acoustic data analyzed in the four experiments.

2.1.2 Gesture annotation

By means of ELAN (Brugman and Russel, 2004) we annotated gestures in their different phases, i.e., preparation, stroke, apex, hold, and recoil (there were not any retraction phase in the samples) in a similar way as described in Jiménez-Bravo and Marrero-Aguiar (2020). The annotation was conducted for all gestures, regardless of the articulator they were performed with—whether it was hands, head or eyebrows separately or in combination (see Figure 1). Nonetheless, it was hand gestures that consistently appeared in all stimuli, often accompanied by a head nod and/or an eyebrow raise, and which often included all phases, since head nods and eyebrow raises often lacked some of the phases performed with hand gestures, e.g., preparation, recoil. Thus, gestures that included head nods and eyebrow raises consistently included the phases of stroke and apex, and occasionally head nods also included the phase of preparation. As can be seen in Figure 1, the phases of stroke and apex were annotated as coincidental for the several articulators that were involved in the realization of a certain gesture. We omitted any further categorization of gestures—e.g., beats, deictic, iconic, or metaphoric, or referential vs. non-referential—and did not compute the number of occurrences performed by each articulator separately or in combination.

2.2 Participants

The responses of 240 participants were gathered online, 60 for each experiment, i.e., 30 of them rated a given stimulus in one of the modalities—audio-only or audiovisual—, while the remaining 30 did so in the other modality. They were not financially compensated and were mainly recruited through social media to take part in a study that was advertised as an online study on memory and perception. None of the subjects reported any hearing disorders.

Several criteria were used to assure the reliability of the collected data. Firstly, it was ensured that participants had Castilian Spanish as their mother tongue and were settled in Spain at the moment of participation (according to their IP-addresses). Secondly, it was also ensured that they took at least 6 min to complete the experimental task, but no more than 13 min, with mean time for completion being 9 min 17 s (*SD* = 2 min 32 s), since we wished to avoid effects due to fatigue (Feenstra et al., 2017) and also wanted to prevent participants from overly relying on logical inferences in the prominence marking task. Initially, a total of 312 naıve listeners completed the four experiments conducted online, but after applying these criteria the answers provided by 240 of them (68 men and 172 women) were selected for the statistical analyses—30 per modality and experiment—, which added up to total of 12,960 rated words.

The main sources of individual variability being reported in previous studies were controlled for, i.e., age, gender and musical training (see Supplementary Figures 1–3). The declared age of participants ranged between 18 and 66 years (M = 36.98, SD = 10.55; M_{men} = 39.80, SD = 10.75; M_{women} = 35.86, SD = 10.26), with a predominance of participants under 50 years of age. Participants were also questioned about whether they had ever studied music and for how long (*¿Has estudiado alguna vez música? ¿Durante cuánto tiempo?*). According to their responses, they were grouped for the analyses into three categories: none, little, and much musical training. More precisely, those participants with some dexterity at playing a musical instrument and/or having up to 5 years of constant formal musical training were grouped as a having little musical training, while those who declared themselves as professional musicians or as non-professional musicians having over 5 years of constant formal musical training up to the moment of the experiment were considered as having much musical training.

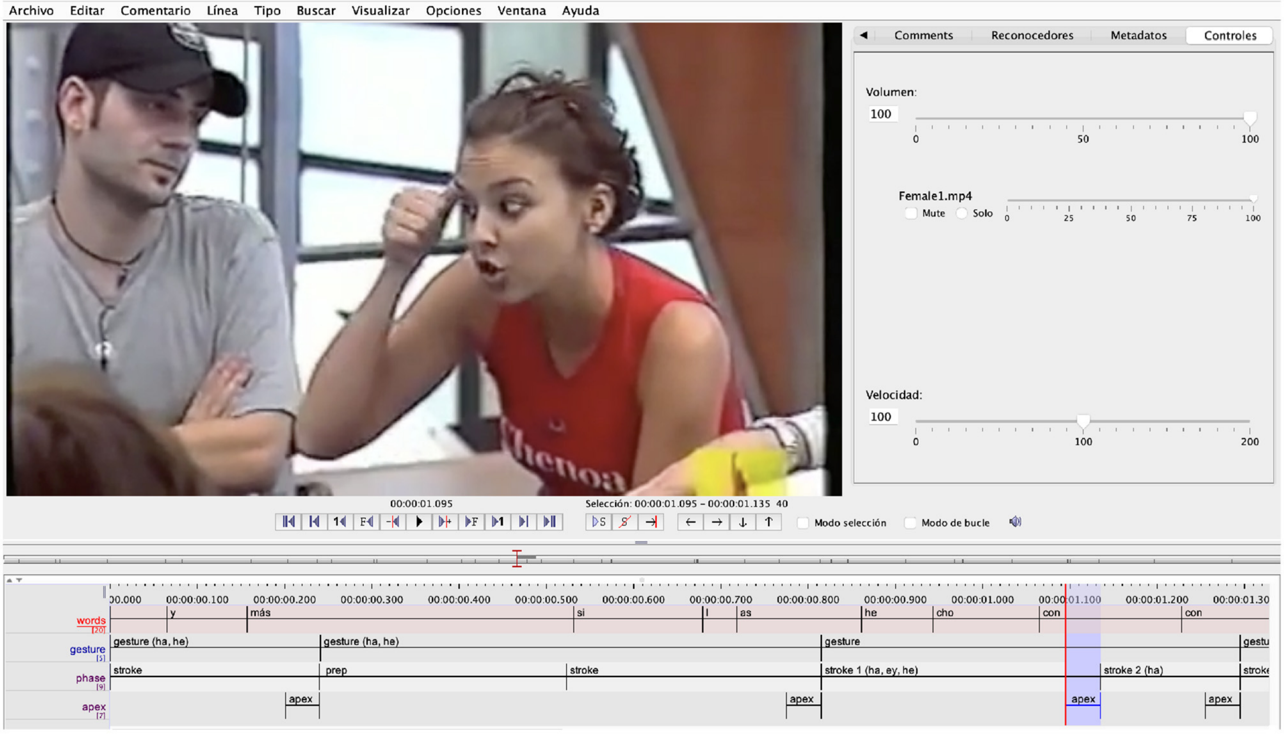
2.3 Procedure

A major challenge faced by every online experiment is the need to ensure the reliability and validity of the results. To achieve this, several requirements are necessary. Thus, in addition to controlling for individual variables, which were gathered through a questionnaire—i.e., age, gender, place of birth, mother tongue, and level of musical training—, questions were also asked during the experimental task to ensure that participants actually paid attention to the auditory and audiovisual stimuli, as detailed below.

The study was conducted using the online survey software *SmartSurvey*. At the recruitment stage, participants were asked to take around 10 min to conduct the experiment on a computer and use headphones in a quiet environment. Then, for the experiment they were firstly presented with a brief set of instructions explaining that they were expected to carry out two different tasks: sometimes they were asked to mark all "the words that are pronounced with more emphasis" in the sentence they were presented with; and some other times, they had to answer about a certain visual element that

TABLE 1 Ranges [in square brackets], mean and standard deviation (in parenthesis) of the analyzed acoustic values in each experiment.

Experiment	Manipulation	Values			
		F0 (Hz)	Intensity (dB)	Duration (s)	H1-H2
Exp0	None	[104, 307] 3 (40.0)	[62, 80] 73.1 (3.5)	[0.027, 0.379] 0.110 (0.069)	[−9.4, 21.6] 3.56 (5.81)
Exp1	F0 within a 2-st range	–	as in Exp0	as in Exp0	[−15.6, 10.4] 0.70 (4.58)
Exp2	Intensity at 69 dB	as in Exp0	–	as in Exp0	[−4.3, 14.3] 3.45 (4.92)
Exp3	F0 2-st + intensity 69 dB	–	–	as in Exp0	[−11.2, 12.0] 1.04 (4.22)



Volumen: 100

Velocidad: 100

00:00:01.095

Selección: 00:00:01.095 – 00:00:01.135 40

words	y	más	si	l	as	he	cho	con	con

gesture: gesture (ha, he) | gesture (ha, he) | gesture | gestu

phase: stroke | prep | stroke | stroke 1 (ha, ey, he) | stroke 2 (ha) | stroke

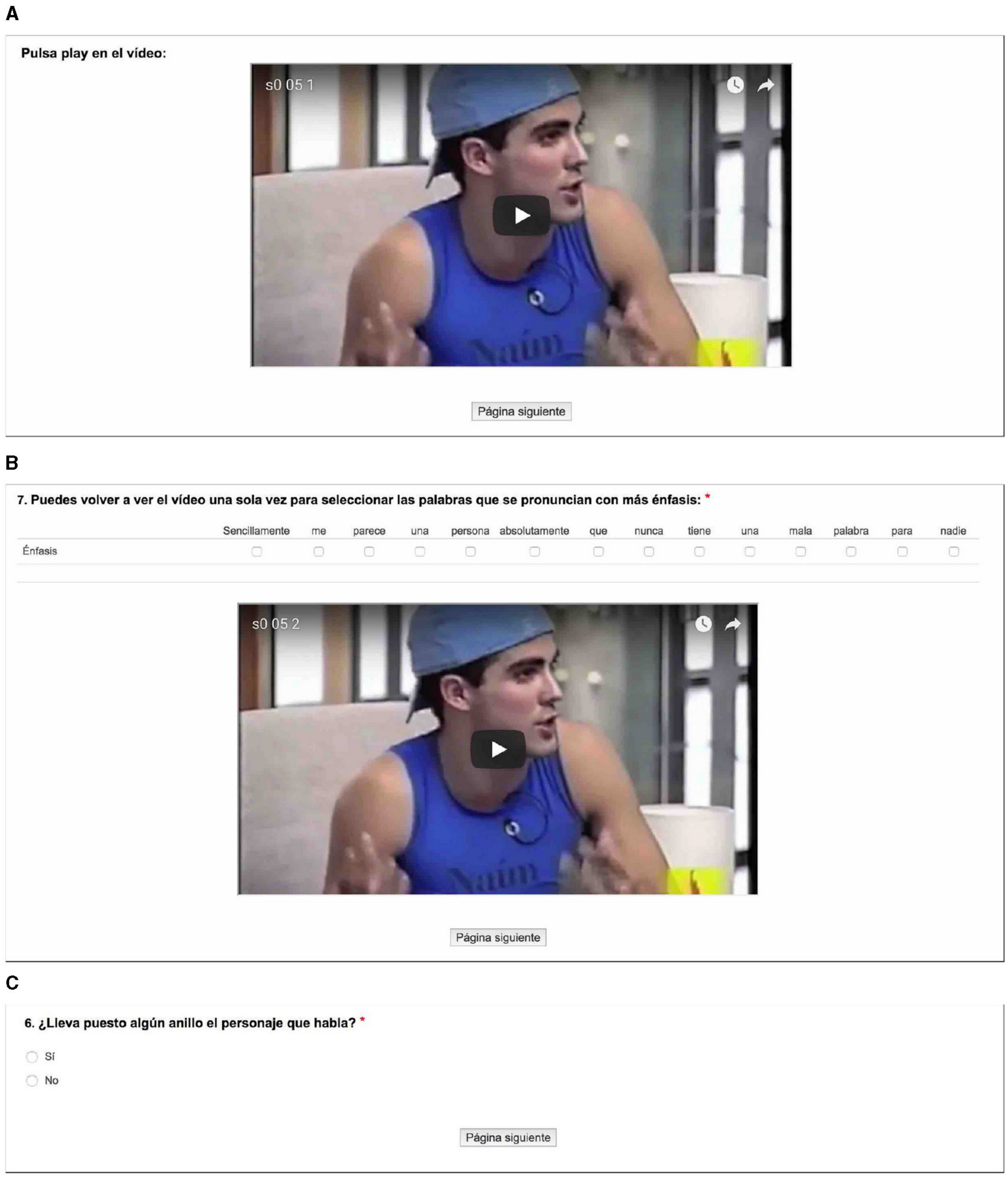
apex: apex | apex | apex | apex

FIGURE 1
Sample screen of gesture annotation in ELAN. Speaker on the right performs an eyebrow rise in its effort peak (apex) together with a head and hand movement. ha, hand; he, head; ey, eyebrows.

had been displayed on the screen or a certain word of the sentence they had just heard in the audio-only modality. Participants had the opportunity to get acquainted with the experiment in a series of trials. Then, after a first self-paced stimulus presentation, a second screen revealed either the experimental task or the filler question (Figure 2). In case they were presented with the experimental task, they were allowed to play back the clip just once more.

2.4 Statistical analyses

Firstly, the differences in the number of prominence marks in each experiment were analyzed by means of chi-square tests, and inter-rater agreement was calculated as the mean Cohen's kappa (1960) for all pairs of participants taking part in each experiment (n = 30). Additionally, a reference value was provided by the prominence marks given by two phonetically trained listeners. Subsequently, in order to achieve a more fine-grained scale of prominence the marks given by participants were distributed across experiments and modalities for all sentences and pooled over participants following the procedure of previous studies (e.g., Swerts, 1997; Cole et al., 2014). Thus, a prominence score (*P*-score) ranging between 0 and 1 was expressed as the proportion of participants who marked a certain word as prominent. More precisely, the number of marks given to a word in a sentence for each experiment and modality was divided by the number of possible total marks for that word, that is, equal to the number of participants. In order to uncover non-random error, kappa



Pulsa play en el vídeo:

7. Puedes volver a ver el vídeo una sola vez para seleccionar las palabras que se pronuncian con más énfasis: *

	Sencillamente	me	parece	una	persona	absolutamente	que	nunca	tiene	una	mala	palabra	para	nadie
Énfasis	□	□	□	□	□	□	□	□	□	□	□	□	□	□

6. ¿Lleva puesto algún anillo el personaje que habla? *

○ Sí

○ No

FIGURE 2
Three sample screens corresponding to the two experimental tasks in the audiovisual modality. Screens **(A, B)** correspond to the same stimulus for prominence marking: **(A)** shows a videoclip for participants to watch just once; after clicking on *Página siguiente* ("Next page") at the bottom, the screen showed in **(B)** displayed check-boxes to mark prominence for a given sentence that could be played back once more (in this example, "Simply put, he seems to me someone that never has a bad word for anyone"). Screen **(C)** shows an example of the filler task, with a question about an element seen in the video sequence—or a word heard in the audio clip in the audio-only modality—, with the purpose of making participants pay attention to the audiovisual information available to them.

coefficients were combined with Pearson's correlation between the *P*-score of the marks provided by two phonetically trained listeners and the rest of the participants in the four experimental conditions and the two modalities (e.g., Hunt, 1986).

Secondly, a global analysis for the four experiments combined was conducted through a set of generalized linear mixed models (GLMMs). These models were fitted with a logit link function for a binomial distribution in R Development Core Team (2023) by means of the *lme4* package (Bates et al., 2015). The declaration of an initial model was motivated both by the research questions and by previous results with which comparisons could be drawn (Jiménez-Bravo and Marrero-Aguiar, 2020), so as to validate the results obtained for the subcorpus used in this study. The dependent variable *prominence* was initially modeled (model G1) as a function of the fixed effects of *modality* in interaction with the rest of the predictors: *experiment*, which contained 4 levels—Exp0 (control experiment), TL (the marks given by two trained listeners), Exp1 (flat F0), Exp2 (flat intensity), Exp3 (flat F0 and flat intensity)—, the presence or absence of each gesture phase—i.e., *preparation*, *stroke*, *apex*, *hold*, *retract*—, and the standardized acoustic values of fundamental frequency (*z.ff*), intensity (*z.intensity*), duration (*z.dur*), and spectral balance (*z.H1H2*). The standardization of the acoustic variables was made per sentence, since participants marked prominent words depending on the phrasal environment in which these were uttered, while fundamental frequency was also standardized per speaker to avoid bias in pitch resulting from gender differences.

In this global analysis, random effects were declared in the initial model only with varying intercepts for both by-subjects and by-items, although varying slopes were later declared as model building proceeded. By-subjects random effects included *participants*—i.e., the raters of prominence in the perceptual experiments—and the by-items included the nested variables of *word* within *sentence* within *speaker*. In every model, optimization was carried out with *bobyqa* (Powell, 2009) to avoid non-convergence errors. In this global analysis, and in order for our results to be more directly comparable with previous results, the predictors *gender* and *musical training* were not included. Then, model selection proceeded by the progressive removal of factors irrelevant for, and justified by, the research. For this the Akaike Information Criterion (AIC) served as a way to rank models (Akaike 1973; see Burnham and Anderson, 2002, for a review), so that a comparison of their AIC values through the package *AICcmodavg* (Mazerolle, 2023) eventually revealed the minimal adequate model, the one with the lowest AIC value. Nonetheless, all models within <2 Δ-points from this minimal adequate model were also reported.

Thirdly, in a second analysis, which allowed a deeper assessment of the variables of interest, generalized linear mixed models (GLMMs) were fitted separately for each experiment and modality. In this case, the dependent variable *prominence* was predicted either as a function of the visual cues in the audiovisual modality or as a function of the audio cues in the audio-only modality. In this second analysis, all models also included two more variables that controlled for the musical training of participants and their gender.

3 Results

3.1 Marks of prominence

Each of the 240 participants marked 54 words, so that the total number of marked words was 12,960, out of which 3,202 (24.1%) received a mark of prominence. Prominence marks per sentence ranged between 2.76 and 3.97 (M = 3.33, SD = 1.98), with no difference between both speakers [χ^2 (1) = 0.40, $p > 0.05$].

3.1.1 Prominence and gesture phases

The audio-only modality served as the baseline to compare the marks given to words coinciding with each gesture phase in the audiovisual modality. In the control experiment, Exp0 (non-manipulated stimuli), the phases that received more marks of prominence in the audiovisual modality were strokes (+9.5%) [χ^2 (1) = 17.55, $p < 0.001$], and apexes (+8.9%) [χ^2 (1) = 12.81, $p < 0.001$]. However, words coinciding with strokes showed significant differences also in Exp2 (flat intensity) [χ^2 (1) = 15.31, $p < 0.00$]. Differences for holds were only observed in Exp2 (flat intensity) [χ^2 (1) = 10.42, p = 0.001]. No differences were found for preparation nor recoil phases in any experiment (Table 2).

In short, the chi-square tests revealed that, under normal acoustic conditions, as in Exp0 (control experiment), participants seemed to consider more words as prominent when strokes and apexes—typically performed with hands, but also possibly coinciding with a head nod and/or an eyebrow raise—co-occurred with prominence signaled by auditory cues. However, the manipulations in the audio signal often overrode this effect, as seen in the fewer marks of prominence given to these gesture phases in Exp1, Exp2, and Exp3.

3.1.2 Prominence and acoustic cues

In this section the results are detailed per experiment and modality, as can be seen in Table 3, and Figure 3. Overall, the control experiment (non-manipulated stimuli) received fewer marks of prominence than any of the three experiments involving manipulated stimuli. Furthermore, when Exp0 was compared with the experiments whose signal was manipulated, the marks given in the audio-only modality consistently increased: in Exp1 (flat F0) [χ^2 (1) = 9.19, p = 0.002]; in Exp2 (flat intensity) [χ^2 (1) = 22.72, $p < 0.00$]; and in Exp3 (flat F0 and flat intensity) [χ^2 (1) = 6.49, p = 0.01] (Figure 3A). However, this trend was reversed in the audiovisual modality, where the visual information co-occurring with degraded acoustic cues reached a maximum in Exp0 but decreased in all three experiments, yielding significant differences in Exp2 [χ^2 (1) = 6.36, $p < 0.01$], and in Exp3 [χ^2 (1) = 5.16, $p < 0.02$] (Figure 3B). Consequently, a possible interpretation is that, when only the audio is available, the degradation of the acoustic signal makes prominent words stand out less from their environment and creates uncertainty as to where prominence lies, while the visual information allows to maintain a higher degree of confidence in this judgement, resulting in fewer marks of prominence. Furthermore, when crossing experiments and modalities, significant differences between both modalities were found within Exp0 (non-manipulated signal) [χ^2 (1) = 10.93, $p < 0.001$], and within Exp2 (flat intensity) [χ^2 (1) = 9.55, p = 0.001], yet in a different direction in each case (Figure 3C).

Marks of prominence given by participants per experiment and modality.

In sum, these results show that the visual information interacts in a complex way with the acoustic cues of prominence and that the visual cues of prominence may not have an additive effect on the marks of prominence but, rather, a subtractive one. Probably, when gestural information is available, a coincidence between the visual and auditory information is required, otherwise words tend not to be considered prominent.

TABLE 2 Different gesture phases accompanying words marked as prominent in all experiments and modalities.

Phase	Words	Marked as prominent (%)					
		Exp0 (control)			Exp1 (flat F0)		
		A	AV	AV (±%)	A	AV	AV (±%)
Preparation	272	25 (8.3)	29 (9.6)	+1.3	37 (12.3)	33 (11.0)	−1.3
Stroke	422	27 (6.0)	70 (15.5)	+9.5	57 (12.6)	74 (16.4)	+3.8
Apex	538	214 (41.9)	260 (51.0)	+8.9	228 (44.7)	218 (42.7)	−2.0
Hold	208	54 (30.0)	55 (30.5)	+0.5	62 (34.4)	52 (28.8)	−5.6
Recoil	180	21 (11.6)	19 (10.5)	−1.1	31 (17.2)	28 (15.5)	−1.7
Total	1,620	341 (21.0)	433 (26.7)	+ 5.7	415 (25.6)	405 (25.0)	−0.6
		Exp2 (flat intensity)			Exp3 (flat F0 + intensity)		
		A	AV	AV (±%)	A	AV	AV (±%)
Preparation	272	43 (14.3)	29 (9.6)	−4.7	19 (6.3)	27 (9.0)	+2.7
Stroke	422	85 (18.8)	43 (9.5)	−9.3	54 (12.0)	50 (11.1)	−0.9
Apex	538	238 (46.6)	237 (46.4)	−0.2	230 (45.1)	215 (42.1)	−3.0
Hold	208	68 (37.7)	39 (21.6)	−16.1	63 (35.0)	58 (32.2)	−2.8
Recoil	180	25 (13.8)	22 (12.2)	−1.6	37 (20.5)	26 (14.4)	−6.1
Total	1,620	459 (28.3)	370 (22.8)	−5.5	403 (24.8)	376 (23.2)	−1.6

Values for the audio-only modality, where no visual information was available, served as the baseline to compare the marks given by listeners in the audiovisual modality. Values in parentheses express percentage.

TABLE 3 Marks of prominence given by participants per experiment and modality.

Modality	Words	Marked as promient (%)			
		Exp0 (control)	Exp1 (flat F0)	Exp2 (flat intensity)	Exp3 (flat F0 + intensity)
A+AV	3,240	774 (23.8)	820 (25.3)	829 (25.5)	779 (24.0)
A	1,620	341 (21.0)	415 (25.6)	459 (28.3)	403 (24.8)
AV	1,620	433 (26.7)	405 (25.0)	370 (22.8)	376 (23.2)
AV (±%)		+5.7	−0.6	−5.5	−1.6

Values in parentheses express percentage.

3.1.3 Inter-rater agreement and prominence score (P-score)

Inter-rater agreement among participants, as can be seen in Table 4, was highest in the audio-only modality for non-manipulated stimuli (Exp0, $\kappa = 0.412$), although the degree of agreement decreased in the audiovisual modality ($\kappa = 0.343$), as was also the case for the marks provided by the two phonetically trained listeners (audio-only, $\kappa = 1.00$; audiovisual, $\kappa = 0.89$). Such agreement values for non-manipulated stimuli were not very different from those obtained in other similar experiments (e.g., Mo et al., 2008; Bishop et al., 2020).

The manipulation of the acoustic cues of prominence in the three independent experiments yielded a poorer inter-rater agreement than in Exp0 in the audio-only modality, suggesting that as the acoustic signal degraded, so did agreement. However, in the audiovisual modality this pattern did not occur, and agreement was almost as high in Exp2 (flat intensity) as in Exp0. This could suggest that, at least in the experiments conducted in this study, visual cues may compensate to some extent for the loss of acoustic information.

Finally, prominence scores (*P*-scores) showed a high degree of consistency between the prominence marks given by participants when compared to those given by phonetically trained listeners, while Pearson's correlation between the *P*-score of the trained listeners and that of participants showed a fairly high correlation in each experiment (Table 5; Supplementary Table 3). The greatest *P*-score was found on clearly prominent words, which co-occurred with the apex of a gesture, such as *nunca* "never" (sentence 1, see Supplementary Figure 4), *unos* "ones/some of them" and *otros* "others" (sentence 2, see Supplementary Figure 5), *ir a saco* "go all out" (sentence 3, see Supplementary Figure 6) and *más* "more" (sentence 4, see Supplementary Figure 7). Although the words in our experiment are in everyday use—i.e., they all appear in the *Corpus del Español del Siglo XXI* (CORPES XXI)—, there is no correlation between their use frequency and the prominence marks they received. For example, the second most marked word, *[a] saco*, has a normalized frequency of 0.33, which is very low. Nor did prominence marks seem to be conditioned by the grammatical category of words, so that among the 10 most marked words we find nouns, proper names, adverbs, articles, and pronouns. It seems

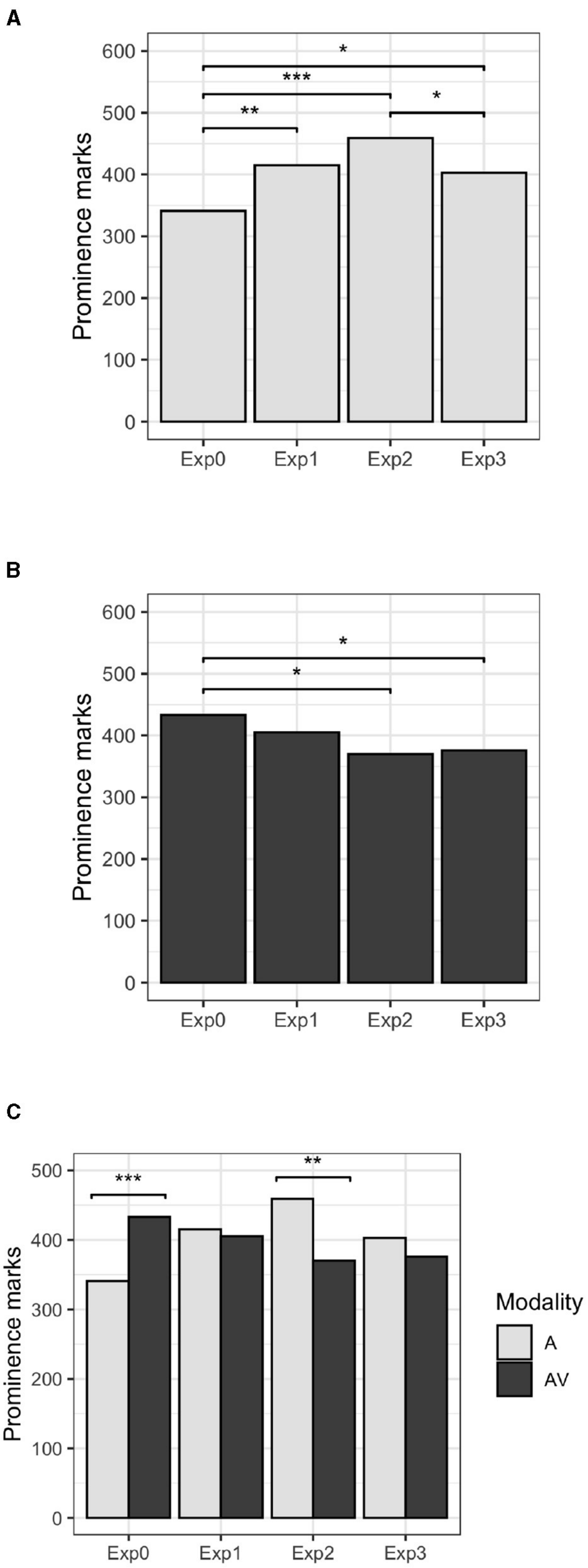


FIGURE 3
Number of prominence marks given by participants per modality (audio-only in top panel, audiovisual in middle panel, and their comparison in bottom panel) in each experiment: Exp0 (control), Exp1 (flat F0), Exp2 (flat intensity), Exp3 (flat F0 + flat intensity). **(A)** Audio-only modality. **(B)** Audiovisual modality. **(C)** Both modalities. $^*p < .05$, $^{**}p < .01$ and $^{***}p < .001$.

that the variables that determine the probability of words of being marked are linked to each specific communicative act and depend on their pragmatic function in that interaction.

3.2 Generalized linear mixed models analyses

3.2.1 Global model

All experiments that included manipulated auditory cues (Exp1, Exp2, and Exp3) were compared to the control experiment Exp0 in a global analysis using the procedure previously described (see Jiménez-Bravo and Marrero-Aguiar 2020, for details). Additionally, the marks provided by two trained listeners were also included to compare with the markings of the 60 participants (30 per modality) taking part in the control experiment, Exp0.

Model building proceeded from models G1 to G27. Initially, the different non-significant predictors were progressively removed, which resulted in a progressive decrease of their AIC value. Next, the variance of by-item random effects was observed to be almost entirely captured by *sentence* and *word*, and the upper level of the nested by-item random effect, *speaker*, was removed from model G13 onwards. Additionally, from model G14 to model G27 slopes for by-items random effects were declared, and they included either the variable *modality* or both variables *modality* and *experiment*. From these subsequent models, G25 yielded the lowest value (AIC = 9,919.99) (see Supplementary Table 4).

The estimates for the predictors of G25 revealed that participants of Exp0 (control experiment) in the audio-only modality did not perform differently from two trained listeners whose marks served as reference (Figure 4). However, when compared to Exp0, participants were more likely to mark words in Exp1 (β = 0.72, SE = 0.27, z = 2.68, p = 0.007), in Exp2 (β = 0.66, SE = 0.26, z = 2.48, p = 0.013), and in Exp3 (β = 0.65, SE = 0.27, z = 2.44, p = 0.014). In this global comparison, this effect was reversed in Exp2 (β = –0.58, SE = 0.37, z = –2.73, p = 0.006), as seen in the interaction between *experiment* and *modality*, with Exp3 falling short of significance (p = .076). Furthermore, participants were more likely to mark prominent words by the overall effect of strokes (β = 1.60, SE = 0.26, z = 6.06, p < 0.001) and holds (β = 1.25, SE = 0.30, z = 4.25, p < 0.001); as well as by the effect of F0 (β = 0.20, SE = 0.04, z = 5.15, p < 0.001) and duration (β = 0.82, SE = 0.14, z = 5.86, p < 0.001). Intensity and spectral balance did not seem to generally contribute to predict the marks of prominence given by participants in this global model.

3.2.2 Models by experiment

3.2.2.1 Exp0 (non-manipulated stimuli)

In this control experiment two mixed models—one for each modality—were fitted to predict prominence from the marks given by participants relying on a non-manipulated signal (Figure 5, see Supplementary Table 6). The model for the audio-only modality revealed that, in the absence of the visual cues of prominence, participants made use of all the acoustic cues at their disposal—except for spectral balance—, with duration showing the strongest effect (β = 1.49, SE = 0.21, z = 7.12, p < 0.001), while F0 (β = 0.66,

TABLE 4 Details of inter-rater agreement (mean Cohen's kappa) for all pairs of participants per experiment and modality.

Experiment	Modality			
	Audio-only (A)	A-TL	Audiovisual (AV)	AV-TL
Exp0 (control)	0.412	0.431	0.343	0.289
Exp1 (flat F0)	0.282	0.299	0.269	0.278
Exp2 (flat intensity)	0.275	0.292	0.380	0.388
Exp3 (flat F0 + intensity)	0.342	0.357	0.366	0.369
Trained listeners (reference)	1.00		0.89	

Additionally, values for the same pairwise comparisons of all participants but including also those of two T(rained) L(isteners) are given as A-TL and AV-TL.

TABLE 5 Correlation (Pearson coefficient) among trained listeners and participants per experiment and modality.

Experiment	Modality	
	Audio-only	Audiovisual
Exp0 (control)	0.820	0.822
Exp1 (flat F0)	0.720	0.750
Exp2 (flat intensity)	0.717	0.814
Exp3 (flat F0 + intensity)	0.716	0.742
Mean	0.743	0.782

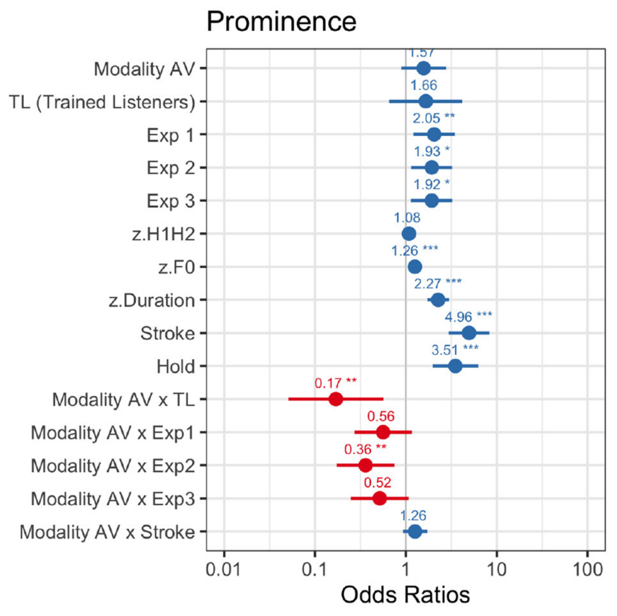


FIGURE 4
Global model. For OR < 1, the effect size equals 1/OR (95% CI).

SE = 0.22, *z* = 2.97, *p* = 0.003) and intensity (β = 0.64, *SE* = 0.24, *z* = 2.67, *p* = 0.007) contributed similarly to the responses given by participants.

Interestingly, the model for the audiovisual modality showed that intensity no longer played a role in cueing prominence when the visual cues were present, while the effect of duration was reduced. Only the effect of F0 remained similar to that in the audio-only modality (β = 0.65, *SE* = 0.17, *z* = 3.78, $p < 0.001$), and that of duration was reduced (β = 1.25, *SE* = 0.17, *z* = 7.58, $p < 0.001$). As for the visual cues of prominence, and as seen by the size of their effect, they seem to have had a stronger perceptual effect than the effect of the auditory cues, with strokes (β = 1.73, *SE* = 0.68, *z* = 2.53, *p* = 0.01) and holds (β = 1.75, *SE* = 0.81, *z* = 2.17, *p* = 0.03) being significant.

3.2.2.2 Exp1 (flat F0)

Two models were fitted to assess how the different variables predict the marks of prominence given by participants in the experiment lacking the prominence-lending properties of F0 (Figure 6, see Supplementary Table 7). Firstly, a stronger effect of duration over intensity in cueing prominence was found in both modalities. In audio-only, prominence marks were predicted mainly by intensity (β = 0.31, *SE* = 0.15, *z* = 1.99, *p* = 0.04) and duration (β = 0.97, *SE* = 0.17, *z* = 5.71, $p < 0.001$). As before, spectral balance was not relevant, probably as a result of the manipulation of F0. In the audiovisual modality, the effect of duration increased (β = 0.95, *SE* = 0.15, *z* = 6.40, $p < 0.001$), and so did the effect of intensity (β = 0.36, *SE* = 0.14, *z* = 2.53, *p* = 0.01).

In this experiment, where minimal variations of F0 were present in the signal, the group of participants with more than 5 years of musical training were more likely to give words a mark of prominence (β = 1.57, *SE* = 0.56, *z* = 2.79, *p* = 0.005), but only in the audio-only modality, while in the audiovisual modality, the responses of participants did not varied as a result of their level of musical training. As for the visual cues of prominence, only the stroke phase of gestures increased the odds of words to be marked as prominent (β = 1.72, *SE* = 0.70, *z* = 2.44, *p* = 0.01).

3.2.2.3 Exp2 (flat intensity)

This experiment tested the effects of the acoustic correlates of F0 and duration after intensity had been flattened at 69 dB (Figure 7, see Supplementary Table 8). In the audio-only modality, duration (β = 1.21, *SE* = 0.18, *z* = 6.72, $p < 0.001$), was found to have a larger effect than F0, the other remaining cue in the signal (β = 0.63, *SE* = 0.16, *z* = 3.87, $p < 0.001$); while in the audiovisual modality, the same larger effect of duration (β = 1.28, *SE* = 0.20, *z* = 6.34, $p < 0.001$), over F0 (β = 0.58, *SE* = 0.19, *z* = 3.11, *p* = 0.001), was observed. Finally, the group of participants with up to 5 years of musical training were less likely to mark words than those with no training at all (β = –1.44, *SE* = 0.46, *z* = –3.09, *p* = 0.002). A *post-hoc* Tukey comparison for this variable revealed no other differences

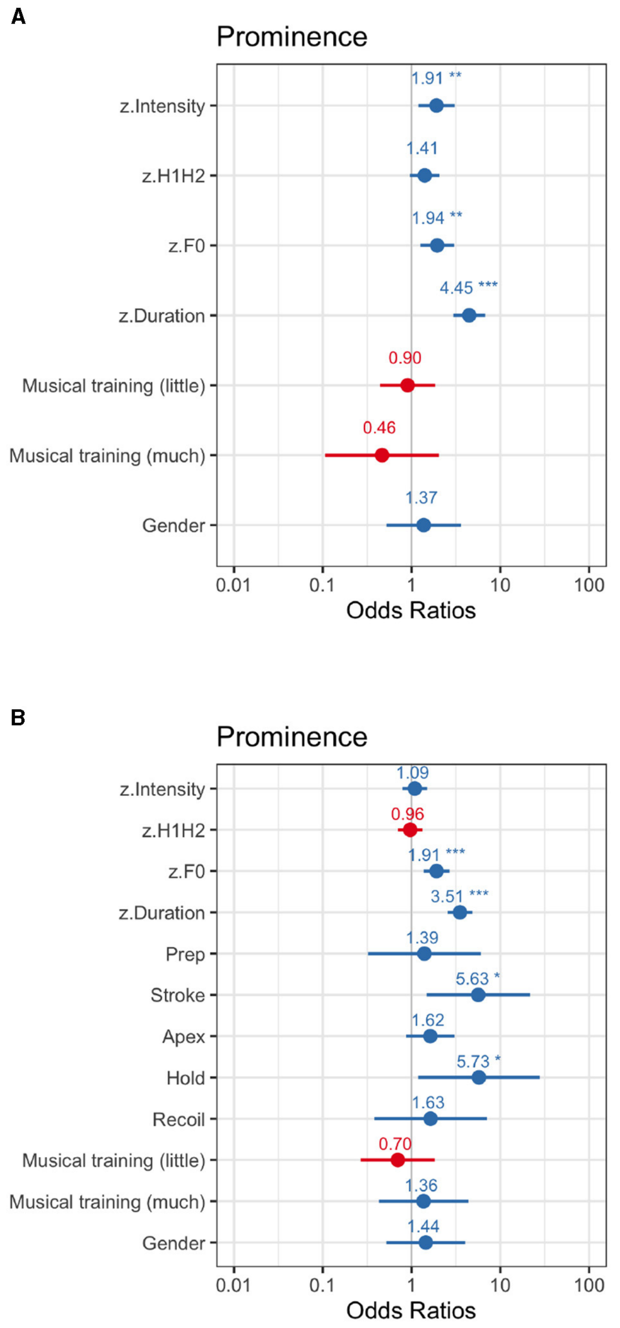


FIGURE 5
Exp0 (control experiment). For OR <1, the effect size equals 1/OR (95% CI). **(A)** Audio-only modality. **(B)** Audiovisual modality. *$p < .05$, **$p < .01$ and ***$p < .001$.

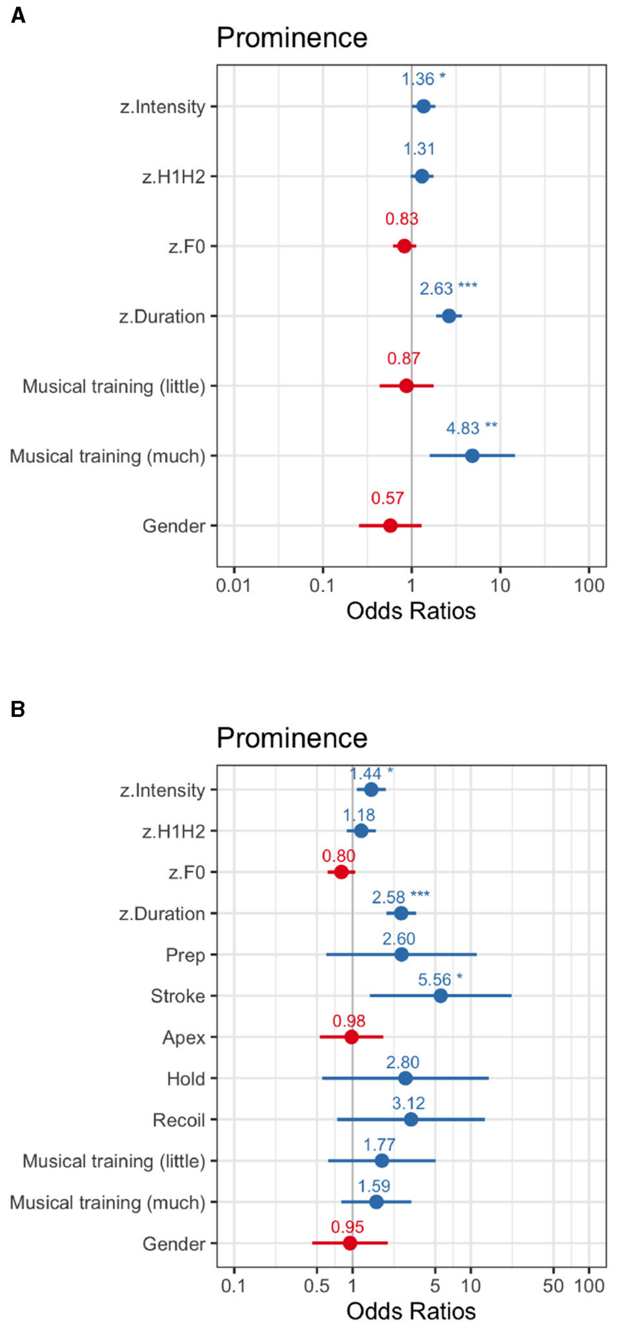


FIGURE 6
Exp1 (flat F0). For OR <1, the effect size equals 1/OR (95% CI). **(A)** Audio-only modality. **(B)** Audiovisual modality.

between those participants with <5 years and those with more than 5 years of musical training.

3.2.2.4 Exp3 (flat F0 and flat intensity)

In this experiment the prominence-lending properties both of F0 and intensity had been reduced (Figure 8, see Supplementary Table 9). The statistical analyses revealed again a consistent effect of syllable duration both in the audio-only (β = 1.34, *SE* = 0.22, z = 6.00, p < 0.001), and in the audiovisual modality (β = 1.26, *SE* = 0.19, z = 6.46, p < 0.001). Among the visual cues of prominence, none of the gesture phases had an effect on the marks of prominence, probably due to the strong degradation of the acoustic cues of prominence.

Furthermore, when controlling for the effect of gender, a difference in the performance between men and women was found in the audiovisual modality (β = 0.95, *SE* = 0.32, z = 2.99, p = 0.002), with women being more likely than men to give words a mark of prominence. This could be related to the fact that

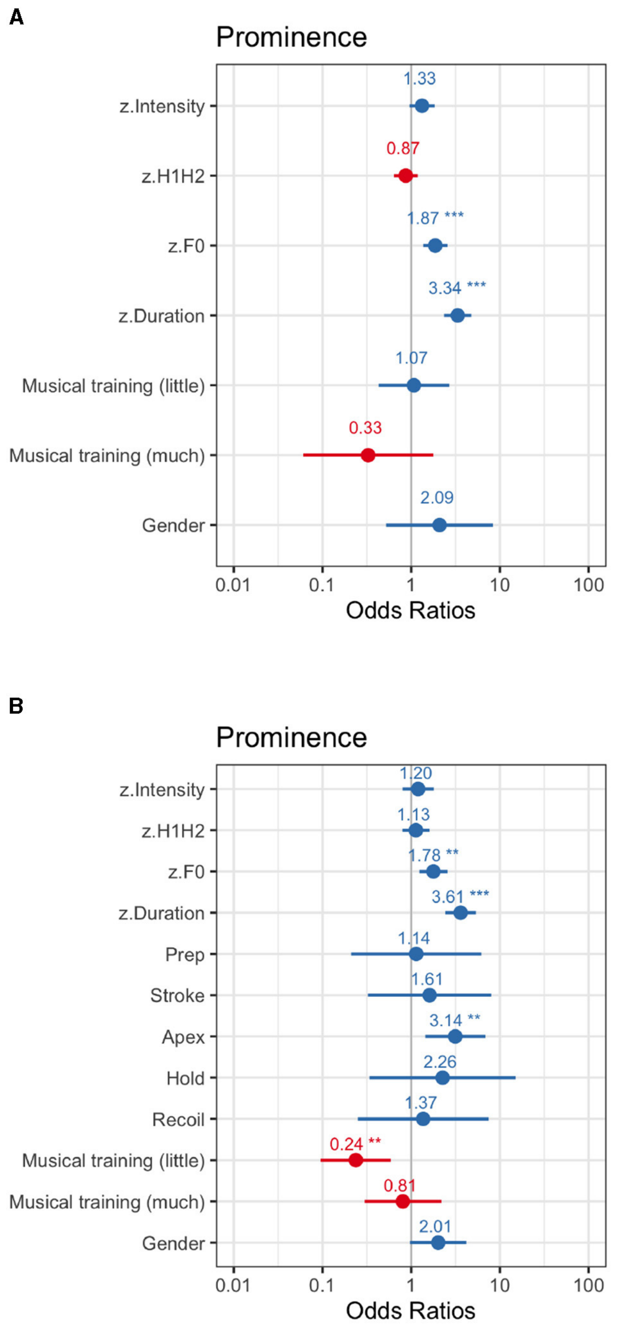


FIGURE 7
Exp2 (flat intensity). For OR <1, the effect size equals 1/OR (95% CI). **(A)** Audio-only modality. **(B)** Audiovisual modality. ***p* < .01 and ****p* < .001.

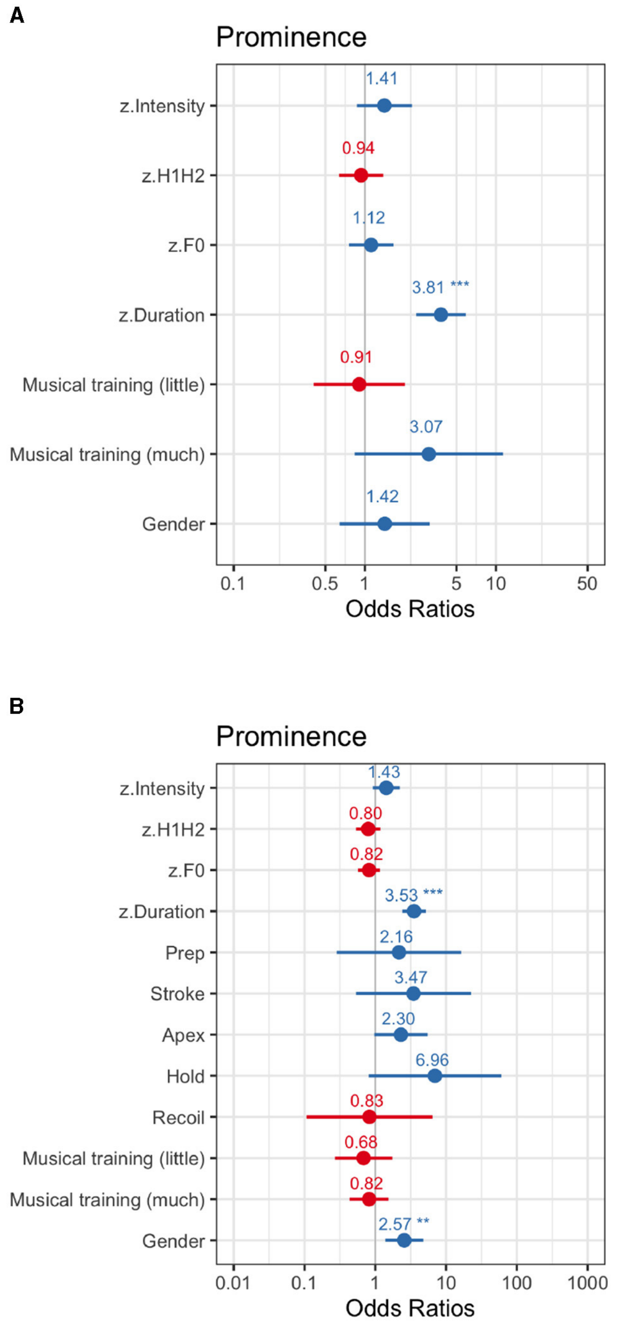


FIGURE 8
Exp3 (flat F0 and flat intensity). For OR <1, the effect size equals 1/OR (95% CI). **(A)** Audio-only modality. **(B)** Audiovisual modality. ***p* < .01 and ****p* < .001.

women are generally considered to be more active gazers than men when processing multimodal information under adverse acoustic conditions (Johnson et al., 1988; Jaeger et al., 1998).

4 Discussion

The aim of this study was to analyze the contribution of gestures in multimodal communication, and also to gain insight into the role and interaction of the acoustic cues of prominence in Castilian Spanish. Building on the methodology described in a previous study (Jiménez-Bravo and Marrero-Aguiar, 2020), four independent experiments in two modalities—audio-only and audiovisual—were conducted online, each involving a different manipulation of the acoustic correlates of prominence, i.e., suppression of the prominence-lending properties of F0 in experiment Exp1; of intensity in experiment Exp2; and both of F0 and intensity in Exp3; the control experiment Exp0 did not involve any manipulation of the spontaneous speech signal.

Firstly, all four experiments were compared in a global analysis, which also included a reference provided by the marks of two phonetically trained listeners on non-manipulated stimuli. The results showed no differences between these marks serving as reference and those of the control experiment. However, an overall effect was observed in the audio-only modality for words to be "overmarked" in all the experiments having manipulated stimuli. On the contrary, an opposite trend was found when participants could rely on the visual cues of prominence in the experiments involving manipulated stimuli, where words tended to receive fewer marks of prominence. Additionally, F0 and duration had an overall effect on the marks given by participants. These results replicated those of a previous study conducted on a larger corpus (Jiménez-Bravo and Marrero-Aguiar, 2020), showing that in our study the use of a larger sample size of participants with a smaller sample size of target sentences yielded similar results. In this sense, the methodological differences in our experiment respect to that by Jiménez-Bravo and Marrero-Aguiar allowed participants to complete the experiments online without suffering a fatigue effect. In addition, the between-subjects design employed here also permitted to conduct independent analyses for both the auditory and the audiovisual cues of prominence for each experiment.

These independent analyses confirmed our first hypothesis, and gestures—whether performed with hands, head, or eyebrows separately or in combination—were used by listeners as a cue of prominence together with the acoustic changes in F0, intensity, or duration. So did our second hypothesis, and in the audio-only modality, lacking any gestural cues, duration served as a sufficient cue to detect prominence, even in the absence of either F0 or intensity. As for our third hypothesis, namely that the apex phase of gestures drives the perception of visual prominence due to its synchronization with the prosody of the verbal signal (Kendon, 1972; Jannedy and Mendoza-Denton, 2005; Loehr, 2012; Esteve-Gibert and Prieto, 2013), it was only partially confirmed, since the stroke was the gesture phase that most often coincided with prominence marks, especially when F0 information was absent.

4.1 Multimodal interaction between gesture and speech

So far most of the previous methodologies applied to the study of the multimodal perception of prominence have used animated agents (e.g., Krahmer et al., 2002a; Al Moubayed and Beskow, 2009) or multimodal stimuli elicited in experimental settings (e.g., Krahmer and Swerts, 2007; Foxton et al., 2010). Very few studies have made use of spontaneous speech, from which only Swerts and Krahmer (2010, experiment 1) conducted a perceptual experiment to study prominence perception, and they did so only in the auditory modality. The limitations inherent to these methods have not allowed yet to study in detail the interplay of the different acoustic correlates of prominence and the exact role played by gestures in relation to them. However, in the present study, spontaneous speech extracted from a talent show ("Operación Triunfo", 1st edition) was used to overcome some of the mentioned limitations, proving that the multimodal perception of prominence can benefit from such a methodological shift.

In fact, our analyses have made evident that visual information interacts with the auditory perception of prominence. In natural non-manipulated stimuli (Exp0), the effect of duration was largely reduced in the audiovisual modality, while intensity stopped cueing prominence. Differently, the effect size of F0 hardly changed in the control experiment by the effect of the visual cues of prominence, and it remained unaffected in the audiovisual modality of Exp2. Such an influence of the visual information on the perception of the auditory signal is supported by the large number of studies accounting both for the strong connection between gesture and speech and for the potential of gestures to enhance the perception of prominence (e.g., Granström et al., 1999; House et al., 2001; Krahmer et al., 2002a,b; Krahmer and Swerts, 2007; Al Moubayed and Beskow, 2009; Scarborough et al., 2009; Foxton et al., 2010; Prieto et al., 2011).

In this regard, visual cues have previously been found to have a stronger perceptual effect than the acoustic cues of prominence (Prieto et al., 2011), which is in line with the general reduction in effect size found here for the acoustic cues when perceived together with visual information in Exp0 (control experiment). By the same token, Jiménez-Bravo and Marrero-Aguiar (2020) observed that when marking for prominence audiovisually, participants did not need stressed syllables to be as high in pitch or as long in duration as in the audio-only modality to give it a mark of prominence.

Additionally, intensity changes may be more affected by visual prosodic information than F0 changes, as intensity is thought to be more correlated with articulatory gestures than F0 (Scarborough et al., 2009; Foxton et al., 2010). For example, Foxton et al. (2010) observed that participants were able to detect both F0 changes and intensity changes in both modalities (audio-only and audiovisual); however, participants could better detect the thresholds of auditory cues when they were accompanied by visual information, and especially so for intensity. Furthermore, intensity has been reported to be processed together with duration as a unit in lexical stress perception, and minimal variations of duration had a larger effect on the perception of loudness than minimal variations of intensity in the perception of syllable length (Turk and Sawusch, 1996). In summary, our results indicate that in non-manipulated stimuli (Exp0) the perceptual effect of intensity can be expendable in the presence of visual information but not that of F0. Conversely, when the acoustic cues of prominence are degraded, the effect of intensity remains constant in the audiovisual modality, as seen in Exp1, where it cued prominence together with duration.

Apart from this, the audiovisual information consistently influenced the number of marks of prominence given by participants, as observed in Exp0, where more marks of prominence were given in the audiovisual modality than in the audio-only modality. Conversely, once the signal was manipulated, participants "overmarked" words in the audio-only modality across experiments. This "overmark effect", however, disappeared in the audiovisual modality (see Table 3). In our view, this is suggestive of a compensatory mechanism in the audiovisual modality prompted by the gestures of prominence, making participants more conservative raters when perceiving audiovisually the same degraded speech signal that is present in the audio-only modality.

In other words, under adverse acoustic conditions—as in Exp1, Exp2, and Exp3—, participants may not be sure of what words are prominent in the audio-only modality, and as a result they tend to mark words more generously than under normal conditions—as in Exp0, non-manipulated signal. This effect is reversed when they can see the speaker, so that the uncertainty introduced by the manipulation of the signal becomes less determinant. Consequently, at that point participants may realize that certain stimuli do not contain so many prominent words after all. That is in line with the better inter-agreement found for the audiovisual modality across experiments when compared to the audio-only modality (see Table 4). In this sense, visual information in the form both of beats (Krahmer and Swerts, 2007) and facial gesturing (House et al., 2001; Swerts and Krahmer, 2008; Dohen and Lœvenbruck, 2009) has been related to a stronger production and perception of verbal prominence, and some studies on the neural integration and processing of gesture and speech have also pointed out that beat gestures might drive listeners' attention and help them to process speech prosody and other relevant aspects of the spoken signal (e.g., Granström et al., 1999; Krahmer and Swerts, 2007; Scarborough et al., 2009; Al Moubayed et al., 2010; Prieto et al., 2011; Kim et al., 2014; Biau et al., 2015).

4.2 Phases of gestures

The several phases of gestures—performed mostly with hands, but also with head and/or eyebrows—that were analyzed in the present study may also offer better insight into how visual information contribute to the perception of prominence. There are two main findings in this respect, showing how degradation of the acoustic signal affects the integration of visual and auditory information.

4.2.1 Natural non-manipulated speech

The model fitted for the audiovisual modality of Exp0 indicated that strokes played an important role, as well as holds, in driving the attention of participants. However, the chi-square tests showed a significant increase in the audiovisual modality only in the number of marks for strokes—as well as for apexes (see Table 2). This suggests that, despite the large effect size observed for holds, when the number of marks given in the audiovisual modality was compared to those marks given in the audio-only modality, it was mostly strokes—and to a lesser extent also apexes—that made participants give words a mark of prominence. Strokes typically coincide with stressed syllables (e.g., Loehr, 2012; Esteve-Gibert and Prieto, 2013; Rohrer et al., 2023) and might be perceptually more salient than apexes, which correspond to the part of strokes that is time-aligned with F0 peaks. In addition, by the size of their effect, the visual cues of prominence seem to have contributed more than the auditory cues to drive the perception of participants. This finding is in line with previous results reporting a stronger perceptual effect of visual over auditory information (Prieto et al., 2011), although some studies reported that either auditory cues play a more important role (e.g., Swerts and Krahmer, 2004, 2008) or that both are perceptually integrated and none is predominant (Dohen and Lœvenbruck, 2009).

4.2.2 Degraded acoustic information

As seen in the statistical models, participants relied on strokes in Exp1 (flat F0), while apexes proved significant only in Exp2 (flat intensity, F0 present). In Exp3 (flat F0 and flat intensity), where duration was the only acoustic cue, none of the phases of the gesture seem to have played any important role, despite the fact that the visual information was available for the marking of prominence. Our interpretation is that in Exp1 and Exp3 the manipulation of F0 can cause a loss of connection between pitch accents and apexes, a connection consistently reported in literature (e.g., Kushch and Prieto Vives, 2016). So, our results suggest that both strokes and apexes can be perceptually relevant in a partially degraded speech signal, as in Exp1 and Exp2, respectively, but as the speech signal becomes more degraded, as in Exp3, none of them suffices to drive the participants' perception.

To our knowledge, the only study with a degraded speech signal lacking the prominence-lending properties of acoustic cues is that of Dohen and Lœvenbruck (2009), who used normal and whispered speech to analyze the perception of prominence. However, it is difficult to compare our results to theirs, since they did not assess the individual cues of prominence either separately or in their interaction with visual information. Nonetheless, it is acknowledged that prominence perception is enhanced when the perceptual effect both of intonation and of gestures are maximized (e.g., Prieto et al., 2011). Regardless of whether visual cues are predominant or not over the auditory ones, what can be interpreted from our results is that the visual cues of prominence do not take over in the absence of clear auditory cues but are themselves reduced in their prominence-lending properties, which points to a strong integration of visual and auditory modalities, as previously suggested.

4.3 Acoustic cues of prominence

An important observation made in this study is that participants resorted across experiments to whichever acoustic cues they could rely on in the signal in order to detect prominence. In the control experiment, where all acoustic cues were intact, F0, intensity, and duration—but not spectral balance—cued prominence in the audio-only modality. Also, under more adverse acoustic conditions—i.e., with degraded acoustic cues—, this mechanism seems to operate in the same way, as participants perceived prominence by means of the combination of the acoustic cues they had at their disposal.

Concerning the role of duration in prominence perception, our results show a stronger effect size of this cue across experiments when compared to other acoustic cues of prominence. This is in line with previous studies supporting the cross-linguistic role played by duration in producing and perceiving phrasal prominence, mostly in combination with the perceptual effects of at least another correlate (Kohler, 2005; Mo, 2008a,b; Vogel et al., 2016). For example, in a similar study on the perception of phrasal prominence conducted in English—but only in the auditory modality—, Mo (2008a) concluded that duration was the cue that most strongly determined the marks of prominence given by participants, although he found a strong effect of spectral balance as well. Nonetheless, our observation that syllable duration alone allowed to detect prominence in the absence of any other cues seems to be at odds with Mo's conclusion that neither duration nor spectral balance suffice by themselves to make participants detect prominence (Mo, 2008b). Although cross-linguistic differences may be relevant in this sense, the important role of duration finds support in Spanish, where the combination of different cues of prominence indicates that, in the context of unstressed vowels, duration has the lead over F0 in cueing phrasal prominence (Vogel et al., 2016). Several authors have also offered evidence for the combined role of duration and intensity/spectral balance in the production and perception of phrasal prominence (e.g., Kochanski et al., 2005; Mo, 2008a; Silipo and Greenberg, 1999, 2000 for English; e.g., Sluijter and van Heuven, 1996b; Sluijter et al., 1997 for Dutch). Even if pitch accents are generally acknowledged to have a lengthening effect on stressed syllables, the combined potential of both cues to signal phrasal prominence cross-linguistically has been called into question (Beckman and Edwards, 1994; Sluijter and van Heuven, 1996a,b; Ortega-Llebaria and Prieto, 2007). When signaling focused constituents, prosodic lengthening has been found to be correlated with higher F0 (Baumann et al. 2007, for German; Watson et al. 2008, for English; Jun and Lee 1998, for Korean). In Spanish, segmental lengthening has also been observed in syllables carrying nuclear stress, which keep the typical low F0 of declarative sentences (Escandell-Vidal, 2011). This observation, made in cases of *verum focus*, has been associated to the values of impatience and insistence introduced by the repetition of given information (Escandell-Vidal et al., 2014) and is suggestive of the independence of duration from F0 when signaling prominence.

Our results for the control experiment do not provide enough evidence to establish a ranking between the other two acoustic cues—F0 and intensity—as auxiliary cues to duration, since both achieved a very similar effect size in predicting prominence in the audio-only modality of the control experiment. This question is reminiscent of the long-standing debate that confronted advocates of the melodic accent against those defending the role of loudness/articulatory effort (e.g., Sievers, 1901; Stetson, 1928; Navarro Tomás, 1964).

Finally, under normal acoustic conditions, in Exp0, spectral balance did not yield significant results. Spectral balance was calculated here as H1-H2 (e.g., Campbell and Beckman, 1997), but it is possible that a different measure and controlling for other sources of variability such as gender or vowel quality (Iseli et al., 2007) might yield different results for its role in the multimodal perception of acoustic prominence (see Kakouros et al., 2018, for a review). Nonetheless, Spanish speakers—differently from Dutch speakers (Sluijter et al., 1997)—do not seem to rely on spectral balance in the perception of unaccented lexical stress but, rather, are more sensitive to overall intensity and duration (Ortega-Llebaria et al., 2007). For example, Heldner (2003) found that spectral balance—next to overall intensity—was an acoustic correlate of phrasal prominence in Swedish, which he measured as the difference between the overall intensity and the intensity in a signal previously low-pass filtered at 1.5 times the F0 (Heldner et al., 1999).

4.4 Controlled variables

Previous studies have highlighted the role of two individual variables in relation to the perception of prominence, namely, age, gender and the level of musical training (e.g., Alm and Behne, 2015; Hutka et al., 2015). When controlling for such variables, our results showed a gender-based difference when duration served as the only acoustic cue in Exp3 (flat F0 and flat intensity), so that women were more likely than men to give words a mark of prominence. Women have been suggested to be more sensitive than men to visual cues in audiovisual speech perception, possibly with a better performance in the context of degraded speech due to a more efficient audiovisual language processing (e.g., Jaeger et al., 1998). As for the influence of musical training, our results were inconclusive: the highly trained participants were more likely to give words a mark of prominence in the absence of F0 in audio-only modality, but in the absence of intensity the group with <5 years of musical training showed an opposite trend.

4.5 Limitations

The main challenges of this study are those of any online experiment. On the one hand, the duration of the task, determining the number of the speech samples, had to be sufficient to address our research questions without causing such a loss of attention or fatigue in the participants as to affect the validity of the results. The way to address this was to start from a wide corpus in a previous study and then select the most appropriate sentences (Jiménez-Bravo and Marrero-Aguiar, 2020) to be analyzed in the present study. On the other hand, the lack of control over the conditions under which the experiments were carried out was minimized by the introduction of filler sentences in the form of attentional tasks, as well as by the collection of anonymous personal demographic data. At any rate, the results obtained for the subcorpus used here were very similar to those reported for a previously validated larger corpus, where participants had taken the experiment in a sound-proof cabin in the laboratory.

Furthermore, methodological differences between our study and previous studies may explain the relatively low inter-rater agreement results. The gesturing performed by speakers in spontaneous speech samples might not always be as clear-cut cues as the head nods and eyebrow raises—and occasionally also hand beats—that can be naturally elicited in experimental settings. In other words, identifying prominent words can be more challenging if the execution of a hand gesture stretches over several words in the course of a spontaneous utterance, even if the most prominent part of the gesture coincides with the stressed syllable of an acoustically prominent word. In this sense, agreement among participants was similar to that obtained for studies conducted with a more homogeneous set of stimuli, for example as that reported by Mo et al. (2008, $\kappa = 0.38$), where prominence was rated only in the audio-only modality for non-manipulated stimuli from a corpus of spontaneous speech. In other cases inter-agreement proved higher than that obtained with natural speech, such as that reported by Bishop et al. (2020, κ = 0.204–0.208). Taking this into account, we can conclude that participants executed the task successfully.

Finally, we consider that the apparent predominance of the visual over the auditory cues observed in this study should be interpreted with caution, especially since the standard error associated to the effect of some of the analyzed gesture phases was large, which in our case often resulted in very large confidence intervals for their true effect size in the population.

5 Conclusions

Firstly, there exists a complex relationship between visual and auditory information, which varies according to the cues available to the listener. In optimal situations, when all cues are available, the gestural information is particularly synchronized with the tonal information conveyed by the movements of F0, but reduces to a certain extent the effect of the remaining cues, i.e., it neutralizes the effect of intensity and reduces that of duration. Additionally, as the acoustic signal degrades, duration alone suffices for prominence to be identified.

Secondly, the preponderance of the different phases of gestures change according to the conditions in which the stimuli are presented. In optimal conditions, the gesture phase driving the perception of prominence is the stroke, only followed by the apex, whether they are performed with hands, head, or eyebrows separately or in combination. However, when F0 information is missing, the apex no longer drives the participants' perception of prominence, while the absence of both F0 and intensity results in none of the phases of the gesture being associated with a greater number of prominence marks.

Thirdly, under normal acoustic conditions, i.e., in non-manipulated speech, participants were able to use the cues available in the speech signal—F0, intensity and duration, but not spectral balance—to detect prominence. Of these, the most robust proved to be duration, the temporal axis of speech, as observed when compared, under adverse acoustic conditions to other available acoustic cues and visual cues were absent. Studies on other prosodic phenomena, such as rhythm, have shown how linguistic typology can determine which acoustic cues are more relevant to speech perception; for example, in Germanic languages, rhythm may be more determined by differences in intensity, while in Romance languages durational differences are more decisive (e.g., Nolan and Jeon, 1996). In this sense, this study contributes to describe the perception of prominence in languages other than English and other Germanic languages.

Finally, to the best of our knowledge, these results on how an acoustically degraded signal affects the integration of visual and auditory information for prominence perception are a novel contribution that may facilitate a better understanding of speech and language processing in conditions other than in the so-called laboratory speech. In real life, the acoustic signal is subject to multiple sources of degradation, such as noise, masking, etc., and knowing that in those cases temporal differences become the most relevant perceptual cue may not only have theoretical but also applied implications, as it is the case for the improvement of hearing aids and implants, as well as for systems of speech recognition.

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://osf.io/m7tfr/?view_only=238699b07bc4429a9353ccccc8f56afa.

Ethics statement

The studies involving humans were approved by Comité de Ética de la Investigación de la UNED (Universidad Nacional de Educación a Distancia). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

MJ-B: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Supervision, Writing – original draft, Writing – review & editing. VM-A: Conceptualization, Supervision, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

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/fcomm.2024.1287363/full#supplementary-material

References

Akaike, H. (1973). "Information theory as an extension of the maximum likelihood principle," in *Proceedings of the 2nd International Symposium on Information Theory*, eds B. N. Petrov, and F. Csaki (Budapest: Akadéemiai Kiadó), 267–281.

Al Moubayed, S., and Beskow, J. (2009). "Effects of visual prominence cues on speech intelligibility," in *Proceedings of the International Conference on Auditory Visual Speech Processing (AVSP09)* (Norwich), 43–46.

Al Moubayed, S., Beskow, J., and Granström, B. (2010). Auditory visual prominence. *J. Multim. User Interf.* 3, 299–309. doi: 10.1007/s12193-010-0054-0

Alm, M., and Behne, D. (2015). Do gender differences in audio-visual benefit and visual influence in audio-visual speech perception emerge with age? *Front. Psychol.* 6:1014. doi: 10.3389/fpsyg.2015.01014

Ambrazaitis, G., and House, D. (2017). Multimodal prominences: Exploring the patterning and usage of focal pitch accents, head beats and eyebrow beats in Swedish television news readings. *Speech Commun.* 95:100–113. doi: 10.1016/j.specom.2017.08.008

Bates, D., Maechler, 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

Baumann, S., Becker, J., and Mücke, D. (2007). "Tonal and articulatory marking of focus in German," in *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS07)* (Saarbrücken).

Beckman, M. E., and Edwards, J. (1994). "Articulatory evidence for differentiating stress categories," in *Phonological Structure and Phonetic Form: Phonology and Phonetic Evidence*, ed P. A. Keating (Cambridge: Cambridge University Press), 7–33.

Besson, M., Schön, D., Moreno, S., Santos, A., and Magne, C. (2007). Influence of musical expertise and musical training on pitch processing in music and language. *Restor. Neurol. Neurosci.* 25, 399–410.

Biau, E., Torralba, M., Fuentemilla, L., de Diego Balaguer, R., and Soto-Faraco, S. (2015). Speaker's hand gestures modulate speech perception through phase resetting of ongoing neural oscillations. *Cortex* 68, 76–85. doi: 10.1016/j.cortex.2014.11.018

Bishop, J., Kuo, G., and Kim, B. (2020). Phonology, phonetics, and signal-extrinsic factors in the perception of prosodic prominence: evidence from rapid prosody transcription. *J. Phon.* 82:100977. doi: 10.1016/j.wocn.2020.100977

Boersma, P., and Weenink, D. (2023). *Praat: Doing Phonetics by Computer*. Computer software. Version 6.3.14. Available online at: http://www.praat.org/ (accessed August 4, 2023).

Brugman, H., and Russel, A. (2004). "Annotating Multimedia/Multi-modal resources with ELAN," in *Fourth International Conference on Language Resources and Evaluation (LREC - 2004). ELAN (Version 6.7) Computer software. (2023).* Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Available online at: https://archive.mpi.nl/tla/elan (accessed August 4, 2023).

Burnham, K. P., and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: a Practice Information-Theoretic Approach*. Berlin; Heidelberg; New York, NY: Springer-Verlag.

Campbell, N., and Beckman, M. E. (1997). Accent, stress, and spectral tilt. *J. Acoust. Soc. Am.* 101, 3195–3195. doi: 10.1121/1.419208

Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., and Espesser, R. (1996). "About the relationship between eyebrow movements and f0 variations," in *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP96)* (Philadelphia, PA), 2175–2179.

Cheema, G. S., Hakimov, S., Müller-Budack, E., Otto, C., Bateman, J. A., and Ewerth, R. (2023). Understanding image-text relations and news values for multimodal news analysis. *Front. Artif. Intell.* 6:1125533. doi: 10.3389/frai.2023.1125533

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46. doi: 10.1177/001316446002000104

Cole, J., Mahrt, T., and Hualde, J. I. (2014). "Listening for sound, listening for meaning: task effects on prosodic transcription," in *Proceedings of the 7th International Conference on Speech Prosody (SP2014)* (Dublin), 859–863.

Cole, J., Mo, Y., and Baek, S. (2010b). The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech. *Lang. Cogn. Process.* 25, 1141–1177. doi: 10.1080/01690960903525507

Cole, J., Mo, Y., and Hasegawa-Johnson, M. (2010a). Signal-based and expectation-based factors in the perception of prosodic prominence. *Lab. Phonol.* 1, 425–452. doi: 10.1515/labphon.2010.022

Contreras, H. (1964). ¿Tiene el español un acento de intensidad? *Boletín Instituto Filología Universidad Chile* 16, 237–239.

Dohen, M., and Lævenbruck, H. (2009). Interaction of audition and vision for the perception of prosodic contrastive focus. *Lang. Speech* 52, 177–206. doi: 10.1177/0023830909103166

Enríquez, E., Casado, C., and Santos, A. (1989). La percepción del acento en español. *Lingüística Española Actual* 11, 241–269.

Escandell-Vidal, V. (2011). Verum focus y prosodia: cuando la duración (sí que) importa. *Oralia* 14, 181–201. doi: 10.25115/oralia.v14i.8186

Escandell-Vidal, V., Marrero Aguiar, V., and Pérez Ocón, P. (2014). "Prosody, information structure and evaluation," in *Evaluation in Context (Pragmatics and Beyond New Series, 242)*, eds G. Thompson, and L. Alba-Juez (Amsterdam: John Benjamins), 153–178.

Esteve-Gibert, N., and Prieto, P. (2013). Prosodic structure shapes the temporal realization of intonation and manual gesture movements. *J. Speech Lang. Hear. Res.* 56, 850–864. doi: 10.1044/1092-4388(2012/12-0049)

Feenstra, H. E., Vermeulen, I. E., Murre, J. M., and Schagen, S. B. (2017). Online cognition: factors facilitating reliable online neuropsychological test results. *Clin. Neuropsychol.* 31, 59–84. doi: 10.1080/13854046.2016.1190405

Foxton, J. M., Riviere, L.-D., and Barone, P. (2010). Cross-modal facilitation in speech prosody. *Cognition* 115, 71–78. doi: 10.1016/j.cognition.2009.11.009

Granström, B., House, D., and Lundeberg, M. (1999). "Prosodic cues in multimodal speech perception," in *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS99), Vol. 1* (San Francisco, CA), 655–658.

Heldner, M. (2003). On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish. *J. Phon.* 31, 39–62. doi: 10.1016/S0095-4470(02)00071-2

Heldner, M., Strangert, E., and Deschamps, T. (1999). "A focus detector using overall intensity and high frequency emphasis," in *Proceedings of the International Congress of Phonetic Sciences (ICPhS99)* (San Francisco, CA), 1491–1493.

House, D., Beskow, J., and Granström, B. (2001). "Timing and interaction of visual cues for prominence in audiovisual speech perception," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH - 2001* (Aalborg), 387–390.

Hualde, J., Cole, J., Smith, C. L., Eager, C. D., Mahrt, T., and de Souza, R. N. (2016). "The perception of phrasal prominence in English, Spanish and French conversational speech," in *Proceedings of the 8th International Conference on Speech Prosody (SP2016)* (Boston, MA), 459–463.

Hunt, R. J. (1986). Percent agreement, Pearson's correlation, and kappa as measures of inter-examiner reliability. *J. Dent. Res.* 65, 128–130. doi: 10.1177/00220345860650020701

Hutka, S., Bidelman, G. M., and Moreno, S. (2015). Pitch expertise is not created equal: Cross-domain effects of musicianship and tone language experience on neural and behavioural discrimination of speech and music. *Neuropsychologia* 71, 52–63. doi: 10.1016/j.neuropsychologia.2015.03.019

Iseli, M., Shue, Y.-L., and Alwan, A. (2007). Age, sex, and vowel dependencies of acoustic measures related to the voice source. *J. Acoust. Soc. Am.* 121, 2283–2295. doi: 10.1121/1.2697522

Jaeger, J. J., Lockwood, A. H., van Valin, R. D., Kemmerer, D. L., Murphy, B. W., and Wack, D. S. (1998). Sex differences in brain regions activated by grammatical and reading tasks. *Neuroreport* 9, 2803–2807. doi: 10.1097/00001756-199808240-00022

Jannedy, S. and Mendoza-Denton, N. (2005). "Structuring information through gesture and intonation," in *Interdisciplinary Studies on Information Structure*, eds S. Ishihara, M. Schmitz, and A. Schwar (Potsdam: Universitätsverlag Potsdam), 199–244.

Jiménez-Bravo, M., and Marrero-Aguiar, V. (2020). Multimodal perception of prominence in spontaneous speech: a methodological proposal using mixed models and AIC. *Speech Commun.* 124, 28–45. doi: 10.1016/j.specom.2020.07.006

Johnson, F. M., Hicks, L. H., Goldberg, T., and Myslobodsky, M. S. (1988). Sex differences in lipreading. *Bull. Psychon. Soc.* 26, 106–108. doi: 10.3758/BF03334875

Jun, S.-A., and Lee, H.-J. (1998). "Phonetic and phonological markers of contrastive focus in Korean," in *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP 98)* (Sydney, VIC), 1295–1298.

Kakouros, S., Räsänen, O., and Alku, P. (2018). Comparison of spectral tilt measures for sentence prominence in speech. *Speech Commun.* 103, 11–26. doi: 10.1016/j.specom.2018.08.002

Kendon, A. (1972). "Some relationships between body motion and speech: an analysis of an example," in *Dyadic Communication*, eds A. W. Siegman, and B. Pope (New York, NY: Pergamon), 177–210.

Kendon, A. (2004). *Gesture: Visible Action as Utterance.* Cambridge: Cambridge University Press.

Kim, J., Cvejić, E., and Davis, C. (2014). Tracking eyebrows and head gestures associated with spoken prosody. *Speech Commun.* 57, 317–330. doi: 10.1016/j.specom.2013.06.003

Kita, S. (1993). *Language and Thought Interface: A Study of Spontaneous Gestures and Japanese Mimetics* (PhD thesis). University of Chicago, Chicago, IL, United States.

Kochanski, G., Grabe, E., Coleman, J., and Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *J. Acoust. Soc. Am.* 118, 1038–1054. doi: 10.1121/1.1923349

Kohler, K. J. (2005). "Form and function of non-pitch accents," in *AIPUK, Vol. 35a*, eds K. J. Kohler, F. Kleber, P. Benno (Kiel: IPDS), 97–123.

Krahmer, E., Ruttkay, Z., Swerts, M., and Wesselink, W. (2002a). "Pitch, eyebrows and the perception of focus," in *Proceedings of the 1st International Conference on Speech Prosody (SP2002)* (Aix-en-Provence), 443–446.

Krahmer, E., Ruttkay, Z., Swerts, M., and Wesselink, W. (2002b). "Perceptual evaluation of audiovisual cues for prominence," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH - 2002* (Denver, CO), 1933–1936.

Krahmer, E., and Swerts, M. (2007). The effects of visual beats on prosodic prominence: acoustic analyses, auditory perception and visual perception. *J. Mem. Lang.* 57, 396–414. doi: 10.1016/j.jml.2007.06.005

Kushch, O., and Prieto Vives, P. (2016). "The effects of pitch accentuation and beat gestures on information recall in contrastive discourse," in *Proceedings of the 8th International Conference on Speech Prosody (SP2016)* (Boston, MA), 922–925.

Leemann, A., Kolly, M. J., Li, Y., Chan, R. K. W., Kwek, G., and Jespersen, A. (2016). "Towards a typology of prominence perception: the role of duration," in *Proceedings of the 8th International Conference on Speech Prosody (SP2016)* (Boston, MA), 445–449.

Llisterri, J. M. J., de la Mota, C., Riera, M., and Ríos, A. (2003). "The perception of lexical stress in Spanish," in *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS03), 2023-26* (Barcelona).

Loehr, D. P. (2004). *Gesture and Intonation* (PhD thesis). Georgetown: Georgetown University.

Loehr, D. P. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Lab. Phonol.* 3, 71–89. doi: 10.1515/lp-2012-0006

Luchkina, T., Puri, V., Jyothi, P., and Cole, J. (2015). "Prosodic and structural correlates of perceived prominence in Russian and Hindi," in *The Scottish Consortium for ICPhS 2015, editor, Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS15)* (Glasgow: The University of Glasgow), 1–5. Available online at: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0793.pdf (accessed March 18, 2024).

Madsen, S. M. K., Whiteford, K. L., and Oxenham, A. J. (2017). Musicians do not benefit from differences in fundamental frequency when listening to speech in competing speech backgrounds. *Sci. Rep.* 7:12624. doi: 10.1038/s41598-017-12937-9

Mazerolle, M. J. (2023). *AICcmodavg: Model Selection and Multimodel Inference Based on (Q)AIC(c). R Package Version 2.3.2*. Available online at: https://cran.r-project.org/package=AICcmodavg (accessed March 18, 2024).

McClave, E. (1998). Pitch and manual gestures. *J. Psycholinguist. Res.* 27, 69–89. doi: 10.1023/A:1023274823974

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748. doi: 10.1038/264746a0

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal About Thought*. Chicago, IL: University of Chicago Press.

Mo, Y. (2008a). Acoustic correlates of prosodic prominence for naïve listeners of American English. *Annual Meet. Berk. Linguist. Soc.* 34, 257–267. doi: 10.3765/bls.v34i1.3574

Mo, Y. (2008b). "Duration and intensity as perceptual cues for naïve listeners' prominence and boundary perception," in *Proceedings of the 4th International Conference on Speech Prosody (SP2008)* (Campinas), 739–742.

Mo, Y., Cole, J., and Lee, E.-K. (2008). "Naïve listeners' prominence and boundary perception," in *Proceedings of the 4th International Conference on Speech Prosody (SP2008)* (Campinas), 735–738.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., and Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility. *Psychol. Sci.* 15, 133–137. doi: 10.1111/j.0963-7214.2004.01502010.x

Muñoz-Coego, S., Florit-Pons, J., Rohrer, P. L., Vilà-Giménez, I., and Prieto, P. (2022). "The prosodic and gestural marking of the information status of referents in children's narrative speech: a longitudinal study," in *Proceedings of the 8th International Conference on Speech Prosody (SP2022)* (Lisbon), 401–405.

Navarro Tomás, T. (1964). La medida de la intensidad. *Boletín Instituto Filología Universidad Chile* 16, 231–235.

Niebuhr, O. (2009). F0-based rhythm effects on the perception of local syllable prominence. *Phonetica* 66, 95–112. doi: 10.1159/000208933

Nolan, F., and Jeon, H. S. (1996). Speech rhythm: a metaphor? *Philos. Transact. R. Soc. B Biol. Sci.* 369:20130396. doi: 10.1098/rstb.2013.0396

Novack, M. A., and Goldin-Meadow, S. (2017). Gesture as representational action: a paper about function. *Psychon. Bull. Rev.* 24, 652–665. doi: 10.3758/s13423-016-1145-z

Ortega-Llebaria, M. (2006). "Phonetic cues to stress and accent in Spanish," in *Selected Proceedings of the 2nd Conference on Laboratory Approaches to Spanish Phonology*, ed M. Díaz-Campos (Somerville: Cascadilla Press), 104–118.

Ortega-Llebaria, M., and Prieto, P. (2007). "Disentangling stress from accent in Spanish: Production patterns of the stress contrast in deaccented syllables," in *Segment al and Prosodic Issues in Romance Phonology (Current Issues in Linguistic Theory, 282)*, eds P. Prieto, J. Mascaró, and M.-J. Solé (Amsterdam: John Benjamins), 155–176.

Ortega-Llebaria, M., and Prieto, P. (2011). Acoustic correlates of stress in central Catalan and Castilian Spanish. *Lang. Speech* 54, 73–97. doi: 10.1177/0023830910388014

Ortega-Llebaria, M., Prieto, P., and Vanrell, M. (2007). "Perceptual evidence for direct acoustic correlates of stress in Spanish," in *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS07)* (Saarbrücken), 1121–1124.

Pamies, A., Fernández, A. M., Martínez, E., Ortega, A., and Amorós, M. C. (2002). "Umbrales tonales en el espa nol peninsular," in *Actas del II Congreso de Fonética Experimental*, ed Díaz García (Sevilla), 272–278.

Patel, A. D., and Iversen, J. R. (2007). The linguistic benefits of musical abilities. *Trends Cogn. Sci.* 11, 369–372. doi: 10.1016/j.tics.2007.08.003

Pelachaud, C., Badler, N., and Steedman, M. (1996). Generating facial expressions for speech. *Cogn. Sci.* 20, 1–46. doi: 10.1207/s15516709cog2001_1

Powell, M. J. D. (2009). *The BOBYQA Algorithm for Bound Constrained Optimization Without Derivatives*. Cambridge: Department of Applied Mathematics and Theoretical Physics.

Prieto, P., Puglesi, C., Borràs-Comes, J., Arroyo, E., and Blat, J. (2011). "Crossmodal prosodic and gestural contribution to the perception of contrastive focus," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH - 2011* (Florence), 977–980.

Quilis, A. (1971). Caracterización fonética del acento español. *Travaux Linguistique Littérature* 9, 53–72.

R Development Core Team (2023). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Computer program. Version 4.3.0. Available online at: https://www.R-project.org/ (accessed August 31, 2023).

Renwick, M., Shattuck-Hufnagel, S., and Yasinnik, Y. (2004). The timing of speech-accompanying gestures with respect to prosody. *J. Acoust. Soc. Am.* 115, 2397–2397. doi: 10.1121/1.4780717

Rohrer, P. L., Delais-Roussarie, E., and Prieto, P. (2023). Visualizing prosodic structure: Manual gestures as highlighters of prosodic heads and edges in English academic discourses. *Lingua* 293:103583. doi: 10.1016/j.lingua.2023.103583

Sandler, W. (2022). Redefining multimodality. *Front. Commun.* 6:758993. doi: 10.3389/fcomm.2021.758993

Scarborough, R., Keating, P., Mattys, S. L., Cho, T., and Alwan, A. (2009). Optical phonetics and visual perception of lexical and phrasal stress in English. *Lang. Speech* 52, 135–175. doi: 10.1177/0023830909103165

Shattuck-Hufnagel, S., Yasinnik, Y., Veilleux, N., and Renwick, M. (2007). "A method for studying the time alignment of gestures and prosody in American English: 'Hits' and pitch accents in academic-lecture-style speech," in *Fundamentals of Verbal and Nonverbal Communication and the Biometric Issue, NATO Publishing Sub-Series E: Human and Societal Dynamics, Vol. 18*, eds A. Esposito, M. Bratanić, E. Keller, and M. Marinaro (Brussels: IOS Press), 1079–1098.

Sievers, E. (1901). *Grundzüge der Phonetik. Bibliothek indogermanischer Grammatiken 1*. Leipzig: Breitkopf und Härtel.

Silipo, R., and Greenberg, S. (1999). "Automatic transcription of prosodic stress for spontaneous English discourse," in *Proceedings of 14th International Congress of Phonetic Sciences (ICPhS99)* (San Francisco, CA), 2351–2354.

Silipo, R., and Greenberg, S. (2000). "Prosodic stress revisited: reassessing the role of fundamental frequency," in *Proceedings of the NIST Speech Transcription Workshop* (College Park, MD).

Sluijter, A. M. C., and van Heuven, V. J. (1996a). Spectral balance as an acoustic correlate of linguistic stress. *J. Acoust. Soc. Am.* 100, 2471–2485. doi: 10.1121/1.417955

Sluijter, A. M. C., and van Heuven, V. J. (1996b). "Acoustic correlates of linguistic stress and accent in Dutch and American English," in *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP96)* (Philadelphia, PA), 630–633.

Sluijter, A. M. C., van Heuven, V. J., and Pacilly, J. J. A. (1997). Spectral balance as a cue in the perception of linguistic stress. *J. Acoust. Soc. Am.* 101, 503–513. doi: 10.1121/1.417994

Smith, C. and Edmunds, P. (2013). "Native English listeners' perceptions of prosody in L1 and L2 reading," in *Proceedings of the Annual Conference of the International Speech Communication Association, 1006 INTERSPEECH—2013* (Lyon, France), 235–238.

Solé, M. J. (1984). Experimentos sobre la percepción del acento. *Estudios Fonética Exp.* 1, 134–243.

Stetson, R. H. (1928). *Motor Phonetics*. La Haye: Martinus Nijhoff.

Stöckl, H., and Pflaeging, J. (2022). Multimodal coherence revisited: notes on the move from theory to data in annotating print advertisements. *Front. Commun.* 7:900994. doi: 10.3389/fcomm.2022.900994

Strand, J., Cooperman, A., Rowe, J., and Simenstad, A. (2014). Individual differences in susceptibility to the mcgurk effect: links with lipreading and detecting audiovisual incongruity. *J. Speech Lang. Hear. Res.* 57, 2322–2331. doi: 10.1044/2014_JSLHR-H-14-0059

Streefkerk, B. M., Pols, L. C. W., and ten Bosch, L. F. M. (1997). "Prominence in read-aloud sentences, as marked by listeners and classified automatically," in *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* (Amsterdam), 101–116.

Sturgeon, B. A., Hubbard, R. J., Schmidt, S. A., and Loucks, T. M. (2015). High f0 and musicianship make a difference: pitch-shift responses across the vocal range. *J. Phonet*. 51, 70–81. doi: 10.1016/j.wocn.2014.12.001

Swerts, M. (1997). Prosodic features at discourse boundaries of different strength. *J. Acoust. Soc. Am*. 101, 514–521. doi: 10.1121/1.418114

Swerts, M., and Krahmer, E. (2004). "Congruent and incongruent audiovisual cues to prominence," in *Proceedings of the 2nd International Conference on Speech Prosody (SP2004)* (Nara), 69–72.

Swerts, M., and Krahmer, E. (2005). Audiovisual prosody and feeling of knowing. *J. Mem. Lang*. 53, 81–94. doi: 10.1016/j.jml.2005.02.003

Swerts, M., and Krahmer, E. (2008). Facial expression and prosodic prominence: effects of modality and facial area. *J. Phon*. 36, 219–238. doi: 10.1016/j.wocn.2007.05.001

Swerts, M., and Krahmer, E. (2010). Visual prosody of newsreaders: effects of information structure, emotional content and intended audience on facial expressions. *J. Phon*. 38, 197–206. doi: 10.1016/j.wocn.2009.10.002

't Hart, J. (1981). Differential sensitivity to pitch distance, particularly in speech. *J. Acoust. Soc. Am*. 69, 811–821. doi: 10.1121/1.385592

Terken, J., and Hermes, D. (2000). "The perception of prosodic prominence," in *Prosody: Theory and Experiment. Studies Presented to Gösta Bruce*, ed M. Horne (Dordrecht: Kluwer Academic Publishers), 89–127.

Thompson, W. F., Schellenberg, E. G., and Husain, G. (2004). Decoding speech prosody: do music lessons help? *Emotion* 4, 46–64. doi: 10.1037/1528-3542.4.1.46

Turk, A., and Sawusch, J. (1996). The processing of duration and intensity cues to prominence. *J. Acoust. Soc. Am*. 99, 3782–3790. doi: 10.1121/1.414995

Vogel, I., Athanasopoulou, A., and Pincus, N. (2016). "Prominence, contrast and the functional load hypothesis: an acoustic investigation," in *Dimensions of Phonological Stress*, eds J. Heinz, R. Goedemans, and H. van der Hulst (Cambridge: Cambridge University Press), 123–167.

Wagner, P., Origlia, A., Avesani, C., Christodoulides, G., Cutugno, F., D'Imperio, M., et al. (2015). "Different parts of the same elephant: a roadmap to disentangle and connect different perspectives on prosodic prominence," in *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS15)* (Glasgow: University of Glasgow).

Watson, D. G., Arnold, J. E., and Tanenhaus, M. K. (2008). Tic Tac TOE: EFFECTS of predictability and importance on acoustic prominence in language production. *Cognition* 106, 1548–1557. doi: 10.1016/j.cognition.2007.06.009