The Role of the Auditory and Visual Modalities in the Perceptual Identification of Brazilian Portuguese Statements and Echo Questions

Luma Miranda, Marc Swerts, João Antônio de Moraes, Albert Rilliard

▶ **To cite this version:**

Luma Miranda, Marc Swerts, João Antônio de Moraes, Albert Rilliard. The Role of the Auditory and Visual Modalities in the Perceptual Identification of Brazilian Portuguese Statements and Echo Questions. Language and Speech, 2021, 64 (1), pp.3-23. 10.1177/0023830919898886 . hal-02456308

HAL Id: hal-02456308
https://hal.science/hal-02456308

Submitted on 27 Aug 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

This submitted paper has been processed and then published in the journal Language and Speech: https://doi.org/10.1177%2F0023830919898886

The role of the auditory and visual modalities in the perceptual identification of Brazilian Portuguese statements and echo questions

Journal:	*Language and Speech*
Manuscript ID	LAS-19-0075.R2
Manuscript Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Miranda, Luma; Universidade Federal do Rio de Janeiro, Acoustic Phonetics Laboratory Swerts, Marc; Tilburg University, Communication and Cognition de Moraes, João Antônio; Universidade Federal do Rio de Janeiro, Acoustic Phonetics Laboratory Rilliard, Albert; Laboratoire d'Informatique pour la Mecanique et les Sciences de l'Ingenieur, Acoustic Phonetics Laboratory; Universidade Federal do Rio de Janeiro
Keywords:	Audiovisual perception, Prosody, Statement, Echo question, Brazilian Portuguese
Abstract:	This paper presents the results of three perceptual experiments investigating the role of auditory and visual channels for the identification of statements and echo questions in Brazilian Portuguese. Ten Brazilian speakers (five male) were video-recorded (frontal view of the face) while they produced a sentence ("Como você sabe"), either as a statement (meaning "As you know.") or as an echo question (meaning "As you know?"). Experiments were set up with the two different intonation contours. Stimuli were presented in conditions with clear and degraded audio as well as congruent and incongruent information from both channels. Results show that Brazilian listeners were able to distinguish statements and questions prosodically and visually, with auditory performances being dominant over visual ones. In noisy conditions, the visual channel improved the interpretation of prosodic cues robustly, while it degraded them in conditions where the visual information is incongruent with the auditory information. This study confirms the previous findings on auditory and visual integration within speech perception, also when applied to prosodic patterns.

Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.
Fig. 3-Assertion production of a female speaker.mp4 Fig. 3-Echo question production of a female speaker.mp4 Fig. 4-Echo question production of a male speaker.mp4 Assertion with SNR of 0dB produced by a female speaker.mp4 Echo question with SNR of 0dB produced by a female speaker.mp4 Assertion with SNR of -6dB produced by a female speaker.mp4 Echo question with SNR of -6dB produced by a female speaker.mp4

SCHOLARONE™
Manuscripts

The role of the auditory and visual modalities in the perceptual identification of Brazilian Portuguese statements and echo questions

Luma Miranda[1], Marc Swerts[2], João A. Moraes[1], Albert Rilliard[3,1]

[1] Laboratory of Acoustic Phonetics, Federal University of Rio de Janeiro, Brazil

[2]Department of Communication and Information Sciences, Tilburg University, The Netherlands

[3] LIMSI, CNRS, Université Paris Saclay, France

For Peer Review

Abstract

This paper presents the results of three perceptual experiments investigating the role of auditory and visual channels for the identification of statements and echo questions in Brazilian Portuguese. Ten Brazilian speakers (five male) were video-recorded (frontal view of the face) while they produced a sentence ("*Como você sabe*"), either as a statement (meaning "*As you know*.") or as an echo question (meaning "*As you know*?"). Experiments were set up with the two different intonation contours. Stimuli were presented in conditions with clear and degraded audio as well as congruent and incongruent information from both channels. Results show that Brazilian listeners were able to distinguish statements and questions prosodically and visually, with auditory performances being dominant over visual ones. In noisy conditions, the visual channel improved the interpretation of prosodic cues robustly, while it degraded them in conditions where the visual information is incongruent with the auditory information. This study confirms the previous findings on auditory and visual integration within speech perception, also when applied to prosodic patterns.

Keywords: Audiovisual perception. Prosody. Statement. Echo question. Brazilian Portuguese.

Introduction

When people process auditory information, like incoming speech, they tend to also rely on visual input, such as facial expressions a speaker produces while talking. That the visual and auditory channels can have a strong impact on speech perception has been shown in the well-known McGurk effect (McGurk & McDonald, 1976): that is, the presentation to participants of incongruent auditory and facial cues about specific segments tends to lead to an illusionary percept that integrates cues from the two modalities. Visual speech (lipreading) has also been studied as a source of information to help listeners with cochlear implants comprehend speech, while more recent studies have demonstrated that the visual channel is beneficial for listeners with normal hearing as well (Rosenblum, 2005).

Previous studies in this area of research focused on the recognition of individual speech sounds, but there is a growing awareness that auditory and visual cues may also interact at higher levels. For instance, the speaker's face has been documented as a reliable indicator of emotions and attitudes in speech (Barkhuysen, Krahmer and Swerts, 2010; Crespo Sendra, Kaland, Swerts and Prieto, 2013; Ekman, Friesen and Hager, 2002; Moraes, Miranda and Rilliard, 2012). Studies have also shown that the visual channel can serve as a marker of communicatively relevant information that is often associated with prosodic features. The visual channel may also signal prominence (Hadar, Steiner, Grant, and Clifford Rose, 1983; Krahmer & Swerts, 2007; Swerts & Krahmer, 2008), that is the marking of some words as being more salient than others in an utterance, as well as signal

For Peer Review

the focus of a sentence (Dohen & Loevenbruck, 2009; Krahmer, Ruttkay, Swerts and Wesselink, 2002). In this way, visual cues may signal the information status of sentence elements, for example, to distinguish broad from contrastive focus (e.g., see Lambrecht, 1994, for a discussion). Visual cues are also used for the identification of statements versus questions (Borràs-Comes & Prieto, 2011; Cruz, Swerts and Frota, 2017; House, 2002; Nicholson, Baum, Kilgour, Koh, Munhall and Cuddy, 2003; Srinivasan & Massaro, 2003). In other words, given the multimodal nature of speech production, listeners perceive linguistic functions through the combination of acoustic and visual cues, both at segmental and suprasegmental levels of speech (Gili Fivela, 2018).

In this paper, the bimodal perception of statements and echo questions in Brazilian Portuguese (henceforth, BP) is investigated. The choice of these two sentence types is motivated by the fact that (i) in BP the same syntactic structure is used for both sentence modes, which makes it a function prominently carried out by prosodic cues and (ii) few studies have analyzed the prosodic production of echo questions. This paper is based on insights from previous studies that have investigated the distinction between statements and questions in different languages. These studies are consistent in claiming that, although the auditory signal tends to be a more informative resource to determine the pragmatic status of an utterance, listeners also extract information from the visual channel to identify sentence types.

A number of studies experimentally investigated audiovisual processing of sentence types. House (2002) analysed the visual recognition of questions in Swedish, using an animated talking head whose visual cues were systematically manipulated to convey statements and questions in an interaction with humans. The experiments showed that listeners relied more on intonational than on visual cues for the identification of questions. For American English, Srinivasan and Massaro (2003) ran a set of experiments with both

For Peer Review

natural and virtual speakers based on one sentence produced as a statement or as an echo question; the lexico-syntactic structure of the sentence was kept constant in order to make sure that listeners did not rely on possible syntactic or semantic cues. Similarly to the Swedish study, the authors concluded that the intonation was more informative than the facial expressions; yet, in the final remarks of their work, they suggested that a larger *corpus* with natural speakers should be used to obtain additional auditory and visual cues that were absent in their experiments.

The integration of the auditory and visual cues in speech is also supported by experiments measuring reaction times. In European Portuguese, Cruz et al. (2017) compared the production and perception of statements and yes-no questions. The results showed that reaction times of the identification task were lower in the audiovisual condition compared to the monomodal conditions, including the auditory condition.

Borràs-Comes and Prieto (2011) analysed the audiovisual recognition of echo questions and statements with contrastive focus using a McGurk style paradigm (presenting incongruent auditory and visual cues) and collecting reaction times from Catalan listeners. Identification was faster when participants were exposed to congruent stimuli, supporting the bimodal integration of speech perception. The study also concluded that the visual channel was predominant over intonational cues for Catalan listeners' decisions. In their discussion, the authors stress the importance of the integration of visual and intonational cues for an accurate and rapid interpretation of these prosodic meanings.

The exceptionally high level of recognition of visual cues in Borràs-Comes and Prieto's (2011) study may be linked to the fact that the authors selected stimuli that "best characterized the contrast" (Borràs-Comes & Prieto, 2011, p. 362) between the sentence types upon which they focused. While this result provides insight into the integration of

For Peer Review

audiovisual cues for sentence type identification, it may not really be representative as it makes use of very prototypical and possibly exaggerated expressions that may not generalize to more natural data. It would thus be useful to investigate to what extent similar findings could be obtained with recordings of more natural and spontaneous occurrences of the facial gestures accompanying the production of echo questions and statements with contrastive focus.

Most of the previously mentioned research reveals that visual cues have some effect on sentence type classification, but that their contribution is relatively small when compared to auditory cues. This raises the question of how the effect would play out for a language as BP, which is peculiar in that it does not mark yes-no questions by syntactic means (e.g., word order), but only by intonation. Moraes (1998) showed that the perceptually pertinent opposition is made by means of intonation, with a fundamental frequency (F0) change on the final pre-stressed and stressed syllables of a sentence nucleus used to mark assertive (F0 rise on the pre-stressed syllable followed by a fall on the stressed one) and interrogative (F0 fall on the pre-stressed syllable followed by a rise on the stressed one) meanings. Moraes (2008) described nuclear pitch accents for these two modes, following the Autosegmental-Metric notation (Pierrehumbert, 1980), as H+L*L% for statements and L+<H*L% for yes-no questions.

As mentioned before, instead of neutral yes-no questions, which are information-seeking questions, that is, the speaker asks for an information unknown to him (Frota et al., 2015), we decided to analyze echo questions, because this type of yes-no question is underexplored in BP. According to Frota et al. (2015), the pragmatic function of echo-questions is twofold: to trigger a yes-no answer, just like the neutral yes-no question, and to signal a lack of comprehension of a previous utterance. When speakers produce echo questions, they repeat the previous utterance in order to verify if it was correctly

For Peer Review

understood. In European and Brazilian Portuguese varieties, a similar nuclear rising F0 for both types of questions was found (Frota et al., 2015). So, given the attested use of intonation patterns for distinguishing statements and echo questions in BP, it remains to be seen whether, in addition to these prosodic cues, and to what extent BP speakers may rely on facial cues for such an identification task.

In a pilot project, Peres, Raposo de Medeiros, Ferreira Netto and Baia (2011) analysed the role of the visual channel for the identification of statements and yes-no questions in BP. Their first experiment utilized a McGurk set-up whereby they created stimuli by combining the audio from one source with the video from another source. A second experiment proposed a video-only condition in which listeners had to assess muted versions of recorded clips. The first experiment indicated a dominance of intonational information over visual information to identify statements and questions, while the second showed that yes-no questions and statements could be recognized by Brazilian participants through the visual channel only. Still, the number of participants of this study was limited to ten listeners, for all the experiments. Hence, the authors' conclusions about the contribution of each channel for the perception of statements and questions in BP may be considered preliminary.

The current study extends the analysis of the prosodic and visual cues that have been proposed to be valuable for the identification between statements and questions in BP. In comparison with prior work, we increased the number of speakers, participants and presentation conditions of audiovisual stimuli in order to obtain a better understanding of the multimodal perception process involved in such a task. As it was our aim to compare the relative capacities of prosodic and visual cues in conveying this distinction, we choose to keep the lexical and syntactic material constant, despite a possible bias, or lack of variability, this may introduce. Note that we decided to include

For Peer Review

stimuli from multiple speakers to make sure that the perceptual results had general and ecological validity, as opposed to previous studies that were based on evaluations of a schematic or highly stylized animated talking head with specific auditory and visual parameter settings. Additionally, we semi-randomly selected speakers' productions from their ten realizations, without explicitly choosing their "best performances".

The phonetic analysis of this study served to describe relevant prosodic cues that can be also used to identify the statements and echo questions in the initial and final regions of the intonational contours regarding not only the F0, but also the intensity and duration parameters. A visual analysis of the recorded data was conducted in order to identify and categorize prototypical gestures that were related to the production of each sentence type in BP and which listeners could recognize in perceptual tasks.

In addition, considering that the intonational cues for these contours in BP are known to be highly intelligible, the role of the visual channel was also explored through the identification tasks of statements and questions using degraded audio conditions. As information from the face is known to improve speech intelligibility in noise (Benoît & Le Goff, 1998; Ouni, Cohen, Ishak and Massaro, 2006; Sumby & Pollack, 1954), the acoustic signal of the recorded data was degraded with two levels of Signal-to-Noise Ratio (SNR), using a noise based on a multitalker babbling speech recording.

The main goal of this paper is to understand the contribution of the speaker's face in the perceptual integration of auditory and visual channels for the identification of statements and echo questions in Brazilian Portuguese. This study aims to investigate the following research questions:

(i) To which extent the visual cues–alone or combined with auditory cues–allow for the correct identification of statements and questions? On the basis of the literature on BP, we may hypothesize a higher recognition

For Peer Review

rate in the bimodal presentation condition as well as a lower identification ratio for visual cues than for auditory ones, but still above chance.

(ii) What is the extent of the influence of the visual channel when the auditory information is impoverished? Based on prior studies, we expect a higher contribution of the visual channel in this condition.

(iii) Which channel will lead the subjects' interpretation when the voice and the face are signaling different information? Here again, a dominant role of auditory cues is foreseen.

The article is divided into the following sections: we first describe how we obtained audiovisual recordings and how we analysed the prosodic and visual properties of these data. Then, the designs of Experiments 1 and 2 are explained followed by a combined statistical analysis of the results of both tests. Next, the setup of Experiment 3 is presented along with the statistical analysis, combining the results of Experiments 1 and 3. Finally, the outcomes of the three experiments are discussed and followed by a general conclusion.

For Peer Review

Methods

Structure of the *corpus* and participants

We invited ten native speakers of BP (five male) to produce different versions of the sentence "*Como você sabe*", both as an assertion (with the meaning "*As you know.*") or as an echo question (with the meaning "*As you know?*", which can also be rephrased as "Is it '*as I know*' that you are asking me?"). Note that the sentence does not contain any lexical or morphosyntactic indices related to the mode; studies so far (Frota et al., 2015;

Moraes, 1998, 2008) have shown that this pragmatic distinction can be carried out by intonation-only cues.

Speakers having a formal training in language sciences were recruited, to maximize the chances that they would understand the communication situations linked to each sentence mode. Although it may induce interpretative behaviors that differ from their natural ones, we think that the fact of recording ten speakers would neutralize most of these drawbacks. All speakers were graduate and undergraduate students at the Federal University of Rio de Janeiro (UFRJ, henceforth), except for one speaker who is a professor at UFRJ and one of the authors. All participants are speakers of Rio de Janeiro variety of BP. They were aged between 19 and 65 years old at the time of the recording session.

Recording procedure

Speakers were asked to read and sign a consent form in which the audiovisual recording process and the future use of the data were explained. Speakers were seated against a dark background in a sound-attenuated room at the Acoustic Phonetics Laboratory of UFRJ. None of them wore glasses. They were recorded using a SONY NEX-F3 video camera, placed 90 cm from the speakers. The camera was positioned to frame the upper body and face. To increase the quality of the acoustic signal, a Zoom H4 recorder was used to jointly record each voice, with a microphone located 20 cm from the speaker's mouth, outside the frame of the camera. Although the speakers are not actors, they were told to express themselves as naturally as possible. The experimenter was in the room during the whole recording session, instructing the speaker about the sentence types. Each subject repeated the sentence ten times, alternating statements and echo questions. This process resulted in 200 recorded utterances. After the recording sessions,

For Peer Review

the audio waves were synchronized with the video in the software Vegas Pro 14.0 (Magix, 2014), using a handclap produced at the beginning of each recording session.

Acoustic analysis

The 200 recorded sentences were manually segmented at the phonemic level using the Praat software (Boersma & Weenink, 2016). Prosodic correlates of fundamental frequency (F0, expressed in semitones (ST) relative to 1 Hz, using Praat's default algorithm), intensity (expressed in dB) and syllabic duration (in seconds) were estimated from the acoustic signal. Some pre-processing of the raw measurements was made: a Praat script produced time normalized F0 contours (Arantes, 2015) that allowed the comparison of F0 prominences between speakers and modes. The same script was adapted to time normalize the intensity measures. Another Praat script was used for the duration analysis by transforming the raw phonemic duration values in a smoothed z-score values that removed intrinsic differences in phoneme duration (Barbosa, 2013).

The results of these measurements indicate that the assertive and interrogative contours bear distinct F0 movements. These distinguishing patterns are quite consistently expressed across our recordings: to implement the declarative contour, all speakers produce a rising F0 movement from the beginning of the contour until the nuclear pre-stressed syllable ("cê"), followed by a fall on the stressed one ("sa"), whereas the echo question is characterized by a low F0 from the prenuclear region of the contour until the pre-stressed syllable ("cê"), followed by a rise on the stressed ("sa") and a fall on the post-stressed syllable ("be"), as illustrated in Fig. 1.

[Insert Figure 1]

The patterns presented in Fig. 1 are consistent with the descriptions in the literature on BP intonation (Frota et al., 2015; Miranda, 2015; Moraes, 2008). Moreover, it can be

For Peer Review

noticed that the final pre-stressed syllable ("cê") in the declarative contour is high, whilst the same syllable is low in the interrogative contour, marking the contrast between the final pre-stressed syllables ("cê") in both contours. This behavior of the final pre-stressed syllable ("cê") is corroborated by Couto, Silva and Miranda (2017) who analysed the declaratives and interrogative contours in three BP varieties–Rio de Janeiro, Fortaleza and Salvador. This difference between the nuclear region of the contours supports Moraes's descriptions (1998, 2008) that can lead in representing the pitch accents of statements as H+L* L% and those of questions as L+<H*L%, as already stated in the introduction of this study.

The acoustic analysis also shows that the intonational contours of both the statement and the echo question are accompanied by specific intensity patterns and similar duration patterns (cf. Fig. 2).

[Insert Figure 2]

Regarding intensity, the main differences between statements and echo questions are found around the tonic syllable ("sa"), with (i) a stronger pre-tonic syllable ("cê") and (ii) a weaker post-tonic syllable ("be") for the assertive contour, as compared to the interrogative one. The tonic syllable ("sa") is the stronger and the longer in both cases, showing comparable values across modes. In addition, based on the bottom panel of Fig. 2, it is possible to observe an increasing syllabic duration along sentences for both modes, up to the tonic syllable ("sa") of the nucleus.

The analysis of the prosodic parameters shows that there are clear prosodic features discriminating statements from echo questions in BP, most clearly in terms of their F0 contours.

Facial movements analysis

For Peer Review

Analyses of facial movements during the production of statements and questions in different languages generally reveal a recurrent set of visual cues. For interrogatives, the literature shows that the head and the eyebrows movements are commonly employed (Cavé, Guaïtella, Bertrand, Santi, Harlay and Espesser, 1996; House, 2002) and provide reliable cues when manipulated in an animated talking head (Granström, House and Swerts, 2002; House, 2002). For instance, eyebrow raising and head tilt were the visual cues found in English echo questions by Srinivasan and Massaro (2003), while eyebrow lowering and vertical head tilting conveyed yes-no questions in the Swedish study conducted by House (2002), using an animated talking head.

In multimodal communication studies, various body movements have been considered as well (hand gestures, head movements, speakers' gaze, etc.). Torreira and Valtersson (2015) investigated conversational data in French to check whether polar questions and continuation statements, which are similar in morphosyntactic and intonational forms, could be discriminated by prosodic and visual cues. They concluded that the distinction of the sentence types is based on gaze towards the addressee during questions, but eyebrow raising, head movements and other interactive manual gesture types were also identified.

In the analysis of our video data, movements on the speakers' faces were described using the FACS annotation system (Ekman et al., 2002) which has been shown to be able to taxonomize and label facial muscular movements made by the speakers. The following eighteen Action Units (AU, henceforth) were selected to describe the facial movements: inner brow raiser (AU 1), outer brow raiser (AU 2), brow lowerer (AU 4), upper lid raiser (AU 5), cheek raiser and lid compressor (AU 6), lid tightener (AU 7), lip corner depressor (AU 15), chin raiser (AU 17), blink (AU 45), head turn left (AU 51), head turn right (AU 52), head up (AU 53), head down (AU 54), head tilt left (AU 55), head tilt right (AU 56),

For Peer Review

head forward (AU 57), head back (AU 58) and, finally, up and down head movement (AU 85). Only the presence of AUs was recorded in this study; intensity was not measured.

Two of the authors independently analysed forty videos, corresponding to two repetitions of statements and echo questions produced by ten speakers, and annotated the facial movements in terms of AUs. The selected stimuli correspond to those used in the perceptual experiments described hereafter. A moderate agreement between the two annotators was reached, as measured by Cohen's Kappa ($\kappa = 0.488$). Although significantly above 0 (i.e., no agreement above chance level), this kappa is described as "moderate" by Landis and Koch (1977). Thus, to maintain the possibility of accounting for possible divergences between the two annotators, the AU analysis was made keeping the annotations of both authors separated. A correspondence analysis was run on the count of AUs observed by speech act, speaker and annotator. The analysis extracts the main dimensions that explain attribution of AUs to the stimuli. The ten first dimensions, which explain about 85% of the variance, were selected to run an algorithm of hierarchical clustering which was used to group together stimuli that have similar characteristics (Husson, Lê, Pagès, 2017). The cluster analysis proposed a three-cluster solution.

Cluster I is composed of stimuli showing AU 4 (brow lowerer) and AU 7 (lid tightener) and expressing echo question. Cluster II contains AU 1 (inner brow raiser), AU 2 (outer brow raiser) and AU 55 (head tilt left) associated with the production of assertions. An example of these two facial expressions can be seen in Fig. 3.

[Insert Figure 3]

Cluster III links AU 17 (chin raiser) and AU 57 (head forward) to the production of echo question by speaker 7. These last two AUs may be considered part of this speaker's idiosyncrasy and are illustrated in Fig. 4.

For Peer Review

[Insert Figure 4]

These three clusters are essentially related to the two speech acts under investigation, plus an idiosyncratic behavior for echo questions by a particularly visually expressive speaker. Note that some of the gestural features produced in co-occurrence with echo questions seem to share similarities with the recurrent "shrug" gestures described in Debras (2017). In addition, differences in labelers' perception do not introduce a major bias at this broad level of clustering.

This analysis of the facial movements produced by Brazilian speakers uttering statements and echo questions shows that Brazilian speakers in our dataset convey the meaning of statements and questions with different and recurrent facial gestures. Therefore, in principle, these speech acts may thus be possibly identified not only by prosodic cues (especially regarding F0 patterns), but also by visual cues. The video recordings of Fig. 3 and 4 can be accessed in the supplementary materials of this paper.

The findings of our visual analysis are partially in line with the visual description of statements and yes-no questions in European Portuguese (EP) provided by Cruz et al. (2015). EP speakers produce statements with an up-down head movement, whereas yes-no questions are produced with either up-down or back-forward head movements plus eyebrow raising. Thus, the head movement is argued to be a visual cue produced by EP speakers for both statements and yes-no questions. In our data, BP speakers also produced head movement with the assertion meaning, just like in EP. On the other hand, the eyebrow raise was observed in the production of BP assertions, instead of questions, as described in EP. In BP speakers' production of echo questions, the eyebrows are typically lowered and the eyes tightened.

Therefore, our results show that eyebrow lowering can be related to question marker in BP, as opposed to the eyebrow raise described by Cruz et al. (2015) for EP. It is worth

For Peer Review

to note that this difference in the facial performance can be related to the speech acts explored in both studies, since echo-questions were analyzed in BP and neutral yes-no questions in EP. Although an echo-question is also considered a type of yes-no question (Frota et al., 2015), as stated in the introduction of this study, the visual production of echo-questions may express a component of doubt or uncertainty, which was shown in Fig. 4 with the facial expression of cluster III.

Summarizing the prosodic and visual descriptions, brow lowering along with tightening of the eyes is associated with the echo questions that present a rising nuclear pitch accent and, in terms of intensity prominence, a weaker final pre-stressed syllable ("cê") as well as a stronger final post-stressed syllable ("be"). The eyebrow raising plus head movement is related to the assertion contour with a falling pitch accent and a stronger final pre-stressed syllable ("cê") plus a weaker final post-stressed syllable ("be"). The final stressed syllable ("sa") is the strongest in relation to intensity and the longest in both contours.

For Peer Review

Perceptual experiments

In this section, the described prosodic and visual cues related respectively to statements and echo questions are perceptually validated. The design of Experiments 1 and 2 are presented separately and the results of both are commented together. Afterwards, the setup of Experiment 3 is exposed along with the outcome that was also combined with Experiment 1.

Experiment 1: Audiovisual recognition with clear audio

The first experiment was designed to verify the relative importance of visual cues in comparison to the auditory ones as signals of the statement and echo question distinction in Brazilian Portuguese.

Participants

Experiment 1 was applied to 64 participants (21 in the audiovisual, 22 in the video-only and 21 in the audio-only condition) with mean age of 29,1 years old. Listeners originated from various regions of Brazil, though predominantly from Rio de Janeiro. Although the majority of participants were either students or professors in Linguistics at UFRJ, some of them were working or studying in different scientific fields (e.g., Philosophy and Engineering). They participated in the experiment through a web interface, using either their personal computer or one provided by the Acoustic Phonetics Lab at UFRJ. For each presentation condition, a new group of listeners was recruited. None reported hearing or sight impairments. Use of headset was required for the auditory conditions implying audio presentations.

Stimuli

The eighth and ninth repetitions of the assertion and echo question produced by all speakers were selected. This way, we gained a randomization of the video material. Unlike Borràs-Comes and Prieto (2011), we did not select the visual data qualitatively. In other words, we did not select the most expressive productions from each speaker to avoid potentially over-acted data. The video clips were cut in the software Vegas Pro so that their length was approximately two seconds. It is worth noting that the production of the whole utterance contained in the clip was often preceded or followed by a short silent pause. In total, forty utterances were selected for the perceptual experiments.

For Peer Review

Procedure

We used the Qualtrics platform to run this perceptual experiment. Stimuli were presented one by one, in a randomized order as generated by the software for each subject. Subjects were tested individually. They had to identify the sentence mode between the two alternatives (forced choice), by pressing a "statement" or "question" button. Participants were told that, depending on the experimental condition, they would see, hear, or see and hear speakers uttering either a statement or a question and had to judge the intended meaning of the utterance "*Como você sabe*".

Experiment 1 presented forty utterances in three conditions: audio-only, video-only and audiovisual, with each condition presented to a different group of participants. This experiment had a between-subject design to avoid possible learning effects that could come into play if participants were presented with all experimental conditions. Each group of participants in the three different conditions were presented with forty trials. The remaining number of trials was always indicated at the top of the screen. A typical run lasted approximately fifteen minutes.

For Peer Review

Experiment 2: Audiovisual recognition with degraded audio

The goal of the second experiment was to analyze the robustness of the audiovisual presentations. Thus, the role of the prosodic and visual cues in the recognition of statements and echo questions in Brazilian Portuguese was assessed in degraded audio conditions.

Participants

Experiment 2 involves 43 participants (21 in the audiovisual and 22 in the audio-only condition) with mean age of 24 years old. Differently from the previous experiment, listeners were only from Rio de Janeiro. Participants used a computer of the Acoustic Phonetics Lab at UFRJ to complete the perceptual task. They were either students or professors of Linguistics at UFRJ. Since listeners were not allowed to participate in multiple experiments, for each presentation condition, a new group of listeners was recruited. None of the participants reported hearing or sight impairments. Again, the use of headset was required for the auditory conditions of the experiment.

Stimuli

The video material of Experiment 2 is the same used in Experiment 1, except that the auditory channel is degraded by mixing it with a multi-speaker babble noise taken from Steeneken & Geurtsen (1988). Note that the speakers were not recorded in noisy conditions. In this study, we are interested in noise for its capacity to degrade the given prosodic cues, so to measure the relative power of visual cues. Recording the speakers in noise would probably change their communication strategies as a consequence of the Lombard effects, leading to an increased F0, duration and energy as well as bigger motions of the face and the head (Fitzpatrick, Kim and Davis, 2015; Kim, Sironic and Davis, 2011), which is not our aim here.

Regarding the addition of noise in the recording material, Cooke (2006) recommends a fixed signal-to-noise ratio (SNR, henceforth) as being more effective when multi-speaker babble noise rather than a single competing talker is employed. Furthermore, the expected limits of SNR for individuals with normal hearing suggested by a study with Brazilian listeners (Costa, Daniel and dos Santos, 2011) showed a range of the SNR that goes from –3,03 dB to –7,55 dB with a mean of –5,29 dB. Based on the latter study and

For Peer Review

after a pilot experiment using SNR of +3 dB, 0 dB and –6 dB, we selected the last two levels of SNR (0 dB and –6 dB) , and applied the continuous multi-speaker babble noise to the original utterance: a favorable SNR of 0 dB and an adverse SNR of –6 dB.

The degraded audio was produced in Praat and later combined again within the original videos in Vegas Pro software. To keep the task feasible, only one repetition from each speaker was used, thus presenting forty stimuli with degraded audio to listeners. In the supplementary materials of this paper, four samples of the audiovisual stimuli with degraded audio can be accessed (two statements and echo questions with SNR of 0 dB and –6 dB produced by one speaker). It is worth remembering that the segmental material did not change in these experiments: listeners had to decode prosody, a different type of information than the segmental one studied in literature using such masking noise.

Procedure

The second perceptual experiment followed the same procedure as described in Experiment 1. The only difference is that Experiment 2 presented forty stimuli with two SNR levels (0 dB and –6 dB) in only two conditions: audiovisual and audio-only conditions, each condition presented to a different group of participants as well. This experiment also had a between-subject design, avoiding potential learning effects for the same reason presented in Experiment 1. Each group of subjects was presented with forty trials and, just like the previous experiment, the remaining number of trials was always indicated at the top of the screen. Participants took fifteen minutes to complete the perceptual task as well.

Results of Experiments 1 and 2

For Peer Review

Results of Experiment 2 were analyzed along with Experiment 1. Subjects' responses were expressed as *success* or *failure* to identify the presented mode (assertive or interrogative). The proportion of success was expressed for each of the seven presentation conditions in Experiments 1 and 2 (audio-only/AO; video-only/VI; audiovisual/AV; audio-only in 0 dB SNR/AO-0; audio-only in –6 dB SNR/AO-6; audiovisual in 0 dB SNR/AV-0; audiovisual in –6 dB SNR/AV-6—see Fig. 5 that shows this ratio in each condition as well as each type of sentence, assertive or interrogative). The ratio of success was used as the dependent variable for a logistic regression model (quasibinomial errors were used to deal with overdispersion in binomial model residuals), using two categorical explanatory factors (*Condition* and *Type*, and their interaction) as predictors.

The logistic regression shows that the interaction between both factors is not significant ($F_{(6,290)} = 1.6$, $p = 0.15$), while both *Condition* and *Type* factors do have a significant impact on the accuracy of judging the sentence mode (Type: $F_{(1,290)} = 9.28$, $p < 0.01$; Condition: $F_{(6,290)} = 10.14$, $p < 0.01$). A *post-hoc* Tukey test was applied to the *Condition* factor so to test for differences between levels of the two factors, showing that three presentation conditions (video-only/VI, audio-only in 0 dB/AO-0 and audio-only in –6 dB/AO-6) do significantly degrade the listener's identification performances compared to the best presentation condition: audio-only (AO). A comparison of participants' performance in each experimental condition of Experiments 1 and 2 is shown in Fig. 5:

[Insert Fig. 5]

The results show a group of presentation conditions containing the AO, AV, AV-0 and AV-6 conditions with homogeneous and high identification capabilities. The audio presentation conveys identification information about sentence mode: the AO condition reaches a high level of correct answers (90% of correct answers, with a confidence interval that nears a perfect score)–thus there is a strong suspicion of ceiling effect in this

For Peer Review

case. The fact that the AV, AV-0 and AV-6 conditions did not significantly differ from the AO condition also advocates for a more general "ceiling effect" for all these conditions: they shall thus all be regarded as presenting comparably high recognition scores. Bimodal presentations (AV) achieve a comparable performance level, with non-significant variations (that thus may be related to, e.g., subjects sampling). The bimodal presentations with degraded audio (AV-0, AV-6) also reach a level of performance comparable to the audio-only and audiovisual conditions. On the contrary, conditions based on visual information alone (VI) and degraded audio alone (AO-0, AO-6) reach a significantly lower identification level. This result supports the view of a cross-modal fusion of information to the interpretation of sentence mode.

Regarding the factor *Type*, assertions presented an identification accuracy ratio about six percent higher than echo questions, which was a significant improvement. The fact that there is no interaction between the *Type* and *Condition* factors indicates a general tendency of subjects to prefer the “assertion” answer in any presentation condition. One may speculate that assertion functions as a default answer when no clear or interpretable information is available for participants.

For Peer Review

Experiment 3: Congruent and incongruent modalities

The third experiment presented congruent and incongruent audio and visual modalities to evaluate which channel exerted a dominant influence on listeners’ identification of sentence type.

Participants

Experiment 3 was applied to 24 participants with mean age of 29,4 years old. Once again, listeners were either students or professors of Linguistics from UFRJ. They

participated in the experiment through a web interface in the computer of the Acoustic Phonetics Lab at UFRJ. The participants do not report any hearing or sight impairments from the participants. The use of headset was also required for the audiovisual presentation of this experiment.

Stimuli

The set of 40 videos (2 repetitions x 2 sentence types x 10 speakers) was used to mix modalities from different stimuli. Using the Vegas Pro software, the audio utterances were matched with video materials from other recordings. The congruent videos were based on the audio of the first repetition synchronized with the video of the second repetition of the same mode; that is, they consisted of both an audio and video channels that cued the same sentence type. Incongruent videos are based on the audio of the first repetition of one mode matched with the video from the second repetition of the other mode, so that there was a mismatch between what the audio and video signals would cue. We decided to construct mixed stimuli for both the congruent and incongruent cases in order to generate a total set of data that were artificially constructed. Using this procedure, we obtained forty artificial videos, twenty with congruent auditory and visual streams, twenty with incongruent streams.

Procedure

The same procedure used in Experiments 1 and 2 was applied in the third experiment, but here all participants would hear and see the videoclips. Experiment 3 presented forty stimuli, containing either congruent or incongruent auditory and visual information (twenty stimuli for each of these two types). The congruent audiovisual condition (AVc) is in principle similar to the audiovisual presentation (AV) of Experiment 1, except that

For Peer Review

the audio and visual channels were taken from two different repetitions of the same speaker for the same targeted sentence mode. The incongruent condition (AVi) consisted of combinations of audio and visual modalities that were retracted from distinct sentence modes (e.g., the auditory information from an echo question paired with the visual information from a statement). The forty stimuli were presented in audiovisual condition to one group of 24 participants (within-subjects design). There were forty trials in this experiment and, just like the previous perceptual tests, the remaining number of trials was also indicated at the top of the screen. A typical run also lasted fifteen minutes.

Results

Results of Experiment 3 are analysed in relation to the outcome of Experiment 1. Answers were expressed as *success* or *failure* to correctly identify the assertive or interrogative mode of the acoustic part of the stimuli (because in incongruent presentations, the visual information signals the alternative category). Results for the two conditions of Experiment 3 are presented in Fig. 6. The proportion of success was calculated for the five presentation conditions (audiovisual/AV, audio-only/AO and video-only/VI from Experiment 1; congruent audiovisual/AVc and incongruent audiovisual/AVi from Experiment 3) and each sentence type (statement or question). Variations in the ratio of success according to *Condition* and *Type* of sentence factors were again analysed by a logistic regression (with quasibinomial errors).

The logistic regression shows that the interaction between the two factors is not significant ($F_{(4,218)} = 0.83$, $p = 0.51$), while the two main factors have a significant impact on ratio of success (Type: $F_{(1,226)} = 8.08$, $p < 0.01$; Condition: $F_{(6,290)} = 8.85$, $p < 0.01$). A *post-hoc* Tukey test was applied to *Condition* factor and shows that the video-only (VI)

For Peer Review

and incongruent audiovisual (AVi) levels do significantly degrade the listener's identification performances, compared to the audio-only (AO) presentations.

As shown in Fig. 6, stimuli in congruent audiovisual presentations were better recognized than in audiovisual incongruent ones, which caused the identification rate of the sentence types to decrease.

[Insert Figure 6]

Based on Experiment 3, we can conclude that the visual channel interferes in the process of listeners' identification of statements and questions, although the auditory component is dominant for the perceptual recognition of these sentence types. This finding means that the speech perception of the intonation of sentence types is bimodal even when the auditory channel is clear.

General discussion

The main goal of this study was to analyze the contribution of the visual channel, in combination with the auditory one, as a possible resource to identify statements and echo questions in BP. To this end, we made use of one carrier sentence produced repetitively by ten speakers of the Rio de Janeiro variety of BP. We chose to use only one sentence because our primary aim is to validate the relative role of the visual modality, controlling for other potential variables (e.g., variation in articulation induced by different sequences of phonemes). We did record various speakers with variable productions to make sure that our findings have general validity; note also that these speakers produced prosodic patterns that have been reported in the literature to be representative of the sentence modes we explored.

For Peer Review

First, the recorded data was described both acoustically and visually. The outcome of the prosodic and visual descriptions brought to light relevant features that could function as cues for this identification in perceptual tasks. The most relevant prosodic parameters to distinguish both modes are the F0 contours, as described in the literature, combined with intensity patterns. The most salient visual cues for questions were the brow lowerer (AU 4) and lid tightener (AU 7), while chin raiser (AU 17) and head forward (AU 54) were more consistently found for one of the speakers, which may be considered an idiosyncratic performance or something unrelated to the distinction of modes under investigation here. For statements, the eyebrow raising movements (AU 1 and AU 2) and head movements (AU 55) turned out to be characteristic for this sentence type. Different eyebrow and head movements seem to visually distinguish statements from echo questions in BP.

Next, we were interested in analyzing the perceptual identification of these two sentence types to evaluate the relative strength of the described auditory and visual cues when clear or impoverished audio signal were presented, as well as when both modalities signaled conflicting information.

Experiment 1 investigated the capacity of participants to identify statements and echo questions in clear audio conditions, in the three presentation conditions: audio-only (AO), video-only (VI) and audiovisual (AV). A ceiling effect was observed in the results, that implies identification performances in audiovisual condition (AV) that could not outperform the audio-only condition (AO) in which the outcome was already quasi perfect. Comparably high-performance levels were found for all the conditions containing clear or degraded audio cues and coherent visual ones, as for clear audio cues only. It is worth mentioning that this type of effect obtained in audio-only (AO) and audiovisual (AV) presentation conditions of Experiment 1 is in line with previous studies that looked

For Peer Review

into the way discourse items are signaled. For instance, as the contrastive focus in French (Dohen & Loevenbruck, 2009) is accurately marked by prosodic cues, the visual cues in bimodal presentations do not have an extra beneficial effect and so could not increase performances that are already at a ceiling level.

One could hypothesize that the visual benefit in the interpretation of the intonation of sentence types is higher when the analyzed intonational contours are similar (i.e., more difficult to identify) compared to contours with clearly different F0 configurations, as in our study. For instance, in Borràs-Comes and Prieto (2011), the difference between a contrastive focus statement and an echo question in Catalan is given by pitch range, since both contours have the same F0 contour (rising-falling nuclear configuration). In the audiovisual experiments, the authors concluded that the visual cues had a stronger effect than the audio cues in the perceptual identification of both utterance types. Also in Catalan, Tubau, González-Fuente, Prieto and Spinal (2015) found that the identification of confirmation and contradicting yes-answers, both with falling F0 patterns, was higher in the audiovisual modality than in the audio-only modality. Similarly, in Miranda, Moraes and Rilliard (2019), the distinction between wh-questions and wh-exclamations in BP is characterized by the inclination of the falling pitch accent in the nuclear region of both contours, which is steeper in wh-questions than in wh-exclamations. Results from the perceptual identification test showed an improved recognition in the audiovisual modality compared to monomodal ones (video-only and audio-only). All these studies demonstrated that facial gestures enhanced the interpretation of the intonation of the sentence types, without limitation due to a "ceiling effect".

In sum, results of Experiment 1 lend support to the conclusions of previous studies that highlight that the same functions are performed by prosody and facial gestures, although visual cues alone are less efficient than the prosodic ones (House, 2002; Srinivasan &

For Peer Review

Massaro, 2003). Nonetheless, the speaker's face can still be considered a reliable source of prosodic information at the sentence level. For instance, in Fisher's study (1969) regarding the "visibility of final pitch" variation in English, listeners had to indicate in a perceptual experiment the direction of terminal pitch contour in the production of questions with either a rising or a falling contour, while watching the visual performances of speakers without sound. Results showed that listeners were able to identify the direction of the intonational contours based on the visual cues. Our findings also support the view that participants are able to identify the intonational contours of different utterance types by relying only on the speaker's face. Fisher attributes the listeners' visual discrimination of final pitch contours to certain motor activities of the face such as the movement of the lips. Still, he also suggests in the final remarks of the paper that the whole face could be a source of information for the listener.

In Experiment 2, participants again had to identify statements and echo questions but this time with noisy stimuli. Noisy audio-only presentations (audio-only in 0 dB SNR/AO-0 and audio-only in –6 dB SNR/AO-6) show a significant decrease in identification performances compared to the audio-only (AO) ones, but there still exist some cues to identify mode, with performances comparable to video-only (VI) presentations. The auditory (AO-0, AO-6) and visual (VI) channels share their respective cues in the audiovisual in 0 dB SNR (AV-0) and audiovisual in –6 dB SNR (AV-6) conditions to attain identification levels comparable to the ideal situation of a clear audio (or audiovisual) presentation. The visual enhancement of the identification of intonational contours in adverse auditory conditions found in our study supports findings of earlier work on visual speech. For instance, Benoît & Le Goff (1998) and Nicholson et al. (2003) state that there is a lower impact of the visual channel on speech perception compared to a clear acoustic signal, but the visual speech can increase identification scores when the

For Peer Review

auditory channel is impoverished. Therefore, our findings confirm that speech intelligibility in a noisy environment is improved when participants have access to the speaker's face.

The results of Experiment 2 also showed that babble noise can efficiently degrade prosodic cues to sentence mode (and probably to other prosodic functions), opening a possible type of evaluation for the contribution of multimodal cues to linguistic meaning (cf. Dohen & Lovenbruck's, 2009, for an opposite prediction). However, since the speakers of our study were not recorded in noisy conditions, their productions do not fit those that would have been produced in a noisy environment. In such a situation, a Lombard speech effect would probably have occurred. It has been found that in speakers' recordings in adverse auditory conditions (Fitzpatrick et al., 2015; Kim et al., 2011), there are differences regarding not only the prosody, such as increased energy, F0, duration and flatter spectral tilt, but also the facial gestures, like bigger motions of the face and head. These are different from the more neutral expressions we elicited in our set of data. Meanwhile, it is not our goal here to evaluate the capacity of a speaker to communicate in adverse conditions, but rather to measure the relative role of both modalities.

From a perception point of view, Vatikiotis-Bateson, Barbosa, Chow, Oberg, Tan and Yehia (2007) verified that both the auditory intelligibility and the visual enhancement are greater when speech is recorded in noise compared to speech recorded in quiet and then added noise to it. However, in the same study, when stimuli were presented to listeners in an "easy listening level", the authors found results in which the recognition of words in the condition of speech recorded in quiet with addition of noise was similar to the condition of stimuli recorded in noise. Vatikiotis-Bateson et al. (2007) state that, depending on the used listening levels of SNR, results can be substantially altered.

For Peer Review

Although the results of Experiment 2 cannot be used to evaluate the degradation of prosodic cues in noisy conditions, the findings remain interesting to show the role of visual cues to enhance the audio decoding process in adverse auditory conditions from the listeners' point of view. Apart from these caveats, Experiment 2 confirms that listeners also rely on the speaker's face to perceive what type of sentence is being uttered even when the speech recorded in quiet has noise added to it, giving support to the multimodal perception of the intonation of sentence types.

The parallel processing of both modalities was shown in Experiment 3, as incongruent presentations degraded the performance of a clear acoustic presentation, introducing significant interferences. Congruent stimuli show accuracy ratio levels comparable to the original stimuli (audio-only/AO and audiovisual/AV), showing that the manipulation does not induce incoherencies. Incongruent presentations of auditory and visual cues show that the visual channel has an impact in the listeners judgment, even if auditory cues are still dominant for the perceptual recognition of sentence mode.

The outcome of Experiment 3 confirms earlier findings regarding the process of information fusion from auditory and visual modalities to form a judgment over pragmatic meanings of sentence types such as statements and echo questions. More specifically, the results are in line with previous research using artificially constructed stimuli in BP (Peres et al., 2011), as well as in other languages, such as Dutch (Swerts & Krahmer, 2008), Catalan (Borràs-Comes & Prieto, 2011) and European Portuguese (Cruz et al., 2017).

Based on these three perceptual experiments, results support previous audiovisual studies that conclude that the auditory channel can convey all the necessary information about sentence mode, while visual channel alone cannot. Yet, at the same time, the results also show that visual cues are decoded by the participants even in the presence of clear

For Peer Review

acoustic presentations, as the visual information degrades their decoding when incoherent. When stimuli are presented in a noisy context, the results reveal that visual cues also add robustness in the speech decoding process.

Therefore, these results are in line with Massaro and Cohen's assumptions (1983), regarding not only the integration of auditory and visual cues as a fundamental process of speech comprehension, but also regarding their viewpoint that we can expect a greater contribution of the visual channel when either the verbal information is ambiguous or the audio information is degraded.

Conclusion

In summary, this article shows that the auditory and visual channels are integrated in the perception of sentence modes in BP. Brazilian listeners identify statements and questions using both auditory and visual cues. Listeners rely more on the auditory cues than the visual ones when presented under clear acoustic conditions. Contrary to our expectations, the recognition rate of the sentences mode in the bimodal condition was not higher than in the audio-only condition. Thus, the first hypothesis written in the introduction section of this study was partially confirmed. However, the visual channel has a beneficial effect when stimuli consists of degraded auditory cues, which confirms the second hypothesis. Finally, although a dominant role of the auditory cues was verified in the experiment with congruent and incongruent audio and visual modalities, the auditory and visual channels are integrated in the perceptual identification process of sentence types, confirming the third hypothesis as well.

We intend to further explore other types of sentences in additional experiments, such as the continuation statement, which is also an underexplored sentence type in BP. For instance, Miranda et al. (2019) analysed the bimodal perception of BP wh-questions and

For Peer Review

wh-exclamations. The results of the perceptual identification test confirm the integration of the auditory and visual channels in signaling wh-questions and wh-exclamations meanings in BP. These positive results suggest a further investigation of the bimodal perception of the intonational contours of other sentence types.

Moreover, it may be possible to go further in the understanding of audiovisual integration by monitoring a participant's reaction time, a measure that was not included in these experimental settings. As shown in the introduction of this paper, previous studies (Borràs-Comes & Prieto, 2011; Cruz et al., 2017) already indicated that listeners' identification of sentence types is faster in audiovisual presentation conditions. Experiments analyzing participant's reaction time to audiovisual recognition of sentence types in Brazilian Portuguese are needed.

For Peer Review

Acknowledgments

We wish to thank Dr. Katy Carlson and the three anonymous reviewers for their thoughtful comments on an earlier version of the manuscript. This research has been funded by the scholarship 88881.134778/2016-01 (awarded by the Brazilian Federal Agency for Support and Evaluation of Graduate Education–CAPES) through a PhD exchange program which allowed the first author to organize a research visit to Tilburg University (NL).

References

Arantes, P. (2015). Time-normalization of fundamental frequency contours: A hands-on tutorial. In: Meireles, A. (Org.). *Courses on Speech Prosody* (pp. 98–123). 1ed. Newcastle upon Tyne, UK: Cambridge Scholars Publishing.

Barbosa, P. (2013). Semi-automatic and automatic tools for generating prosodic descriptors for prosody research. *Proceedings of Interspeech satellite event Tools and Resources for the Analysis of Speech Prosody*, Aix-en-Provence, France, pp. 86–89.

Barkhuysen, P., Krahmer, E. & Swerts, M. (2010). Cross-modal and incremental perception of audiovisual cues to emotional speech. *Language and Speech*, 53 (1), 3–30.

Benoît, C. & Le Goff, B. (1998). Audiovisual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP. *Speech Communication*, 26, 117–129.

Boersma, P. & Weenink, D. (2016). *Praat*: *doing phonetics by computer*. Computer program (Version 5.1.05) available at: http://www.praat.org/.

Borràs-Comes, J. & Prieto, P. (2011). Seeing tunes. The role of visual gestures in tune interpretation. *Journal of Laboratory Phonology*, 2 (2), 355–380.

Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F. & Espesser, R. (1996). About the relationship between eyebrow movements and F0 variations. *Proceedings of Fourth International Conference on Spoken Language Processing*, Philadelphia, Pennsylvania, USA. DOI: 10.1109/ICSLP.1996.607235

Cooke, M. (2006). A glimpsing model of speech perception in noise. *Journal of Acoustical Society of America*, 119 (3), 1562–1573.

For Peer Review

Costa, M. J., Daniel, R. C. & dos Santos, S. N. (2011). Reconhecimento de sentenças no silêncio e no ruído em fones auriculares: valores de referência de normalidade. *CEFAG*, 13 (4), 685–691.

Couto, L. R., Silva, C. G. & Miranda, L. S. (2017). Prosódia dos enunciados declarativos e interrogativos totais nas variedades de Salvador, Fortaleza e Rio de Janeiro. *Revista de Estudos da Linguagem*, 25 (3), 1105–1142. DOI: http://dx.doi.org/10.17851/2237-2083.25.3.1105-1142.

Crespo Sendra, V., Kaland, C., Swerts, M. & Prieto, P. (2013). Perceiving incredulity: the role of intonation and facial gestures. *Journal of Pragmatics*, 47, 1–13.

Cruz, M., Swerts, M. & Frota, S. (2017). The role of intonation and visual cues in the perception of sentence types: evidence from European Portuguese varieties. *Laboratory Phonology*, 8 (1), 23. DOI: http://doi.org/10.5334/labphon.110.

Cruz, M., Swerts, M. & Frota, S. (2015). Variation in tone and gesture within language. *Proceedings of Eighteenth International Congress of Phonetic Sciences*, Glasgow, UK. Retrieved from: http://www.icphs2015.info/pdfs/ Papers/ICPHS0452.pdf.

Debras, C. (2017). The shrug: forms and meanings of a compound enactment. *Gesture*, 16 (1), 1-34. DOI: https://doi.org/10.1075/gest.16.1.01deb

Dohen, M. & Loevenbruck, H. (2009). Interaction of audition and vision for the perception of prosodic contrastive focus. *Language and Speech*, 52 (2–3), 177–206.

Ekman, P., Friesen, W. & Hager, J. (2002). *The Facial Action Coding System*. 2nd ed (CD-ROM). Salt Lake City, Utah, USA: Research Nexus.

For Peer Review

Fisher, C. G. (1969). The visibility of terminal pitch contour. *Journal of Speech and Hearing Research*, 12 (2), 379–382.

Fitzpatrick, M., Kim, J. and Davis, C. (2015). The effect of seeing the interlocutor on auditory and visual speech production in noise. *Speech Communication*, 74, 37–51.

Frota, S., Cruz, M., Svartman, F., Collischonn, G., Fonseca, A., Serra, C.,... & Vigário, M. (2015). Intonational variation in Portuguese: European and Brazilian varieties. In: Frota, S.; Prieto, P. (Org.). *Intonation in Romance* (pp. 235–283). 1ed. Oxford, UK: Oxford University Press.

Gili Fivela, B. (2018). Multimodal analyses of audio-visual information: Some methods and issues in prosody research. In: Feldhausen, I., Fliessbach, J. & Vanrell, M. M. (Eds.). *Methods in Prosody*: *A Romance Language Perspective (Studies in Laboratory Phonology 4)* (pp. 83–122). Berlin, Germany: Language Science Press.

Graf, H. P., Cosatto, E., Volker, S. & Huang, F. J. (2002). Visual prosody: facial movements accompanying speech. *Proceedings of Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, Washington, D.C., USA. DOI: 10.1109/AFGR.2002.1004186

Granström, B., House, D. & Swerts, M. (2002). Multimodal feedback cues in human-machine interaction. *Proceedings of First International Conference on Speech Prosody*, Aix-en-Provence, France, pp. 347–350.

Hadar, U., Steiner, T. J., Grant, E. C. & Clifford Rose, F. (1983). Head movement correlates of juncture and stress at sentence level. *Language and Speech*, 26, 117–129.

Hirst, D. & Espesser, R. (1993). Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix*, 15, 75–85.

For Peer Review

House, D. (2002). Intonational and visual cues in the perception of interrogative mode in Swedish. *Proceedings of Seventh International Conference on Spoken Language Processing*, Denver, Colorado, USA, pp. 1957–1960.

Husson, F., Lê, S., & Pagès, J. (2017). *Exploratory multivariate analysis by example using R*. Chapman and Hall/CRC.

Kim, J., Sironic, A. & Davis, C. (2011). Hearing speech in noise: seeing a loud talker is better. *Perception*, 40 (7), 853–862.

Krahmer, E., Ruttkay, Z., Swerts, M. & Wesselink, W. (2002). Pitch, eyebrows and the perception of focus. *Proceedings of First International Conference on Speech Prosody*, Aix-en-Provence, France, pp. 443–446.

Krahmer, E. & Swerts, M. (2007). The effects of visual beats on prosodic prominence: acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, 57 (3), 396–414.

Lambrecht, K. (1994). *Information structure and sentence form*. Cambridge: Cambridge University Press.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33 (1), 159–174.

Massaro, D. & Cohen, M. (1983). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology Human Perception & Performance*, 41 (5), 751–775.

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746–748. Retrieved from: https://www.nature.com/articles/264746a0.

For Peer Review

Miranda, L. S. (2015). *Análise da entoação do português do Brasil segundo o modelo IPO*. (Master thesis) – Faculdade de Letras da Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2015. Retrieved from: http://www.letras.ufrj.br/posverna/mestrado/MirandaLS.pdf.

Miranda, L. S., Moraes, J. A. & Rilliard, A. (2019). Audiovisual perception of wh-questions and wh-exclamations in Brazilian Portuguese. *Proceedings of Nineteenth International Congress of Phonetic Sciences*, Melbourne, Australia, pp. 2941–2945.

Moraes, J. A. (1998). Intonation in Brazilian Portuguese. In: Hirst, D. & Di Cristo, A. (Eds.) *Intonation Systems: A Survey of Twenty Languages* (pp. 179–194). Cambridge, UK: Cambridge University Press.

Moraes, J. A. (2008). The pitch accents in Brazilian Portuguese: analysis by synthesis. *Proceedings of Fourth Conference on Speech Prosody*, Campinas, Brazil, pp. 389–397.

Moraes, J. A., Miranda, L. S. & Rilliard, A. (2012). Facial gestures in the expression of prosodic attitudes in Brazilian Portuguese. *Proceedings of Seventh GSCP International Conference Speech and Corpora*, Belo Horizonte, Brazil, pp. 157–161.

Nicholson, K. G., Baum, S., Kilgour, A., Koh, C. K., Munhall, K. G. & Cuddy, L. L. (2003). Impaired processing of prosodic and musical patterns after right hemisphere damage. *Brain and Cognition*, 52, 382–389.

Ouni, S., Cohen, M., Ishak, H. & Massaro, D. (2006). Visual contribution to speech perception: Measuring the intelligibility of animated talking heads. *EURASIP J. AUDIO SPEECH MUSIC PROC.*, (2006) 2007: 047891. DOI: https://doi.org/10.1155/2007/47891.

For Peer Review

Peres, D. O., Raposo de Medeiros, B., Ferreira Netto, W. & Baia, M. F. A. (2011). The role of the visual stimuli in the perception of prosody in Brazilian Portuguese. *Proceedings of Fifth Conference on Laboratory Approaches to Romance Phonology*, Somerville, Massachusetts, USA, pp. 136–141.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation*. Ph.D. dissertation, MIT.

Qualtrics software, Version 2017 of Qualtrics. Copyright © [2019] *Qualtrics*. Online platform available at: https://www.qualtrics.com/.

Rosenblum, L. D. (2005). Primacy of Multimodal Speech Perception. In: Pisoni, D. B. & Remez, R. E. (Eds.). *Handbook of Speech Perception* (pp. 52–78). Oxford, UK: Blackwell Publishing.

Vatikiotis-Bateson, E., Barbosa, A. V., Chow, C. Y., Oberg, M., Tan, J. and Yehia, H. C. (2007). Audiovisual Lombard speech: reconciling production and perception. *Proceedings of International Conference on Auditory-Visual Speech Processing*, Hilvarenbeek, The Netherlands, pp. 45–50.

Vegas Pro software, Version 14 of Vegas Pro. Copyright © [2016] *MAGIX*. Software available at: https://www.vegascreativesoftware.com/.

Srinivasan, R. & Massaro, D. (2003). Perceiving prosody from the face and voice: Distinguishing statements from echoic questions in English. *Language and Speech*, 46 (1), 1–22.

Steeneken, H. & Geurtsen, F. W. M. (1988). *Description of the RSG-10 noise database*. TNO Institute for Perception, report IZF, 1988–3.

For Peer Review

Sumby, W. & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of Acoustical Society of America*, 26, 212–215.

Swerts, M. & Krahmer, E. (2008). Facial expressions and prosodic prominence: effects of modality and facial area. *Journal of Phonetics*, 36 (2), 219–238.

Torreira, F. & Valtersson, E. (2015). Phonetic and visual cues to questionhood in French. *Phonetica*, 72, 20–42.

Tubau, S., González-Fuente, S., Prieto, P. and Espinal, M. A. (2015). Prosody and gesture in the interpretation of yes-answers to negative yes/no-questions. *The Linguistic Review*, 32(1), 115–142.

For Peer Review

Figure 1: *Plot of the mean F0 values (in ST) of normalized time over each modality: assertion (above) and echo question (below).*

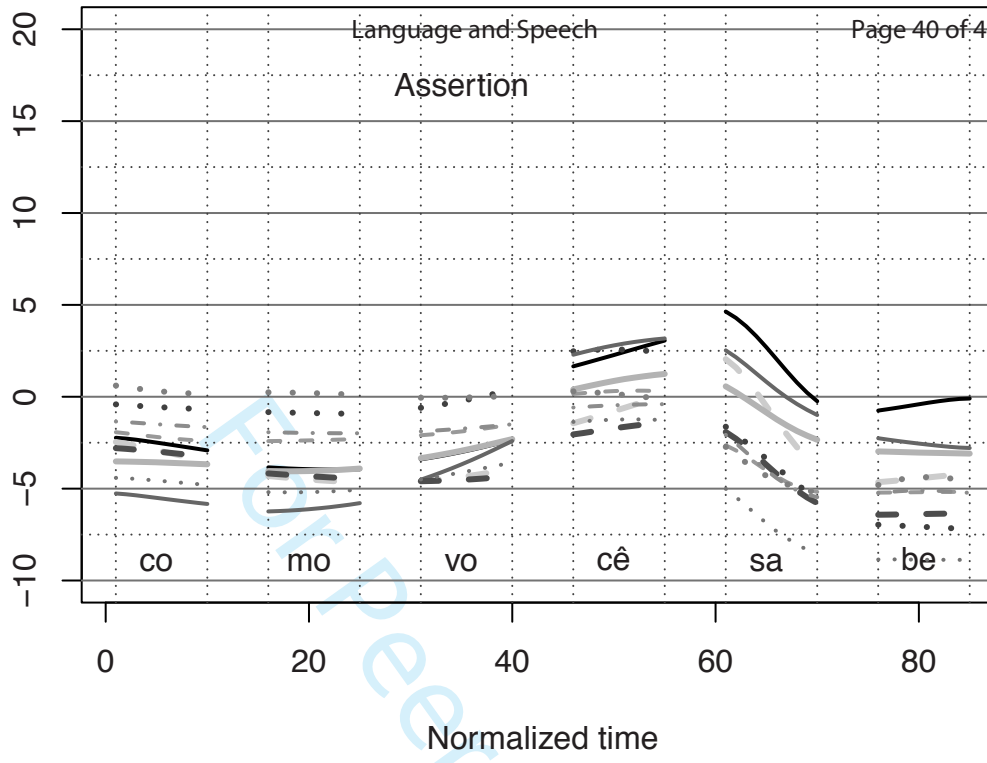
Figure 2: *Plot of the mean intensity (in dB, top panel) and duration (in z-score, bottom panel) over each modality: assertion (left) and echo question (right).*

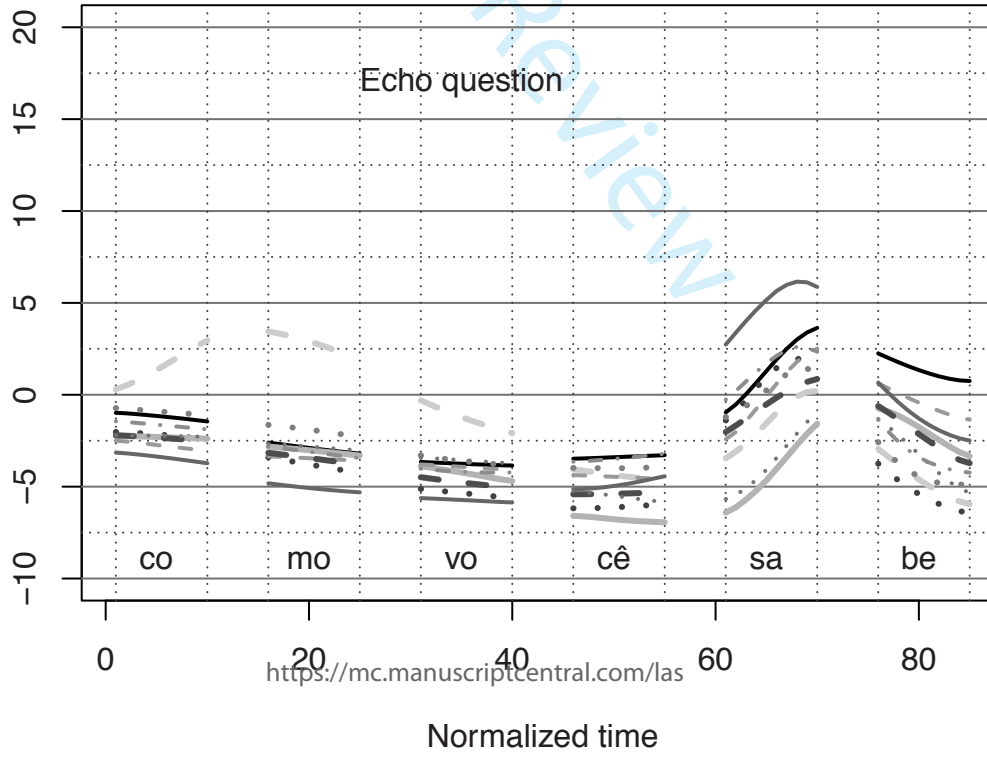
Figure 3*: Stills from a female speaker producing an assertion (above) and an echo question (below).*

Figure 4: *Stills from a male speaker producing an echo question.*

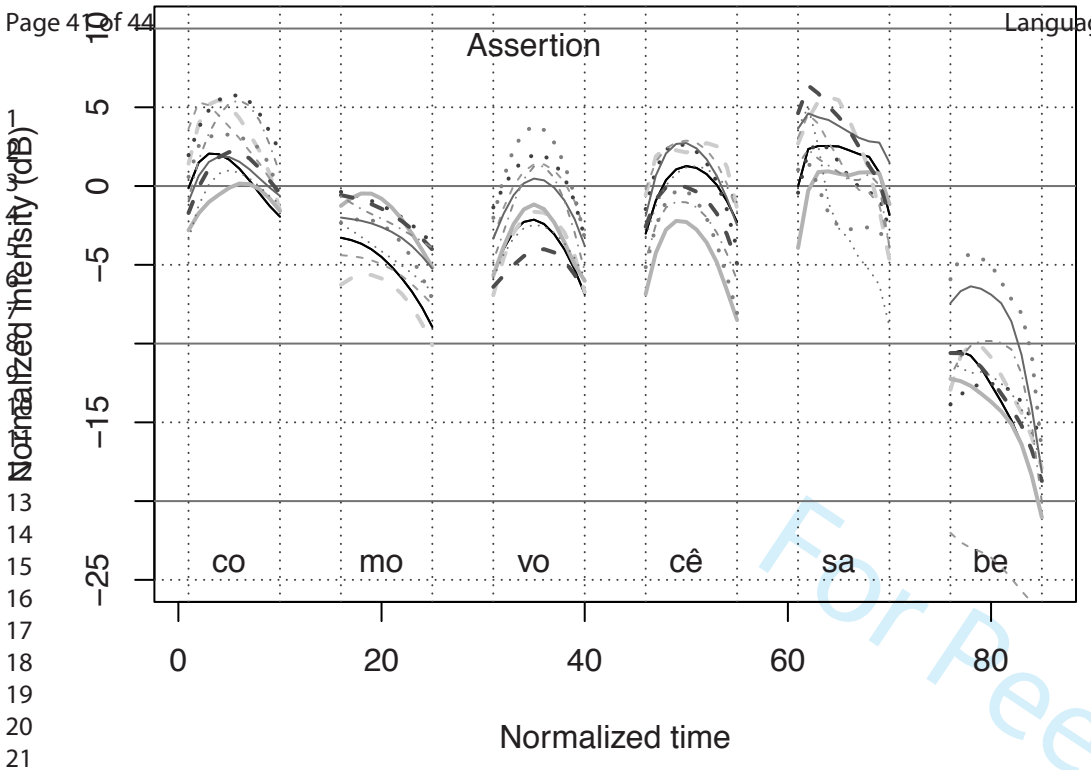
Figure 5: *Recognition rate of Experiments 1 and 2 with the following conditions: audio-only (AO), video-only (VI) and audiovisual (AV) from Experiment 1; audiovisual in 0 dB SNR (AV-0), audiovisual in –6 dB SNR (AV-6), audio-only in 0 dB SNR (AO-0) and audio-only in –6 dB SNR (AO-6) from Experiment 2.*

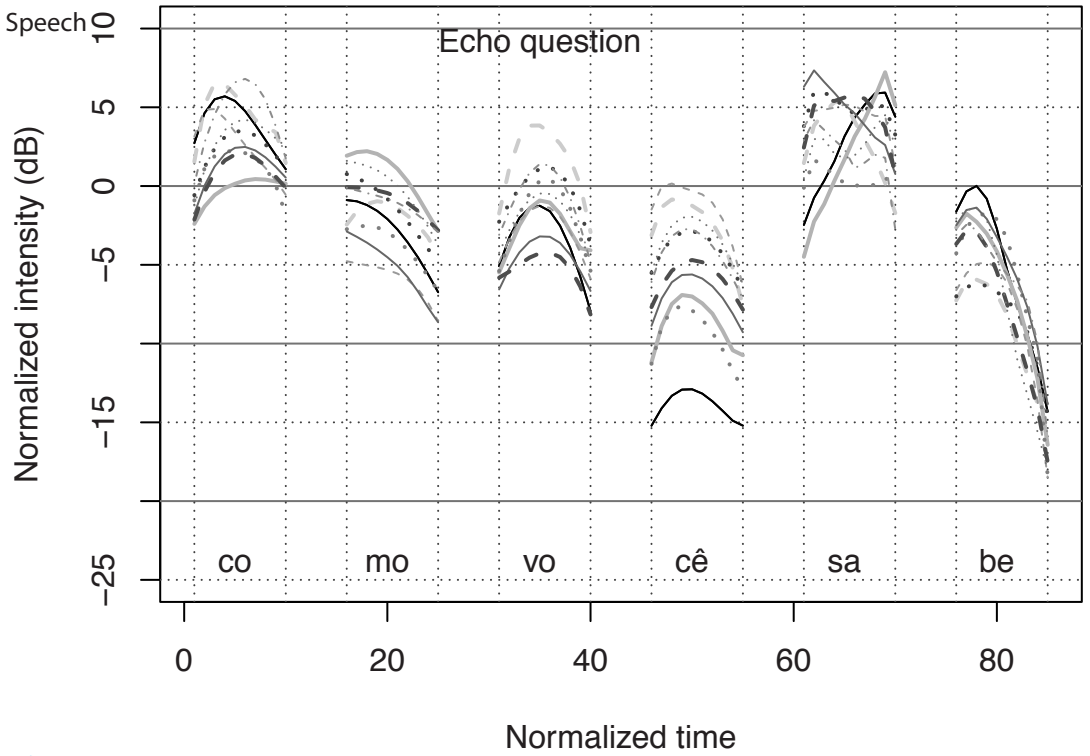
Figure 6: *Recognition rate of Experiments 1 and 3 with the following conditions: audio-only (AO), video-only (VI) and audiovisual (AV) from Experiment 1; congruent audio and video (AVc) and incongruent audio and video (AVi) from Experiment 3.*

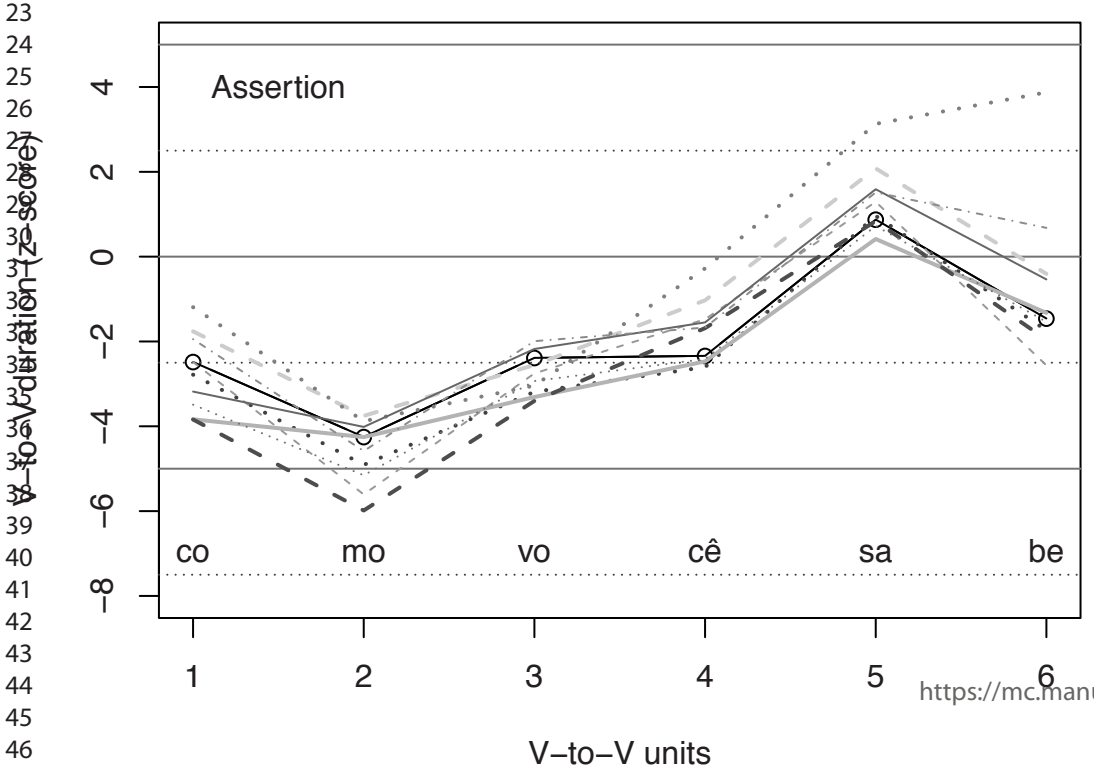


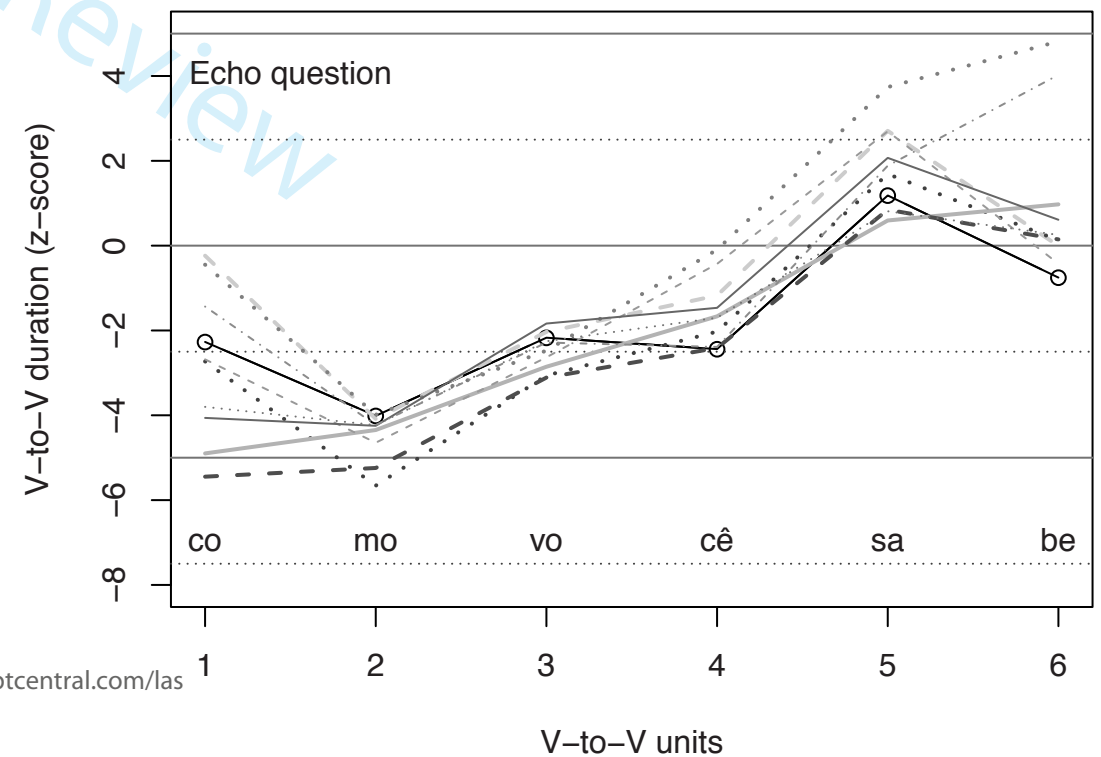


For Peer Review









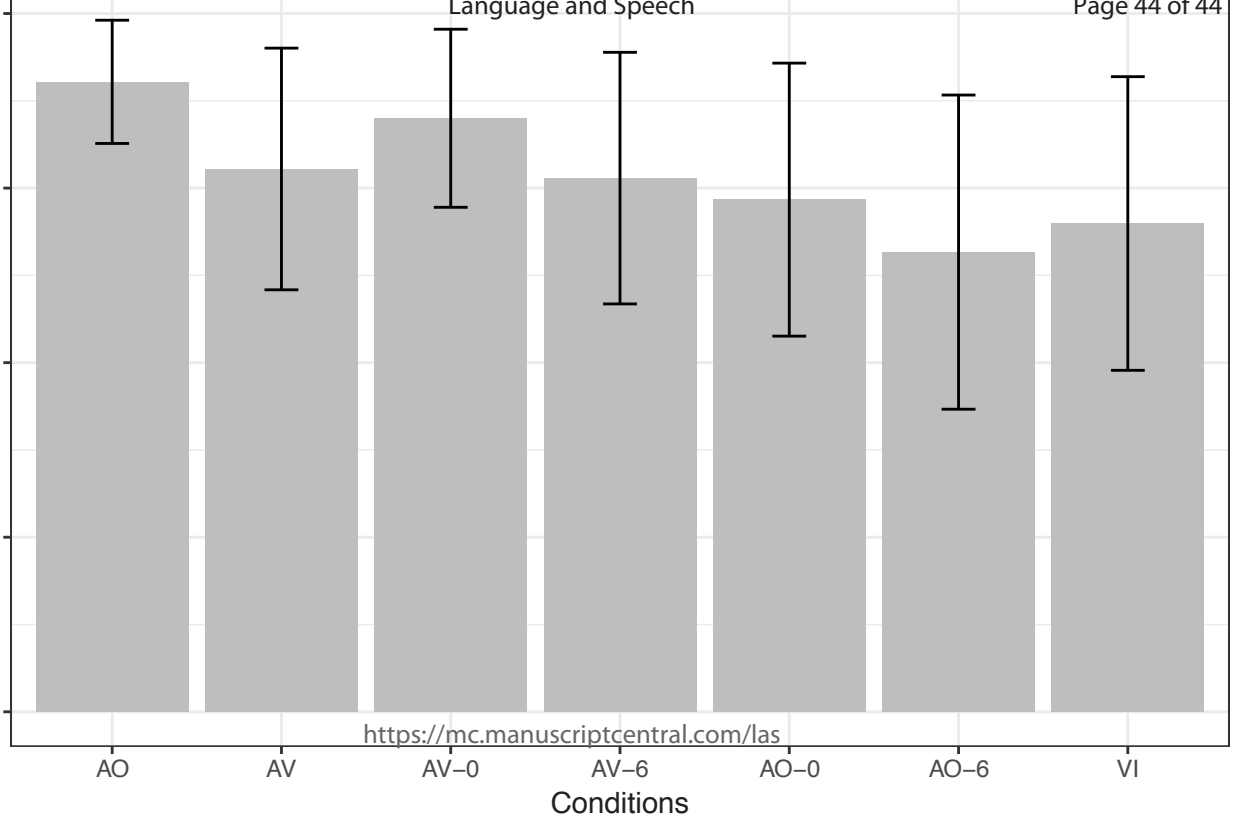
For Peer Review

Stills from a female speaker producing an assertion (above) and an echo question (below).

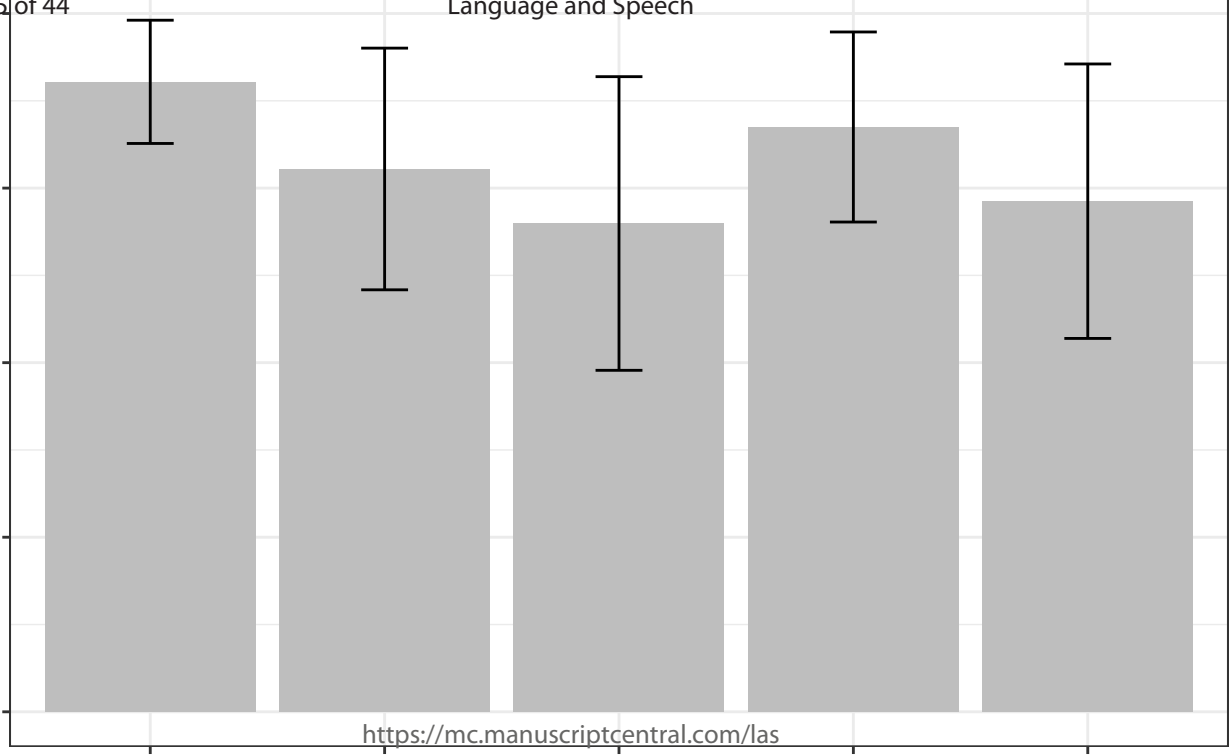
1375x518mm (72 x 72 DPI)

Stills from a male speaker producing an echo question.

1375x254mm (72 x 72 DPI)



https://mc.manuscriptcentral.com/las



https://mc.manuscriptcentral.com/las