Visual channel facilitates the comprehension of the intonation of Brazilian Portuguese wh-questions and wh-exclamations: evidence from congruent and incongruent stimuli

Luma da Silva Miranda, João Antônio de Moraes, Albert Rilliard

Language and Cognition (2024), 1–21
doi:10.1017/langcog.2024.16

ARTICLE

Visual channel facilitates the comprehension of the intonation of Brazilian Portuguese wh-questions and wh-exclamations: evidence from congruent and incongruent stimuli

Luma da Silva Miranda[1], João Antônio de Moraes[2,3] and Albert Rilliard[2,4]

[1]Eötvös Loránd University, Budapest, Hungary; [2]Federal University of Rio de Janeiro, Rio de Janeiro, Brazil; [3]The National Council for Scientific and Technological Development, Brasília, Brazil and [4]Université Paris Saclay, CNRS, LISN, Orsay, France
Corresponding author: Luma da Silva Miranda; Email: miranda.luma@btk.elte.hu

(Received 15 November 2022; Revised 13 November 2023; Accepted 23 February 2024)

Abstract
This paper presents an audiovisual perceptual analysis of the wh-question and wh-exclamation intonation in Brazilian Portuguese using auditory–visual congruent and incongruent stimuli, to investigate the relative importance of each modality in signaling pragmatic meanings. Ten Brazilian Portuguese speakers (five female) were filmed while producing both speech acts 10 times. Next, artificial stimuli were created: audio and visual cues were either matched (audio and video from the same speech act) or mismatched (audio and video from the different speech acts), resulting in 10 congruent and 10 incongruent stimuli of the wh-questions and the wh-exclamations. The perceptual experiment was taken by 36 Brazilians who identified the stimulus as a question or an exclamation. Results from the logistic regression showed that the factor 'congruence' was significant and had a significant interaction with 'speakers', which means that the congruent stimuli increased the comprehension of the Brazilian Portuguese wh-questions and wh-exclamations. In contrast, the incongruent stimuli tended to lower listeners' identification, but to a degree depending on individual speakers' strategies. Although variation in the accuracy of expressing both speech acts was also found across speakers, this study corroborates that the visual channel impacts the perceptual identification of the pragmatic intonation function of distinguishing sentence mode.

Keywords: Brazilian Portuguese; incongruent stimuli; multimodal perception; wh-question; wh-exclamation

1. Introduction

1.1. Multimodal communication

The production and comprehension of human language pass through multiple modalities to convey meaning, including verbal and nonverbal cues (Holler & Levinson, 2019; Kendon, 1980, 2004; Levinson & Holler, 2014; McNeill, 1992). While talking, speakers tend to spontaneously produce co-speech gestures that contribute to the expression of semantic and pragmatic communicative intents. Traditionally, in Pragmatics, the elements within the context of speech production (e.g., the position of objects) are considered to interpret the utterance's meanings. However, several studies have shown that visual cues are available in the communicative act itself that can also be used to reveal the pragmatic meaning of the utterance (Kelly, 2001). The pragmatic contribution of gestures accompanying nonemotional speech has been largely described (Barsalou, 2008; Kendon, 2004; McNeill, 1992), even though the role of gestures in human communication has been traditionally explored in emotional speech (Barkhuysen et al., 2010; Crespo Sendra et al., 2013; Ekman et al., 2002; Moraes et al., 2010, 2012).

In production studies of face-to-face conversation, it has been documented that gestures carry various meanings, including the speaker's communicative intentions. Considering the transmission of pragmatic functions by facial expressions (muscular activity of the face used for communicative intents) during a conversation, Bavelas and Chovil (2018) showed, for instance, that a speaker may produce the 'thinking face' to indicate that (s)he is remembering something or searching for a word or concept. This facial expression comprises a short gaze shift, either away or in the opposite direction of the interlocutor, or eyes closed. A type of facial expression commonly used to express disagreement is the 'not face' (Benitez-Quiroz et al., 2016), composed of eyebrow frowns, compressed chin muscles and pressed lips; it is also used to express emotions such as anger, contempt and disgust. Debras (2017) defined the 'shrug gesture' (Debras, 2017) as a kinetic ensemble possibly expressed through different body movements – the facial expressions, hand or shoulder gestures – to convey mainly interpersonal attitudes. For instance, the 'mouth shrug', in which the eyebrows are raised, and the lip corners are down, is considered a sign of 'disclaimer', expressing meanings such as 'I don't know' or 'I don't understand'. Other studies have demonstrated that smiles may express irony or sarcasm in a conversation (Bavelas & Chovil, 2018; Caucci & Kreuz, 2012; González-Fuente et al., 2015).

As for perception, when the visual channel is added to the auditory channel in the stimuli presentation, the identification accuracy of prosodic meanings increases (House, 2002; Miranda et al., 2021; Srinivasan & Massaro, 2003) and even the processing of the message becomes faster when compared to auditory speech alone (Borràs-Comes & Prieto, 2011; Cruz et al., 2017; Holler et al., 2018). The present study investigates the contribution of the speaker's facial gestures in the perception of pragmatic meanings conveyed by intonation, such as speech acts, which are defined as actions triggered by the speaker when (s)he produces an utterance (Searle, 1969).

In the last few decades, a substantial number of multimodal studies analyzed several functions expressed by intonation, such as demarcation (phrasing the speech stream into smaller units), highlighting (the place of prominence within an utterance) and the distinction of sentence mode (statements, questions, requests, commands etc.). For instance, regarding yes-no/echo questions and assertions, due to their essential function as building blocks in conversation (Holler et al., 2018), these

https://doi.org/10.1017/langcog.2024.16 Published online by Cambridge University Press

speech acts are vastly explored in a multimodal approach in several languages such as Swedish (House, 2002), American English (Srinivasan & Massaro, 2003), British English (Fisher, 1969), Catalan (Borràs-Comes & Prieto, 2011), Mexican Spanish (Gomes da Silva, 2019; Miranda et al., 2020a), Brazilian Portuguese (henceforth BP) (Miranda et al., 2021; Peres et al., 2011) and European Portuguese (Cruz et al., 2017). Overall, these studies support that different facial expressions participate in the transmission of pragmatic meaning conveyed by intonation and that visual cues are perceptually integrated with auditory cues, meaning that listeners benefit from prosodic functions presented in a bimodal condition.

Considering that the multimodal analysis of the expression of speech acts other than yes-no questions and assertions, such as wh-questions and wh-exclamations, are scarce, it leaves an open question of whether the visual signals produced for expressing these pragmatic meanings also facilitate communication. For instance, Gili Fivela (2015) analyzed facial expressions and head movements in a corpus of Italian wh-questions conveying surprise, exclamations and statements. The author found out that the more expressive the information was, the more use of facial gestures in the production of speakers. In Miranda et al. (2019), 10 BP speakers were filmed and recorded while producing the sentence '*Como você sabe*' as wh-questions (meaning, 'How do you know?') and wh-exclamations (meaning, 'How clever you are!'). The acoustic and visual cues of the production data from all speakers were described, and a perceptual experiment was set up. In the perceptual identification task, there were three conditions of stimuli presentation: audio-only, visual-only and audiovisual. Sixty BP speakers participated in the perceptual task. Results revealed that stimuli presented in the audiovisual condition achieved significantly higher identification rates than those presented in unimodal conditions. This showed that the auditory and visual cues of wh-questions and wh-exclamations are integrated during the cognitive process of speech perception.

The present study aims to expand the multimodal analysis of Miranda et al. (2019) by verifying the relative strength of the auditory and visual cues in the perception of Brazilian Portuguese wh-questions and wh-exclamations with a new approach. Stimuli constructed based on mismatched modalities were used in this study's perceptual experiment design, since very little is known about the contribution of facial cues to the perception of BP wh-questions and wh-exclamations. It is important to highlight that, from an evolutionary perspective, Levinson and Holler (2014) argue that the foundation for human multimodal communication comes from the human capacity for structured social interaction, in which the most common niche of language use is face-to-face interaction. This way, language comprehension includes gestures strongly connected to speech, and the experiment conducted in this study was designed to understand this connection better.

1.2. Previous studies on congruent and incongruent modalities

Many methodological approaches exist to study the role differences between auditory and visual modalities. Since information from different modalities is processed at several cognitive or neural levels (Holler & Levinson, 2019), an experimental setup using congruent and incongruent stimuli has been proposed to study their interaction. For instance, spatiotemporal ventriloquism effects have been studied to understand the role of each modality (Spence, 2007). Also, the well-known 'McGurk

https://doi.org/10.1017/langcog.2024.16 Published online by Cambridge University Press

effect' (McGurk & MacDonald, 1976), in which the articulation place of a phone presented bimodally may be the same (congruent presentation) or not (incongruent presentation) across modalities, is famous for demonstrating the integration of auditory and visual cues in speech perception at the phoneme level.

Studies with mismatched audiovisual stimuli have also demonstrated that speech and gesture are integrated at semantic, pragmatic and discursive levels. For instance, McNeill et al. (1994) used mismatched gesture-and-speech input carrying various meanings within narratives, such as spatial, perspective or anaphoric gestures. First, participants of the perceptual experiment had to watch a video of someone telling a story from an animated cartoon in which some gestures were mismatched with speech. Then, they had to retell the story to a second person. The analysis of the results showed that listeners omitted the part of the narrative with mismatched input or tried to form a new coherent interpretation of the excerpt. The study concludes that subjects exposed to artificially mismatched auditory and visual information, which is an unusual condition, still try to reconstruct a single idea, integrating speech and gesture.

Kelly et al. (2015) analyzed how gesture and speech cues are integrated for language comprehension, also using mismatched stimuli. The task of the perceptual experiment was to relate one word with the audio or visual input, which was either a manual gesture or an action on objects, to find out (i) whether there is any difference between the two visual targets used in the study and (ii) to verify if the incongruent audio and visual information would decrease the listener's accuracy and speed. The authors concluded that incongruent stimuli led to more errors in recognizing the spoken sentences accompanied by gestures and actions on objects.

This approach of analyzing speech with mismatched auditory and visual cues is also applied to prosodic patterns. The perception of BP assertions and echo questions using multimodal congruence and incongruence was tested by Miranda et al. (2021). The authors reported that the analyzed speech acts were better recognized when stimuli were congruent. In contrast, in the incongruent presentation of stimuli, listeners' identification scores of BP echo questions and assertions were lowered. Cruz et al. (2017) investigated the role of the visual channel in the perception of yes-no questions and statements, comparing different varieties of European Portuguese. The authors also used congruent and incongruent audiovisual presentations in the experiments. They concluded that the speech acts in congruent conditions were better recognized and identified more rapidly.

In Borràs-Comes and Prieto (2011), Catalan echo questions and focused statements were analyzed. Based on reaction time measures, it was concluded that listeners took longer to process the incongruent stimuli and that both pragmatic meanings received lower identification scores when stimuli were based on mismatched information. Also, regarding lexical stress, Swerts and Krahmer (2004) created artificial stimuli to investigate the perception of stress location in Dutch. Results also indicated that, in incongruent conditions, the visual channel interferes with the perception by lowering the identification rate of stress position. Taken together, all these studies suggest that, whereas the perceptual identification of prosodic functions in congruent presentation conditions is more accurate, incongruent audiovisual stimuli are more difficult to process than congruent stimuli, which means that facial expressions and speech are integrated into language processing.

https://doi.org/10.1017/langcog.2024.16 Published online by Cambridge University Press

1.3. The aim of the current study

This paper reports the design and the application of a perceptual experiment based on manipulating the congruence between modalities to investigate the role of the auditory and visual channels in the perception of wh-questions and wh-exclamations in Brazilian Portuguese. To our knowledge, no study has examined so far how BP listeners perceive these two speech acts using this specific approach; thus, it remains to be determined whether the visual channel will influence the perceptual identification of wh-questions and wh-exclamations when the audio and visual cues are signaling different pragmatic meanings. Additionally, although studies on prosody have been analyzing the auditory channel alone, research gives evidence that the visual channel has a significant role in speech comprehension (Levinson & Holler, 2014) and that human communication is based on one multimodal integrated system (Holler & Levinson, 2019; Massaro & Cohen, 1983; McNeill, 1992). Hence, the analysis of the interaction between auditory and visual cues producing the pragmatic function of intonation sheds light on the understanding of multimodal language processing.

Since Miranda et al. (2019) concluded that both channels are integrated during speech perception of BP wh-question and wh-exclamations with listeners benefiting from the presentation of stimuli with congruent auditory and visual cues, in this paper, the research question and hypothesis of this study are: Will the mismatched facial and audio cues allow the recognition of the speech acts? We hypothesize, based on previous studies, that the incongruent stimuli will hinder the identification of the pragmatic meaning expressed in the auditory modality by intonation, although the comprehension of the speech act will still be possible. Speech and gesture are tightly integrated (Kelly et al., 2002) so that the unusual visual cues in the incongruent condition will force the listeners to interpret the stimuli as a single pragmatic meaning.

2. Auditory and visual cues of Brazilian Portuguese wh-questions and wh-exclamations

2.1. Acoustic cues of BP wh-questions and wh-exclamations

In Brazilian Portuguese, the intonation notably carries the difference between wh-questions and wh-exclamations (see Figure 1).

BP wh-questions tend to be produced with a high initial F0 on the first stressed syllable followed by a falling F0 contour until the sentence's nuclear region, as Moraes (2008) described for the Rio de Janeiro dialect. Such a nuclear F0 fall was also found in other Brazilian Portuguese varieties, such as Vitória da Conquista (Bahia) (Oliveira et al., 2014) and Florianópolis (Santa Catarina) (Zendron da Cunha, 2016) and other Portuguese varieties (e.g., Standard European Portuguese), in Falé (2006) and Mata (1990), as well as for other languages, such as English (Bolinger, 1989), French (Beyssade et al., 2007) and Hungarian (Gyuris & Mády, 2013). However, either in German (Repp, 2015, 2020) or in Cosenza Italian (Sorianello, 2011), a rising F0 can also be observed at the nucleus. In Portuguese, wh-questions with a rising nuclear F0 contour carry a supplementary politeness meaning (Frota et al., 2015). Regarding the prenucleus, as already stated, BP wh-question intonational contour presents a high initial F0 on the first stressed syllable, whereas, in different languages, wh-questions may start either with a high phrase-initial boundary tone as in Hungarian (Gyuris & Mády, 2013), or with a prenuclear H* accent in the wh-word as observed for the

https://doi.org/10.1017/langcog.2024.16 Published online by Cambridge University Press

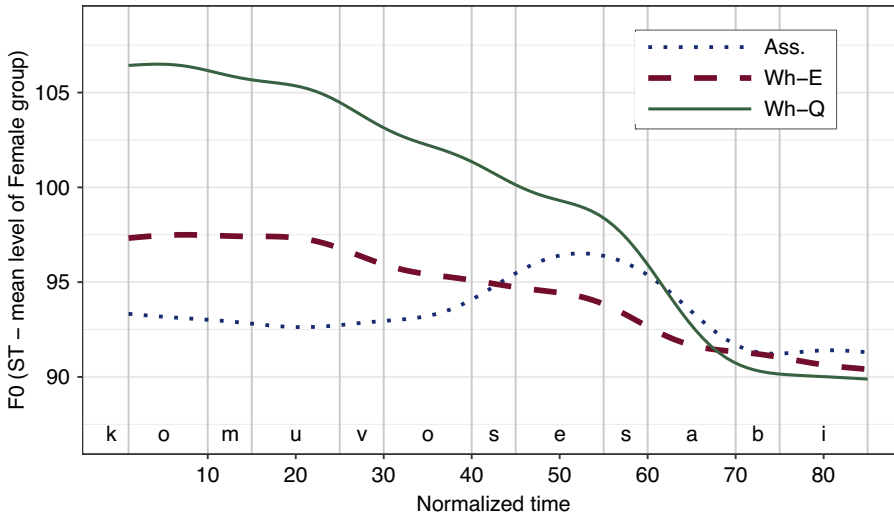

Figure 1. Mean intonational contours of Brazilian Portuguese wh-questions (solid line), wh-exclamations (dashed line) and assertions (dotted line) produced by the 10 speakers of this study.

Italian variety of Cosenza (Sorianello, 2011). When compared to assertions, there are other typical acoustic cues to wh-questions; for instance, in German wh-questions (Brinckmann & Benzmuller, 1999), the pitch range is more extensive, the topline decline is shallower, and there is a higher pitch offset and a higher nuclear peak in German wh-questions (Rohloff & Michalsky, 2018). Similarly, Miranda et al. (2020b) described a higher mean F0 in BP wh-questions compared to other speech acts such as assertions, echo questions and wh-exclamations.

As for the wh-exclamations in BP, different intonational contours were described depending on the wh-word employed in the sentence (Zendron da Cunha, 2016). Wh-exclamations starting with the wh-word '*como*' (how) are characterized by a higher initial F0 in the wh-word and a falling F0 on the nuclear region of the contour, whereas for wh-exclamations starting with the wh-words '*que*' (what) and '*quanto*' (how many), a lower F0 is observed on the wh-word and a rising intonation at the nucleus – an intonational contour that is similar to the one described for yes-no questions in BP (Moraes, 1998, 2008). The present study analyzes wh-exclamations with the wh-word '*como*' (how). Based on the Rio de Janeiro variety, Moraes (2008) described the intonation contour of this type of wh-exclamation with a high initial F0 followed by a falling F0 contour until the nuclear region; however, this fall is less steep than the one found in the wh-question. Findings on wh-exclamations in other languages such as English (O'Connor & Gordon, 1961), German (Batliner, 1988; Repp, 2015, 2020), Italian (D'Eugenio, 1976) and French (Delattre, 1966) also showed a falling F0 intonation on the nucleus. In addition, the prenuclear region in wh-exclamations may also vary across languages: one may observe either an initial extra high pitch, represented by the prenuclear accent %H, such as in Cosenza Italian (Sorianello, 2011, 2012), or an initial low pitch, such as in Hungarian (Gyuris & Mády, 2013).

Other acoustic cues are mentioned in the literature as typical of wh-exclamations; for example, Sorianello (2011) pointed out that, in Cosenza Italian wh-exclamations, there is a lengthening of the nuclear stressed syllable. In Brazilian Portuguese,

https://doi.org/10.1017/langcog.2024.16 Published online by Cambridge University Press

Miranda et al. (2020b) did not find any durational or intensity differences in the overall mean of these two acoustic parameters to distinguish the intonation contour of wh-exclamations from wh-questions. However, the authors noted in the same study that there were differences in the most intense syllables within the intonational contours: whereas the most intense syllables in the wh-exclamation contour are the first and the second syllables ('co' and 'mo'), in the wh-question, the second and the third syllables ('mo' and 'vo') are the most intense. As for duration, although at the beginning of the wh-exclamation contour, compared to the wh-question contour, there is longer duration (which was measured using V-to-V units: the onset of one vowel until the onset of the subsequent vowel; Barbosa, 2006), overall, the duration patterns of both speech acts include a progressive temporal lengthening until the last stressed syllable; after this, syllables have shorter duration.

Additionally, perceptual experiments in Brazilian Portuguese based on the acoustic signal that analyzed the wh-question and wh-exclamation intonation stressed the importance of the F0 movements on the nuclear region of the contour, along with intensity and duration patterns, for the recognition of these speech acts. For instance, Oliveira et al. (2014), by manipulating the F0 contour of BP wh-questions and wh-exclamations, found out that, while stimuli with lower F0 on the nucleus were perceived as wh-questions, stimuli with higher F0 at the nucleus were recognized as wh-exclamations. Miranda et al. (2022), in turn, applied a perceptual experiment with resynthesized stimuli to weigh the relevance of F0 movements versus duration and intensity patterns in the perceptual identification of speech acts. It is worth mentioning that, although the acoustic analysis of Miranda et al. (2020b) indicated that there are significant differences between the two speech acts only in relation to the intensity parameter, the duration and intensity patterns of each intonational contour were analyzed in a combined manner in the authors' perceptual experiment, transplanting the melodic contour onto the rhythmic base of the wh-questions and wh-exclamations, an approach described, for example, by Moraes and Rilliard (2018) and already applied in other BP intonation patterns (Miranda et al., 2022, 2023). Results showed that BP listeners identified stimuli with a falling nuclear F0 movement along with the wh-question patterns of duration and intensity as wh-question, whereas stimuli with a falling F0 followed by a subtle rising nuclear F0 movement plus its duration and intensity patterns from the wh-exclamation contour were identified as wh-exclamation.

In sum, previous studies on BP wh-questions and wh-exclamations provide evidence that there are distinctive auditory cues (nuclear F0 movements and intensity patterns) that listeners could rely on in distinguishing these speech acts. Although there are similarities between the two speech acts, such as the initial high F0 at the prenucleus and the falling F0 movement at the nucleus, wh-questions have a higher mean F0 when compared to wh-exclamations and a steeper falling F0 movement in comparison with the wh-exclamation nuclear contour. Besides, the most intense syllables in the wh-question are in middle of the contour, and, in the wh-exclamation, the most intense syllables are in the beginning of the contour.

2.2. Visual cues of BP wh-questions and wh-exclamations

Several studies have shown that visible bodily movements can contribute to expressing various speech acts. As for questions and assertions, eye gaze and eyebrows

https://doi.org/10.1017/langcog.2024.16 Published online by Cambridge University Press

movements are the most typical visual cues in the literature (Kendrick & Holler, 2017; Torreira & Waltersson, 2015). Direct gaze was described as a signal for questions when it is held until the addressee provides an answer (Argyle & Cook, 1976; Borràs-Comes et al., 2014) and gaze shifts were described in the production of assertions. Regarding eyebrow movements, they are observed for the production of questions and verified frequently in other speech acts such as instructions related to requests (Flecha-García, 2010). Although previous research has indicated that specific facial signals are associated with questions and assertions, few works have targeted wh-questions and wh-exclamations. Miranda et al. (2019) described specific facial gestures in the production of BP wh-questions and wh-exclamations. For wh-questions, a lowered eyebrow plus the head turned right were the most typical combination of gestures, whereas for wh-exclamations, raised eyebrow, lips corner raiser and up and down head movement were found, as seen in Figure 2.

The description of facial gestures for the BP wh-questions is in line with previous studies in both spoken (Gili Fivela, 2015) and sign languages (Cruz et al., 2019; Paiva et al., 2016), in which the eyebrow frowns were described as question markers. In agreement with Nota et al. (2021), one can consider that an eyebrow frown signals a subclass of speech act, as in the case of a wh-question. As for the wh-exclamation facial gestures, one can observe in Figure 2 that other visual cues co-occur with the smile and head nod, such as the lowered eyebrow. Although it is argued that smiles can indicate irony or sarcasm (Bavelas & Chovil, 2018; Gaucci & Kreuz, 2012; González-Fuente et al., 2015), in the production context of a wh-exclamation, the smile may be signaling a positive attitude toward the addressee. In short, the data described by Miranda et al. (2019) showed that specific facial gestures are related to each BP speech act analyzed in this paper.

3. Method

3.1 Recording procedure

The original material used in this paper is the same found in Miranda et al. (2019), however video recordings were used in this study to create artificial stimuli. The corpus of this study is, thus, composed of one BP sentence, '*Como você sabe*', that was produced either as a wh-question (meaning 'How do you know?') or as a wh-

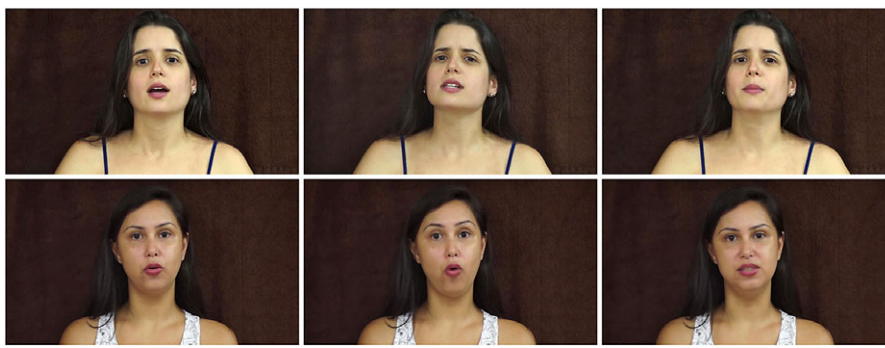

Figure 2. Facial gesture for wh-question produced by speaker 4 (top row) and wh-exclamation produced by speaker 6 (bottom row).

https://doi.org/10.1017/langcog.2024.16 Published online by Cambridge University Press

exclamation (meaning 'How clever you are!'). In terms of syntax, the BP wh-question and wh-exclamation sentences are constructed with a wh-word at the front of the clause; however, the wh-questions in the Portuguese language present a mobility of this item that could also be placed at the end of the utterance (*in situ*): '*O que você falou?* X *Você falou o quê?*' (Both meaning, 'What did you say?'). This sentence with the wh-word '*como*' ('how') in the initial position of the clause allows us to compare the wh-question and the wh-exclamation intonation contour without interference with lexical or morphosyntactic cues.

Ten Brazilian speakers (five female, five male) from Rio de Janeiro with corrected-to-normal vision and no language, hearing or motor disorders were recorded for the corpus. At the time of the recordings, nine out of 10 speakers were graduate or undergraduate students at the Federal University of Rio de Janeiro, except one who was a professor at the same university. The age of the participants ranged from 19 to 64 years old (mean: 28.5). The recording procedure was made as follows. At first, informants had to read and sign a consent form detailing the procedure and the future use of the recordings. Next, the experimenter explained orally the pragmatic context linked to the production of wh-questions and wh-exclamations by describing real-life examples using the sentence '*Como você sabe*'. For instance, the speaker is asking a specific information from an interlocutor in the wh-question context (Speaker 1: We have a test tomorrow. /Speaker 2: '*Como você sabe*?'/ How do you know?'). In the wh-exclamation context, the speaker is expressing his admiration toward the interlocutor's knowledge (Speaker 1 clearly explained a difficult theory. / Speaker 2: '*Como você sabe!*'/ 'How clever you are!'. After the explanation of the two contexts, the experimenter asked the participants to imagine themselves in these contexts and to produce the two speech acts. This was done without the experimenter producing either the intonation or the facial gestures related to these expressions. The recording sessions took place at the Phonetics Laboratory of the Federal University of Rio de Janeiro in a sound-attenuated room. A dark background was positioned behind the seated speakers. A SONY NEX-F3 camera, set 90 cm from the speaker, was used to film the participants, framing the upper body and face. Along with the camera, a Zoom recorder was used to record high-quality audio; located 20 cm from the speakers' mouths and outside the camera frame. During the recording session, the participants were asked not to wear glasses. As part of a larger project, the recording session included the production of four speech acts: assertion, echo-question, wh-question and wh-exclamation. The experimenter instructed the type of speech act to be produced, alternating the order. By the end of the recording session, 400 sentences were recorded since the 10 speakers produced the four speech acts 10 times. This large number of repetitions was intended at acoustical analysis. Some of the results from this acoustical analysis were presented earlier in this paper and they are detailed in Miranda et al. (2020b). In the present study, only the production of the wh-questions and wh-exclamations is used for the perceptual analysis. Besides, the multimodal perceptual analysis of assertions and echo questions can be seen in Miranda et al. (2021).

After the recordings, the software Vegas Pro (Magix, 2016) was used to synchronize the audio and video channels thanks to a handclap at the beginning of each session, and, to cut each sentence into individual videos with a length between 2 and 3 seconds. For the current work, the eighth and ninth repetitions of each speech act produced by the 10 speakers were chosen, so to avoid a selection linked to the

https://doi.org/10.1017/langcog.2024.16 Published online by Cambridge University Press

perceived quality of their performances. Hence, 40 videos (2 speech acts × 2 repetitions × 10 speakers) were selected.

3.2 Perceptual experiment – congruent and incongruent stimuli

The perceptual experiment aims to analyze the role of the visual channel in identifying wh-questions and wh-exclamations, compared to the auditory channel, presenting subjects with stimuli with congruent and incongruent modalities.

3.2.1 Stimuli and procedure

The audio and the visual performances of BP wh-questions and wh-exclamations were combined to create congruent stimuli, that is, stimuli with the same speech act in both modalities, and incongruent stimuli, that is, stimuli with different speech acts in the two modalities. This was done to evaluate the relative importance of each modality for the perceptual identification of these speech acts. The combination of the stimuli was made with the Vegas Pro software (Magix, 2016) following these steps: for the congruent stimuli, the two repetitions of the same speech act were selected, with the audio of the first dubbed into the video of the second. For incongruent stimuli, one repetition of each speech act was used; then, the video of one speech act was dubbed with the audio from the other. The outcome of this procedure was a set of 40 stimuli. On the one hand, there were 20 congruent stimuli for the two speech acts: 10 auditory wh-questions with visual wh-questions and 10 auditory wh-exclamations (Wh-E) with visual wh-exclamations (Wh-Q). On the other hand, 20 incongruent stimuli were also produced: 10 auditory wh-questions with visual wh-exclamations and 10 auditory wh-exclamations (Wh-E) with visual wh-questions (Wh-Q). For clarity, the incongruent stimuli will be referred to by the name of their auditory speech acts. Thus, a congruent 'Wh-E' is based on two 'Wh-E' recordings, one for each modality (resp. an audio 'Wh-E' and a visual 'Wh-E'). An incongruent 'Wh-E' is based on an audio 'Wh-E' and a visual 'Wh-Q' from the same speaker (resp. an audio 'Wh-E' and a visual 'Wh-Q'). It is worth mentioning that it was important to create artificial stimuli even for the congruent stimuli (combining two different repetitions of the same speech act produced by the same speaker), so all the stimuli in the experiment are artificial. Four samples of congruent and incongruent wh-questions and wh-exclamations are provided in the Supplementary materials of this paper.

Using the PsyToolkit website (Stoet, 2010, 2017), a perceptual identification task was designed and applied over the internet. In a forced-choice task, BP listeners had to identify each stimulus with matched or mismatched auditory and visual cues expressing a 'question' or an 'exclamation'. Once the experiment started, each stimulus appeared sequentially on the screen, in a randomized order. Participants could repeat the stimuli as many times as they wanted before selecting the answer, but they were not allowed to go back and change the response of a previous stimuli. An experiment run lasted 10–15 minutes.

3.2.2 Participants

Thirty-six BP speakers participated in the perceptual experiment. Participants (25 women and 11 men) were from various regions from Brazil, yet most from the

https://doi.org/10.1017/langcog.2024.16 Published online by Cambridge University Press

Rio de Janeiro variety. They were either graduate or undergraduate students aged 18–54 (mean: 31.3). They had normal hearing and normal or corrected-to-normal vision and no speech or language disorders history. Before taking the test, they had to read a disclaimer explaining the study's aims and that they could stop their participation at any time without penalty; after accepting the conditions and reading the instruction, the stimuli presentation began.

3.2.3 Statistical analysis

The raw answers were the speech act identified for each stimulus; they were expressed as hit (1) or miss (0) whether the listener identified the mode of the auditory modality (i.e., hits were attributed to wh-question answers for congruent 'Wh-Q' stimuli and incongruent stimuli with 'Wh-Q audio' and to wh-exclamation answers for congruent 'Wh-E' stimuli and incongruent stimuli with 'Wh-E' audio; other answers were coded as misses). This was done as the audio modality is supposed to dominate, and it evaluates the decrease in the identification ratio of the audio modality introduced by the incongruent visual modality.

For each stimulus, the number of hits and misses over all answers was used to express a proportion of hits. This proportion or ratio of correct answers (correct in the sense explained above – related to audio modality) was used as the dependent variable for a logistic regression with, as independent variables, the 'audio speech act' with two levels (wh-question – 'Wh-Q' and wh-exclamation – 'Wh-E'), the 'congruence' between modalities with two levels (congruent – 'C' and incongruent – 'I') and the 'speaker' who produced the stimuli (10 levels), plus the double and triple interaction between these factors. The logistic model was fitted using the R software (R core team 2021). Details of the statistical processing are given in the R script available in the Supplementary materials of this article.

Note that we chose to model 'speaker' as a fixed factor and not as a random factor (as it may be customary in some recent publications) for several reasons: first, the speakers recruited for this designed experiment have been selected for being relatively familiar with the linguistic notions they had to perform – they are thus not representative of the general population. The results of this test shall thus be regarded as limited to the specific variety of laboratory speech they represent and may not generalize typically to spontaneous speech. The constrained nature of the test and the specificity of its stimuli (e.g., requiring multiple productions of the same sentence for strictly aligned dubbed versions of the stimuli) was an overwhelming task to base it on spontaneous speech. Considering each speaker's idiosyncrasy is also important to see how individual choices affect the outcome: the speakers here are not an annoying variable, but their individual strategies must be considered to explain the variation in the perceivers' reactions.

4. Results of the perceptual experiment

Figure 3 presents the mean proportion of correct answers (hit) obtained by the stimuli of both speech acts, wh-exclamation ('Wh-E') and wh-question ('Wh-Q'), for both presentation conditions: congruent ('C') and incongruent ('I'). One can observe a decrease in the proportion of hits in incongruent presentations and lower identification levels for the wh-exclamation speech act.

https://doi.org/10.1017/langcog.2024.16 Published online by Cambridge University Press

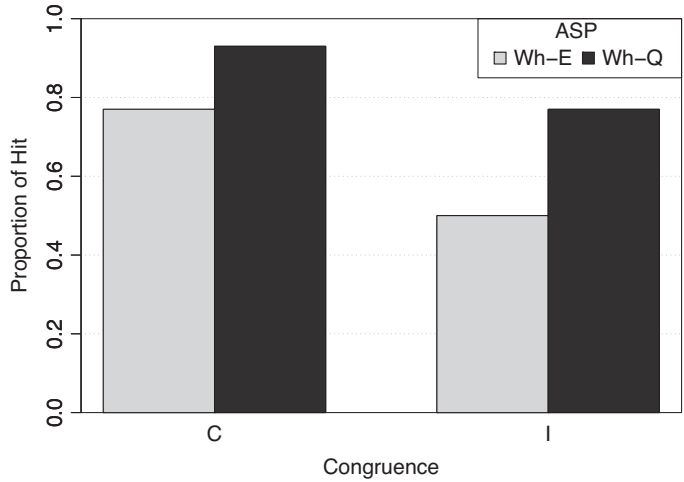

Figure 3. Mean proportion of hit obtained for both audio speech act 'ASP' (wh-exclamation: 'Wh-E' and wh-question: 'Wh-Q') in the congruent ('C') and incongruent ('I') conditions of presentation.

The complete logistic model described in Section 3.2.3 was submitted to a simplification process (Baayen, 2008; Crawley, 2013; Gries, 2013) that ended up removing the non-significant triple interaction ($\chi^2(9) = 14.4, p = 0.11$) and the double interaction between 'audio speech act' and 'congruence' ($\chi^2(1) = 0.009, p = 0.92$). It also allowed grouping speakers receiving similar answers into five groups: speakers (1/7), (3/5/8/10), (2/4), (6), and (9). The minimal optimal model is summarized in Table 1 (output using R's 'parameters' library, Lüdecke et al., 2020); it took thus into consideration the three factors 'audio speech act', 'congruence' and five-level 'speaker group', plus the double interactions between 'speaker group' and 'audio speech act', and 'speaker group' and 'congruence'.

The model shows that the main effect of the 'audio speech act' was not significant. At the same time, the 'congruence' between modalities and the 'speaker groups' had significant main effects – and both 'audio speech act' and 'congruence' interact with 'speaker group'. The two interactions are also detailed in Figures 4 and 5.

A pairwise contrast between levels of the five 'speaker groups' and the two levels of 'congruence' (Bonferroni correction, calculated using R's 'phia' library, De Rosario-Martinez, 2015; see Figure 4) showed two important results. First, the congruent stimuli reached a high and comparable level of identification for all groups (i.e., all performances were identified at a similar level, above 80%). Second, incongruent stimuli tend to lower the identification levels (no incongruent stimuli received a higher score than its congruent counterpart), but in a speaker-dependent fashion. There are three different situations: (i) groups (2/4) and (9) did not show an influence of the visual cues on the identification level of the speech acts. For these three speakers, the identification of the speech act is similar in both congruent and incongruent presentations; (ii) for group (3/5/8/10), the identification score for incongruent presentations significantly lowered, but the identification of the audio modality is still at about 70% (i.e., the 'audio speech act' was identified preferentially for these four speakers) and (iii) for groups (1/7) and (6), the identification score of the 'audio speech act' for incongruent presentation lowered at about chance level

https://doi.org/10.1017/langcog.2024.16 Published online by Cambridge University Press

Table 1. Summary of the model, reporting the regression coefficients (Log-odds) and their standard errors (SE), with the associated confidence intervals (CI), *z*- and *p*-values for each parameter and the constant (other reporting tables are available through the *R* script in the Supplementary materials of this paper)

Parameter	Log-odds	SE	95% CI	*z*	*p*
(Intercept)	2.54	0.36	[1.89, 3.31]	7.04	<.001
ASP [Wh–Q]	–4.66e–03	0.30	[−0.60, 0.59]	−0.02	0.988
Cgr [I]	−2.60	0.37	[−3.38, −1.92]	−7.03	<.001
Spk5 [2/4]	−1.27	0.46	[−2.19, −0.39]	−2.77	0.006
Spk5 [3/5/8/10]	−1.61	0.40	[−2.45, −0.86]	−3.99	<.001
Spk5 [6]	−0.87	0.52	[−1.90, 0.15]	−1.68	0.093
Spk5 [9]	−2.27	0.49	[−3.27, −1.33]	−4.62	<.001
ASP [Wh–Q] × Spk5 [2/4]	2.08	0.55	[1.04, 3.24]	3.76	<.001
ASP [Wh–Q] × Spk5 [3/5/8/10]	2.31	0.40	[1.53, 3.11]	5.76	<.001
ASP [Wh–Q] × Spk5 [6]	−0.78	0.49	[−1.75, 0.18]	−1.59	0.112
ASP [Wh–Q] × Spk5 [9]	2.36	0.58	[1.26, 3.57]	4.03	<.001
Cgr [I] × Spk5 [2/4]	2.07	0.52	[1.07, 3.10]	4.01	<.001
Cgr [I] × Spk5 [3/5/8/10]	1.34	0.44	[0.51, 2.23]	3.07	0.002
Cgr [I] × Spk5 [6]	1.32	0.54	[0.27, 2.39]	2.45	0.014
Cgr [I] × Spk5 [9]	2.00	0.57	[0.89, 3.14]	3.50	<.001

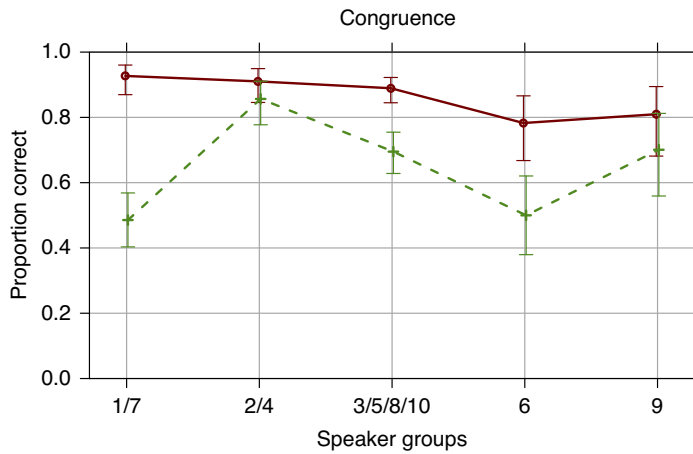

Figure 4. Means and confidence intervals estimated by the regression model for the interactions between factors 'congruence' of audiovisual modalities (separate lines) and 'speaker group' (x-axis).

(around 50%): for these stimuli, listeners attributed both possible speech acts at comparable levels – the two modalities were equivalently weighted for these three speakers.

Regarding the first and third results abovementioned, after an analysis of the visual data, it was concluded that, in terms of visual cues, (i) speakers 1, 7 and 6 used more facial cues than the other speakers, producing wh-questions with more head movements, whereas the wh-exclamations were produced with the combination of the eyebrow, head movements plus the lips movements; and (ii) speakers 2, 4 and 9 presented fewer uses of facial movements in the visual production of wh-questions and wh-exclamations. This may have influenced listeners' perception, so they relied more on the audio cues to distinguish both speech acts for this group of speakers.

https://doi.org/10.1017/langcog.2024.16 Published online by Cambridge University Press

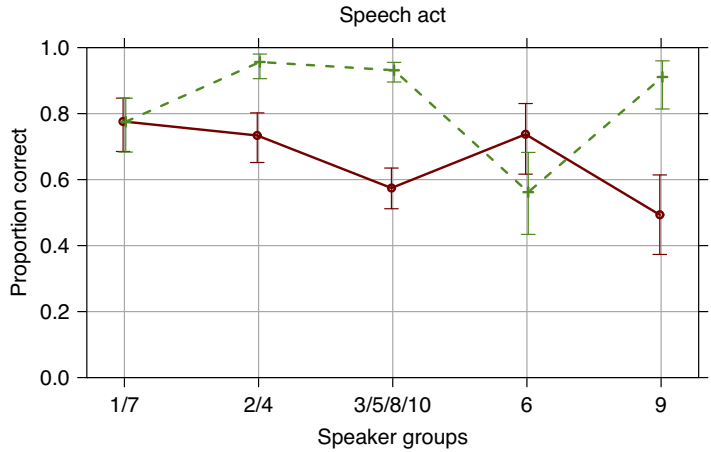

Figure 5. Means and confidence intervals estimated by the regression model for the interactions between factors 'audio speech act' (separate lines) and 'speaker group' (x-axis). Speech acts: wh-question (PQ) and wh-exclamation (EX).

The post hoc contrast run on the interaction between the factors 'audio speech act' and 'speaker group' (see Figure 5) showed that the performances of speakers varied across individuals (with speaker 6 having lower performances) and across speech acts – some speaker groups having a comparable performance for both speech acts (groups 1/7 and 6). In contrast, the other groups (seven speakers in total) had higher identification scores for wh-questions than for wh-exclamations. In some cases (speaker 9, and group 3/5/8/10), the wh-exclamations received identification scores close to chance; still, it is important to recall that this score regroups congruent and incongruent stimuli.

As for the wh-questions produced by speaker 6, the authors of this paper analyzed the acoustics of the production of this speech act and concluded that there was a subtle F0 rise in the final post-stressed syllable 'be' of the utterance. As already noted by Frota et al. (2015), a final rise in the wh-question intonational contour in Portuguese may add an attitudinal meaning of politeness to the question. This might be the reason why listeners did not give a good evaluation of speaker 6's production, which made the wh-question intonation produced by this speaker different from the others. In addition, the wh-exclamation produced by speaker 9, as well as speakers 3, 5 and 10, had a flat F0, in the nuclear region of the contour, instead of a final rise with a small F0 excursion found in the production of the other speakers. Although speaker 8 is grouped with speakers 3, 5 and 10, which means that there is no significant difference between them, it is important to highlight that this speaker produced a wh-exclamation with a falling F0 followed by a rise in the stressed syllable at the end of the contour moderately higher than the other speakers, which might be related to expressing an attitude of surprise.

5. Discussion

This paper analyzed the perception of the auditory and visual cues of wh-questions and wh-exclamations in Brazilian Portuguese using congruent and incongruent

https://doi.org/10.1017/langcog.2024.16 Published online by Cambridge University Press

stimuli. Based on a previous study (Miranda et al., 2019), both speech acts have already been described as having discriminant auditory and visual cues that listeners could rely on to identify the pragmatic meaning of the utterances. The present study confirmed that the congruent presentation of stimuli facilitates the comprehension of speech acts, which supports an integration of the auditory and visual channels for speech perception. As one can observe in the results of Figure 4, in the Section 4 of this article, the identification level for the 'audio speech act' is mostly above 50% – which means that there is indeed a dominance of the audio modality in expressing the difference between these two speech acts. In addition, the ratio of correct identification of the 'audio speech act' in 8 out of 10 speakers was higher in the congruent condition. Despite these caveats, it is important to stress that most of the results presented in this paper are in line with previous studies (Borràs-Comes & Prieto, 2011; Cruz et al., 2017; Miranda et al., 2021; Swerts & Krahmer, 2004).

However, it is also relevant to highlight that the hypothesis of the present study, which predicted that the incongruent stimuli would hinder the identification of the speech acts was partially confirmed. The speaker group '3/5/8/10' showed a decrease in 'audio speech act' identification for incongruent stimuli, but still received mean identification close to 70%. In speakers '2/4' and '9', there was no influence of the mismatched visual cues in the identification rate of the speech acts. One possible explanation for this result was the lower use of visual cues in the speech acts produced by these speakers. On the other hand, in the case of speakers '1/7' and '6', the incongruent presentations received scores down to about 50% of identification ratio for the audio speech act – which means that the visual modality has a considerable weight in those cases. These groups of speakers used more visual cues to produce the speech acts. This outcome illustrates the speaker-dependent use of modalities, and the global primary importance of the auditory cues for expressing these speech acts.

It is also worth mentioning that the variability found in the production of the facial gestures within the speaker groups of the present study is in line with previous studies (Keating et al., 2003), which indicated differences regarding the use of facial expressions produced by speakers. For instance, some speakers may use more head nods, while others rely more on moving either the eyebrow or the mouth. Taking the results of the congruent and incongruent stimuli together, it is possible to conclude that there is a tight link between speech and gesture in language comprehension, because, even when the auditory and visual cues are mismatched, listeners try to integrate cues from both channels to form a single interpretation of the utterance (McNeill et al., 1994).

In addition, as shown in Figure 5, overall, the wh-question is better recognized than the wh-exclamation, a typical result found in previous descriptions. In fact, Zendron da Cunha (2016) also verified in the application of her perceptual experiments that Brazilian listeners have more difficulties in identifying the wh-exclamation intonation contour because this speech act in BP has two different intonational behaviors, as already mentioned in Section 2 of this article, considering the wh-word in the syntactic structure ('*como*' – 'how' x '*que*' – 'what' and '*quanto*' – 'how many'), whereas the wh-questions present the same intonation contour, independently from the wh-word in the structure of the sentence. The exception of this finding is the wh-question produced by speaker 6, which differed from the other speakers. It is possible that her production was not recognized as neutral by the listeners due to a final F0 rise in her wh-question contours, which adds an attitude of politeness in the speech act production, as described in the literature (Frota et al., 2015).

https://doi.org/10.1017/langcog.2024.16 Published online by Cambridge University Press

The fact that the wh-exclamation intonation contour is more challenging to recognize also shows the relevance of visual cues when the auditory cues are weaker or less clear. Massaro and Cohen (1983) stated not only that the integration of auditory and visual cues is part of speech perception but also that the visual channel is a reliable source when the utterances are ambiguous and when the auditory channel is degraded. In the case of the two speech acts analyzed in this paper, the final nuclear contour has a similar falling F0 configuration (Moraes, 2008; Oliveira et al., 2014; Zendron da Cunha, 2016), differently from the study of Miranda et al. (2021), which compared the bimodal perception of BP assertions and echo questions, two speech acts with different pitch accents in the nuclear position of the intonational contour. Hence, the analysis of the interaction between the shape of F0 contours and the use of visual cues is also relevant.

Wagner et al. (2015) affirmed that using different approaches in the analysis of speech seems to reinforce the main findings of the area. In this study, mismatched artificial stimuli were created. In addition, the use of controlled sentences in the production of speech acts that are not so frequent, such as wh-exclamations, is also warranted, because a robust analysis of this type of speech act would require a great amount of recording. Compared with the results of Miranda et al. (2019), the perceptual experiment applied in the present study with congruent and incongruent stimuli corroborates that facial gestures contribute to the perception of pragmatic meanings of intonation. As already mentioned in the introduction, in Miranda et al. (2019), when the visual cues were added to the auditory cues in the audiovisual condition, participants' perception of the wh-question and wh-exclamation intonation significantly increased compared to the monomodal conditions (audio-only and video-only). Hence, both studies brought evidence that language processing is multimodal (Holler et al., 2018; McNeill, 1992), which means that the recognition of the pragmatic function of the intonation is enhanced not only by adding the visual cues to the auditory cues compared to audio alone but also by presenting congruent stimuli of the speech acts. In the incongruent condition, the perception of the speech acts was hindered. Therefore, the pragmatic contribution of the visual cues in the perception of speech acts supports the theories of multimodal language (Barsalou, 2008; Kendon, 2004; McNeill, 1992).

Furthermore, in BP, it is necessary to further analyze the timing of the facial gestures within the utterance production of wh-questions and wh-exclamations as well as other speech acts, such as assertions and echo questions. Nota et al. (2021) described a corpus of dyadic Dutch face-to-face conversation to analyze which facial gestures are displayed in the production of questions and statements as well as the timing of these gestures within both speech acts. The results of the analysis showed that the facial gestures were more likely to occur with questions than assertions, but also that these gestures occurred earlier in the production of questions compared to assertions. The authors explained that the fact that questions are more visually marked than assertions is related to the processing of language in the conversation since the co-speech gestures play an important role in the temporal coordination of turn-taking, in which visual cues facilitate the comprehension of the speaker's speech act, so that the listener can also prepare his/her answer early. It is worth remembering that, regarding answering a question in conversation, a gap longer than the average may signal a dispreferred answer (Kendrick & Torreira, 2015). In other words, the listener must process the speaker's intended speech acts faster to provide an appropriate and timely answer in a conversational interaction, and the visual cues

https://doi.org/10.1017/langcog.2024.16 Published online by Cambridge University Press

contribute to the early recognition of the speaker's communicative intentions in this context.

As already presented in the introduction of this article, several uses of gestures contribute to expressing the speakers' communicative intentions. In the present study, multimodal communication (Barsalou, 2008; Kendon, 2004; McNeill, 1992) is also verified in the expression of pragmatic functions of the intonation, such as speech acts. Nonetheless, the limitations of this study are related to the perceptual analysis of multimodal speech without measuring reaction time and to the use of a limited number of speakers. Besides, analyses involving more naturalistic settings are needed to make proper generalizations about the use of facial gestures in the transmission of the pragmatic meanings investigated in this paper.

6. Conclusion

This study has shown that the visual channel facilitates the comprehension of the intonation of Brazilian Portuguese wh-questions and wh-exclamations speech acts, since the perceptual identification of the stimuli in the congruent condition (i.e., the auditory and visual cues from the same speech act) was higher than that in the incongruent condition (i.e., the auditory and visual cues from different speech acts). In this mismatched condition, the perception of the speech acts was significantly decreased for most of the recorded speakers. The results of the perceptual experiment not only confirm the findings of previous studies for the congruent condition (Borràs-Comes & Prieto, 2011; Cruz et al., 2017; Kelly et al., 2015; Miranda et al., 2021; Swerts & Krahmer, 2004), but also show that, in the incongruent condition, the visual channel affects the interpretation of the speech act, having a considerable weight in comparison with the auditory channel.

Based on this outcome, we can conclude that the perception fusion of the auditory and visual modalities in identifying the pragmatic function of intonation was verified, given that the type of visual input added to the auditory channel impacted the processing of the pragmatic intonation meaning. Also, in the mismatched context, listeners responded differently in the interpretation of the artificial stimuli, depending on the speakers' performance: the more visual cues were employed by the speakers, the more the perception of the speech acts in the mismatched condition was hindered. Another outcome verified in this article is the fact that speakers differ about their use of visual cues in producing these speech acts, which is also in line with previous studies (Keating et al., 2003), as well as the use of auditory cues (Cruz et al., 2015).

Further investigation of the speech acts analyzed in this paper is needed regarding the production and perception of Brazilian Portuguese wh-questions and wh-exclamations in a naturalistic setting of language use, such as in spontaneous speech and conversational interaction.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/langcog.2024.16.

Data availability statement. The data that support the findings of this study are available from the corresponding author (L.d.S.M.), upon reasonable request.

Acknowledgments. The authors are grateful for Dr. Pilar Prieto and for the three anonymous reviewers' valuable comments on the earlier version of the manuscript. Part of this research has been funded by the

scholarship 88882.331896/2015-01 (awarded by the Brazilian Federal Agency for Support and Evaluation of Graduate Education–CAPES) during the first author's Ph.D. course at Federal University of Rio de Janeiro, Brazil.

Competing interest. The authors declare none.

References

Argyle, M., & Cook, M. (1976). *Gaze and mutual Gaze*. Cambridge University Press.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Barbosa, P. (2006). *Incursões em torno do ritmo da fala*. Pontes.

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. https://doi.org/10.1177/0023830909348993

Barsalou, L. (2008). Grounded cognition. *Annual Review of Psychology*, 59, 617–645.

Batliner, A. (1988). Der Exklamativ: Mehr als Aussage oder doch nur mehr oder weniger Aussage? Experimente zur Rolle von Hohe und Position des F0-Gipfels. In H. Altmann (Ed.), *Intonationsforschungen* (pp. 243–271). Niemeyer.

Bavelas, J. B., & Chovil, N. (2018). Some pragmatic functions of conversational facial gestures. *Gesture*, 17(1), 98–127.

Benitez-Quiroz, C. F., Wilbur, R. B., & Martinez, A. M. (2016). The not face: A grammaticalization of facial expressions of emotion. *Cognition*, 150, 77–84.

Beyssade, C., Delais-Roussarie, E., & Marandin, J. M. (2007). The prosody of interrogatives in French. *Nouveaux Cahiers de Linguistique Française*, 28, 163–175.

Bolinger, D. (1989). *Intonation and its uses: Melody in grammar and discourse*. Stanford University Press.

Borràs-Comes, J., Kaland, C., Prieto, P., & Swerts, M. (2014). Audiovisual correlates of interrogativity: A comparative analysis of Catalan and Dutch. *Journal of Nonverbal Behavior*, 38, 53–66.

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

Brinckmann, A., & Benzmuller, R. (1999). The relationship between utterance type and F0 contour in German. In *Proceedings of the sixth European conference on speech communication and technology (Eurospeech 1999), Budapest, Hungary* (pp. 21–24), ESCA. https://www.isca-speech.org/archive/pdfs/eurospeech_1999/brinckmann99_eurospeech.pdf.

Caucci, G. M., & Kreuz, R. J. (2012). Social and paralinguistic cues to sarcasm. *Humor*, 25, 1–22.

Crawley, M. J. (2013). *The R book* (2nd ed.). John Wiley & Sons. https://doi.org/10.1002/9781118448908

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. (2015). Variation in tone and gesture within language. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of eighteenth international congress of phonetic sciences, Glasgow, UK*, 452, 1–5. The University of Glasgow. http://www.icphs2015.info/pdfs/Papers/ICPHS0452.pdf.

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.

Cruz, M., Swerts, M., & Frota, S. (2019). Do visual cues to interrogativity vary between language modalities? Evidence from spoken Portuguese and Portuguese Sign Language. In Calhoun, S., Escudero, P., Tabain M. & Warren, P. (Eds.), *Proceedings of AVSP 2019 – International Conference on Auditory-Visual Speech Processing, August 10–11, 2019, Melbourne, Australia*. Australasian Speech Science and Technology Association Inc.

D'Eugenio, A. (1976). The intonation systems of Italian and English. *Rassegna Italiana di Linguistica Applicata*, 8(1), 57–85.

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

Delattre, P. (1966). Les 10 intonations de base du français. *The French Review*, 40(1), 1–14.

De Rosario-Martinez, H. (2015). phia: Post-Hoc Interaction Analysis. R package version 0.2-1. Retrieved from https://CRAN.R-project.org/package=phia

Ekman, P., Friesen W. V., & Hager J. C. (2002). *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*. San Francisco, CA: Consulting Psychologists Press.

Falé, I. (2006). *Percepção e reconhecimento da informação entoacional em Português Europeu* [Doctoral dissertation, University of Lisbon].

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

Flecha-García, M. L. (2010). Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in English. *Speech Communication*, 52, 542–554.

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

Gili Fivela, B. (2015). L'integrazione di informazioni multimodali: prosodia ed espressioni del volto nella percezione del parlato. In E. Pistolesi, R. Pugliese, & B. Gili Fivela (Eds.), *Parole, gesti, interpretazioni: Studi linguistici per Carla Bazzanella* (pp. 107–127). Aracne.

Gomes da Silva, C. (2019). *A prosódia de atos de fala no espanhol da Cidade do México*. [Doctoral dissertation, Federal University of Rio de Janeiro].

González-Fuente, S., Escandell-Vidal, V., & Prieto, P. (2015). Gestural codas pave the way to the understanding of verbal irony. *Journal of Pragmatics*, 90, 26–47

Gries, S. T. (2013). *Statistics for linguistics with R: A practical introduction* (2nd ed.). De Gruyter Mouton. https://doi.org/10.1515/9783110307474

Gyuris, B., & Mády, K. (2013). Approaching the prosody of Hungarian wh-exclamatives. In Szigetvári, P. (Ed.), *VLLXX: Papers presented to László Varga on his 70th birthday*. Eötvös Loránd University. http://seas3.elte.hu/VLlxx/gyuris-mady.html

Holler, J., Kendrick, K. H., & Levinson, S. C. (2018). Processing language in face-to-face conversation: Questions with gestures get faster responses. *Psychonomic Bulletin & Review*, 25, 1900–1908.

Holler, J., & Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends in Cognitive Sciences*, 23(8), 639–652. https://doi.org/10.1016/j.tics.2019.05.006

House, D. (2002). Intonational and visual cues in the perception of interrogative mode in Swedish. In *Proceedings of seventh international conference on spoken language processing* (pp. 1957–1960). Causal Productions.

Keating, P., Baroni, M., Mattys, S., Scarborough, R., Alwan, A., Auer, E., & Bernstein, L. (2003). Optical phonetics and visual perception of lexical and phrasal stress in English. In *Proceedings of the fifteenth international conference on phonetic sciences, August 3–9, 2003, Barcelona, Spain* (pp. 1–4). Universitat Autónoma de Barcelona.

Kelly, S., Healey, M., Özyürek, A., & Holler, J. (2015). The processing of speech, gesture, and action during language comprehension. *Psychonomic Bulletin & Review*, 22, 517–523. https://doi.org/10.3758/s13423-014-0681-7

Kelly, S. D. (2001). Broadening the units of analysis in communication: Speech and nonverbal behaviours in pragmatic comprehension. *Journal of Child Language*, 28, 325–349. https://doi.org/10.1017/S0305000901004664

Kelly, S. D., Iverson, J. M., Terranova, J., Niego, J., Hopkins, M., & Goldsmith, L. (2002). Putting language back in the body: Speech and gesture on three time frames. *Developmental Neuropsychology*, 22(1), 323–349. https://doi.org/10.1207/S15326942dn2201_1

Kendon A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In: M. R. Key (Ed.), *The relationship of verbal and nonverbal communication* (pp. 207–227). Mouton.

Kendon, A. (2004). *Gesture. Visible action as utterance*. Cambridge University Press.

Kendrick, K. H., & Torreira, F. (2015). The timing and construction of preference: A quantitative study. *Discourse Processes*, 52, 255–289.

Kendrick, K. H., & Holler, J. (2017). Gaze direction signals response preference in conversation. *Research on Language and Social Interaction*, 50(1), 12–32. https://doi.org/10.1080/08351813.2017.1262120

Levinson, S. C., & Holler, J. (2014). The origin of multi-modal communication. *Philosophical Transactions of the Royal Society B*, 369, 20130302. http://doi.org/10.1098/rstb.2013.0302

Lüdecke, D., Ben-Shachar, M., Patil, I., & Makowski, D. (2020). Extracting, computing and exploring the parameters of statistical models using R. *Journal of Open Source Software*, 5(53), 2445. https://doi.org/10.21105/joss.02445

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.

Mata, A. I. (1990). *Questões de entoação e interrogação em português. Isso é uma Pergunta* [Master's thesis, University of Lisbon].

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

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago University Press.

McNeill, D., Cassell, J., & McCullough, K. E. (1994). Communicative effects of speech-mismatched gestures. *Research on Language and Social Interaction*, 27(3), 223–237.

Miranda, L., Gomes da Silva, C., Moraes, J. A., & Rilliard, A. (2020a). Visual and auditory cues of assertions and questions in Brazilian Portuguese and Mexican Spanish: A comparative study. *Journal of Speech Sciences*, 9, 73–92.

Miranda, L., Moraes, J. A., & Rilliard, A. (2019). Audiovisual perception of wh-questions and wh-exclamations in Brazilian Portuguese. In Calhoun, S., Escudero, P., Tabain M. & Warren, P. (Eds.) *Proceedings of the nineteenth international congress of phonetics sciences* (pp. 2941–2945), Melbourne, Australia, 2019. Australasian Speech Science and Technology Association Inc.

Miranda, L., Moraes, J. A., & Rilliard, A. (2020b). Statistical modeling of prosodic contours of four speech acts in Brazilian Portuguese. In *Proceedings of the tenth international conference on speech prosody* (pp. 404–408). University of Tokyo. https://doi.org/10.21437/SpeechProsody.2020-83

Miranda, L., Moraes, J. A., & Rilliard, A. (2022). Effects of F0 movements, intensity, and duration in the perceptual identification of Brazilian Portuguese wh-questions and wh-exclamations. *DELTA. Documentação de Estudos em Lingüística Teórica e* Aplicada, 38(3), 1–29. https://doi.org/10.1590/1678-460X202258882

Miranda, L., Moraes, J. A., & Rilliard, A. (2023). Identificação perceptiva de pistas prosódicas da asserção e da questão-eco no português brasileiro: análise por ressíntese. In: B. Urbán (Org.). *'Se mais mundo houvera, lá chegara' Tanulmányok Rákóczi István 65. Születésnapja tiszteletére* (pp. 95–124). ELTE Eötvös Kiadó.

Miranda, L., Swerts, M., Moraes, J. A., & Rilliard, A. (2021). The role of the auditory and visual modalities in the perceptual identification of Brazilian Portuguese statements and echo questions. *Language and Speech*, 64(1), 3–23. https://doi.org/10.1177/0023830919898886

Moraes, J. A. (1998). Intonation in Brazilian Portuguese. In D. Hirst, & A. Di Cristo (Eds.), *Intonation systems: A survey of twenty languages* (pp. 179–194). Cambridge University Press.

Moraes, J. A. (2008). The pitch accents in Brazilian Portuguese: Analysis by synthesis. In P. A. Barbosa, S. Madureira, & C. Reis (Eds.), *Proceedings of the fourth international conference on speech prosody, May 6–9, Campinas, São Paulo, Brazil* (pp. 389–397). LBASS.

Moraes, J. A., Miranda, L., & Rilliard, A. (2012). Facial gestures in the expression of prosodic attitudes in Brazilian Portuguese. In H. Mello, M. Petrorino, & T. Raso (Eds.), *Proceedings of the VIIth GSPC international conference: Speech and corpora (GSCP 2011)* (pp. 157–161). Firenze University Press.

Moraes, J. A., & Rilliard, A. (2018). Describing the intonation of speech acts in Brazilian Portuguese: Methodological aspects.In I. Feldhausen, J. Fliessbach, & M. M. Vanrell (Eds.), *Methods in prosody: A Romance language perspective* (pp. 229–262). Language Science Press. https://doi.org/10.5281/zenodo.1441347

Moraes, J. A., Rilliard, A., Mota, B., & Shochi, T. (2010). Multimodal perception and production of attitudinal meaning in Brazilian Portuguese. In *Proceedings of the fifth international conference on speech prosody, May 10–14, Chicago, Illinois, USA* (pp. 1–4). University of Illinois.

Nota, N., Trujillo, J. P., & Holler, J. (2021). Facial signals and social actions in multimodal face-to-face interaction. *Brain Sciences*, 11, 1017. https://doi.org/10.3390/brainsci11081017

O'Connor, J. D., & Gordon, F. A. (1961). *Intonation of colloquial English: A practical handbook*. Longman.

Oliveira, J., Pacheco, V., & Oliveira, M. (2014). Análise perceptual das frases exclamativas e interrogativas realizadas por falantes de Vitória da Conquista/BA. *Signum: Estudos Linguísticos*, 17(2), 354–388. https://doi.org/10.5433/2237-4876.2014v17n2p354

Paiva, F. A. S., Martino, J. M., Barbosa, P. A., Benetti, A., & Silva, I. R. (2016). Um sistema de transcrição para língua de sinais brasileiras: O caso de um avatar. *Revista do Gel*, 13(3), 12–48.

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. In: S. M. Alvord (Ed.). *Selected proceedings of the 5th conference on laboratory approaches to romance phonology (LARP)* (pp. 136–141). Cascadilla Proceedings Project.

Repp, S. (2015). On the acoustics of wh-exclamatives and wh-interrogatives: Effects of information structure and sex of speaker. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the eighteenth international conference of phonetic sciences* (pp. 1–5), August 10–14, 2015. The University of Glasgow.

Repp, S. (2020). The prosody of wh-exclamatives and wh-questions in German: Speech acts differences, information structure and sex of speaker. *Language and Speech*, 63(2), 306–361. https://doi.org/10.1177/0023830919846147

Rohloff, M., & Michalsky, J. (2018). Pitch scaling as a question cue in German wh-questions. In M. Belz, C. Mooshammer, S. Fuchs, S. Jannedy, O. Rasskazova, & M. Żygis (Eds.). *Proceedings of the conference on phonetics & phonology in German-speaking countries (P&P 13)* (pp. 169–172). Leibniz-Zentrum Allgemeine Sprachwissenschaft & Humboldt-Universität zu Berlin.

Searle, J. (1969). *Speech acts*. Cambridge University Press.

Sorianello, P. (2011). Aspetti prosodici e pragmatici dell'atto esclamativo. *Studi Linguistici e Filologici Online*, 9, 287–332.

Sorianello, P. (2012). A prosodic account of Italian exclamative sentences: A gating test. In *Proceedings of the sixth international conference on speech prosody, May 22–25, Shanghai, China* (pp. 298–301). Tongji University Press.

Spence, C. (2007). Audiovisual multisensory integration. *Acoustical Science and Technology*, 28(2), 61–70.

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

Stoet, G. (2010). PsyToolkit - A software package for programming psychological experiments using Linux. *Behavior Research Methods*, 42(4), 1096–1104.

Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, 44 (1), 24–31.

Swerts, M., & Krahmer, E. (2004). Congruent and incongruent audiovisual cues to prominence. In B. Bel, & I. Marlien (Eds.), *International conference on speech prosody 2004, March 23–26, Nara, Japan* (pp. 69–72).

Swerts, M., & Krahmer, E. (2005). Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, 53(1), 81–94.

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

Vegas Pro software *MAGIX*. (2016). Version 14 of Vegas Pro. https://www.vegascreativesoftware.com/

Wagner, P., Trouvain, J., & Zimmerer, F. (2015). In defense of stylistic diversity in speech research. *Journal of Phonetics*, 48, 1–12.

Zendron da Cunha, K. (2016). *Sentenças exclamativas em português brasileiro: um estudo experimental de interface* [Doctoral dissertation, Federal University of Santa Catarina]. http://fonapli.paginas.ufsc.br/files/2017/06/Tese_KarinaZendrondaCunha1.pdf

Cite this article: Miranda, L. S., de Moraes, J. A., & Rilliard, A. (2024). Visual channel facilitates the comprehension of the intonation of Brazilian Portuguese wh-questions and wh-exclamations: evidence from congruent and incongruent stimuli, *Language and Cognition*, 1–21. https://doi.org/10.1017/langcog.2024.16