AUDIOVISUAL PERCEPTION OF WH-QUESTIONS AND WH-EXCLAMATIONS IN BRAZILIAN PORTUGUESE

Luma da Silva Miranda[1], *João Antônio de Moraes*[1], *Albert Rilliard*[1,2]

[1]Federal University of Rio de Janeiro, Brazil
[2]LIMSI, CNRS, Université Paris Saclay, France
lumah.miranda@gmail.com, jamoraes3@gmail.com, albert.rilliard@limsi.fr

ABSTRACT

This paper examines auditory and visual cues used for the discrimination of wh-question and wh-exclamative speech acts in Brazilian Portuguese. The sentence *Como você sabe* was uttered as a wh-question (meaning *How do you know?*) and as a wh-exclamation (meaning *How clever you are!*) by ten Brazilian Portuguese speakers (five males and five females) from Rio de Janeiro. The acoustic and visual analyses revealed that these two speech acts not only showed different F0 contours and intensity patterns, but also discriminant facial expressions. A perceptual experiment that investigates the role of visual *versus* audio channels with three presentation conditions (audio only, video only and audiovisual) was applied with sixty Brazilian participants (twenty per condition). The results indicate that listeners rely on both channels to perceive the wh-questions and wh-exclamations and that the audiovisual condition was more accurately recognized than the monomodal ones.

Keywords: Audiovisual perception, wh-questions, wh-exclamations, Brazilian Portuguese.

1. INTRODUCTION

In the last few decades, the amount of research on prosody that explores the relationship between visual and auditory channels has increased significantly. Given the multimodal nature of speech production [9], listeners can recognize speech functions through acoustic and visual cues. Several prosodic aspects in speech production such as prominence [19], focus [5,10] and the discrimination of utterance types – statements and questions– [8,18] have already been investigated in bimodal conditions in multiple languages. These studies provided evidence to support the fact that auditory and visual information is integrated into speech perception [18], though the visual contribution can be relatively small for the perceptual recognition of prosodic linguistic meaning in comparison to expressive prosody, such as the manifestation of emotions [3] and attitudes [13].

It is worth mentioning that the studies, which investigated the role of the auditory and the visual channels in the communication of utterance types, usually address the distinction between statements and yes-no questions. The present work expands this analysis to two different speech acts [16] – wh-questions and wh-exclamations–, utterances constructed with a WH-word in the initial position of the clause, allowing a syntactic parallelism for the analysis of intonational contours. Thus, these sentences have the same morphosyntactic and lexical forms, yet they can still be distinguished by their intonational contour.

The description of the wh-question intonational contour in Brazilian Portuguese (BP, henceforth) shows an overall falling F0 movement that extends to the nuclear region. This pattern was verified consistently in various dialects [7]. On the other hand, as for wh-exclamations, there exist Brazilian varieties described with either a nuclear falling F0 movement [14,21] or a nuclear slightly convex F0 movement [12]. In addition, an experimental study [11] that explored these two intonational contours with speakers from Rio de Janeiro using the IPO stylization methods [20] showed that a falling F0 movement on the nuclear syllable of wh-questions and a slight rising F0 on the final stressed syllable of wh-exclamations are perceptually relevant.

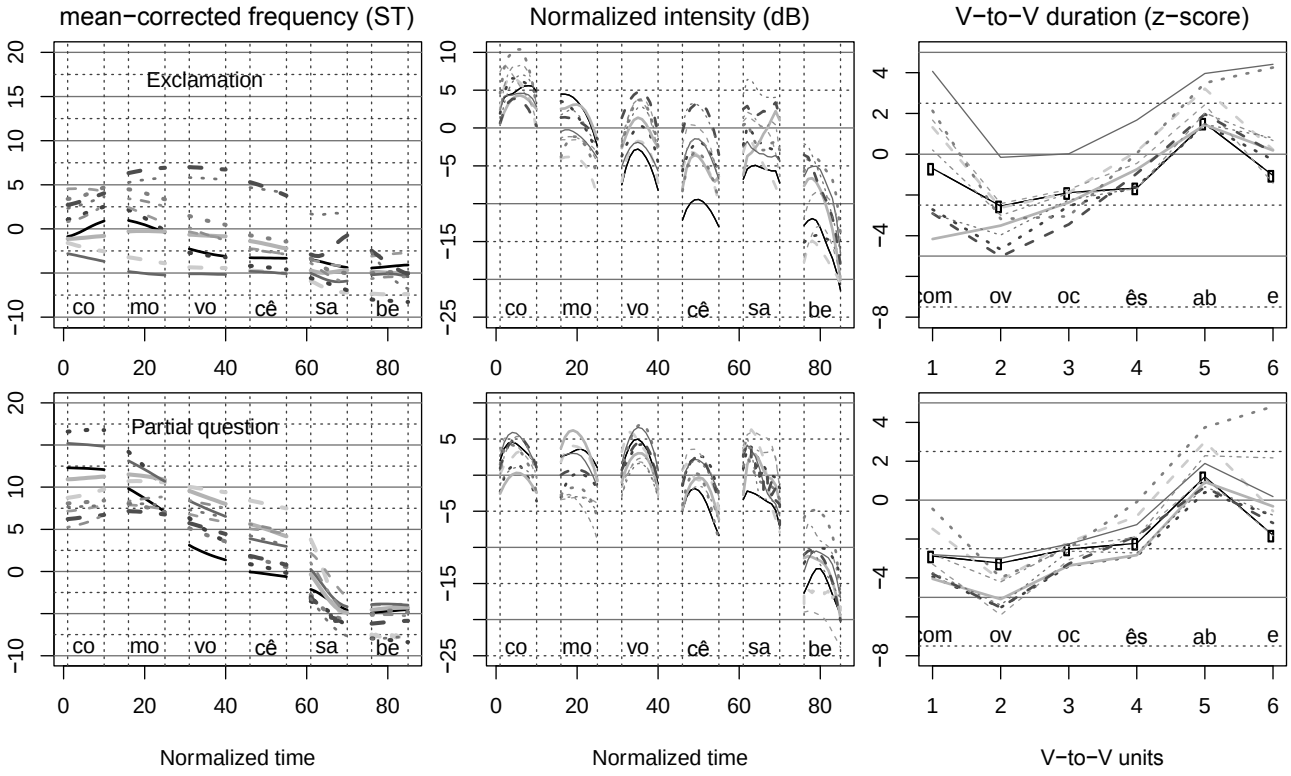
The purpose of this study is to investigate the perceptual recognition of Brazilian Portuguese wh-questions and wh-exclamations through the auditory and visual channels. We hypothesize that adding the visual channel will help listeners to discriminate these two intonational contours, since listeners can integrate information from both channels [18].

2. METHOD

2.1. Data collection

The sentence *Como você sabe* was produced as a wh-question (meaning *How do you know?*) and a wh-exclamation (meaning *How clever you are!*) by ten native BP speakers (five males and five females) from Rio de Janeiro. All the speakers were undergraduate or graduate students of the Federal University of Rio de Janeiro (UFRJ, henceforth), except for speaker 3, who is a professor and one of the authors. Speakers' ages ranged from 19 to 64 years old at the time of the recording session.

Figure 1: Upper row of the panel: speakers' normalized F0 curves, normalized intensity curves and Z-score duration curves for wh-exclamations (Exclamation). Bottom row of the panel: speakers' normalized F0 curves, normalized intensity curves and Z-score duration curves for wh-questions (Partial question).



2.2. Procedure

Before the audiovisual recordings at UFRJ, participants were asked to sign a consent form. The recording sessions took place in a sound-attenuated room at the Acoustic Phonetics Laboratory at UFRJ. SONY NEX-F3 camera was positioned 90 cm far away from the speaker, who was seated on a chair against a dark background. The camera was adjusted to film the front view of the speaker from below the neck to above the top of the head. Sound was recorded with Zoom H4 audio recorder positioned in front of the speaker (about of 20 cm away) outside the field view of the camera. Each speaker followed the instruction to produce the sentence types indicated by the experimenter, who was in the room during the whole recording session. Each speaker repeated the wh-question and wh-exclamation sentences ten times, resulting in 200 utterances collected for the acoustic analysis. Afterwards, we used VEGAS PRO [17] software to synchronize the audio from the recorder with the video from the camera. Then, videoclips of the eighth and ninth repetitions of each sentence type were cut into two-second clips using the same software. Forty utterances were selected, allowing a randomization of the video material that was used for the visual analysis and the perceptual experiment.

2.3. Data acoustic analysis

The F0, intensity and duration measures were extracted automatically from the 200 utterances that were manually segmented at the phoneme level through the application of two Praat scripts. The first one [1] allowed the creation of a time-normalized F0 contour in order to compare the different F0 prominences, eliminating the influences of the microprosody on the F0 contours. Regularly spaced vectors of estimated values of F0 and intensity were taken (10 samples of the vowel sounds and 5 samples of the consonant sounds). This way, we obtained F0 and intensity curves on a normalized time basis. Moreover, the information about the segmental duration was grouped in units V-to-V and submitted to a normalization in order to subtract differences of intrinsic and cointrinsic duration to each phoneme. By using the SG-Detector script [2], the raw duration values of our data were transformed in smoothed z-score values.

Comparing the normalized F0 curves of wh-exclamations (in the left of the upper row) and wh-questions (in the left of the bottom row) of Figure 1, the wh-question presents a higher melodic level in the wh-word and a falling F0 movement throughout the contour with a steeper slope in the nuclear region, whereas the wh-exclamation contour starts in a lower F0 level followed by a F0 fall that is smoother in the nuclear syllable. As for the normalized intensity

patterns of wh-exclamations (in the middle of the upper row) and wh-questions (in the middle of the bottom row) of Figure 1, the clear similarity between the two sentences is the drop of the intensity in the final post-stressed syllable ('be'), which is lower in the wh-question contour. On the other hand, in the wh-exclamation, the intensity of the first stressed syllable ('co') is higher than the one found in the wh-question. Regarding the duration patterns of wh-exclamations (in the right of the upper row) and wh-questions (in the right of the bottom row) of Figure 1, the values are presented in a z-score distribution. The main difference between these contours is found in the prenuclear region in which the first stressed syllable of the wh-question is located two standard deviations below the mean, whereas in the wh-exclamation, this syllable is between one standard deviations above the mean and two standard deviation below the mean, which indicates more variation compared to the wh-question. Both contours showed a similar duration distribution in the nuclear region of the sentence.

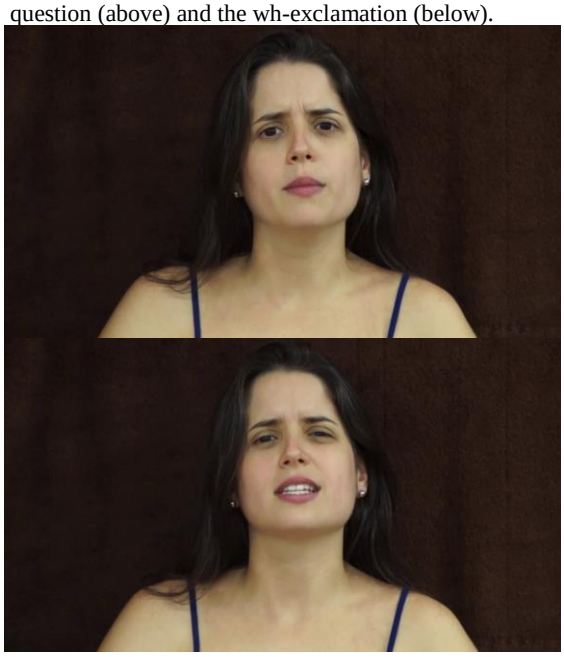
The phonetic description of our data set indicates that wh-questions and wh-exclamations have different behaviours and that they are best distinguished by the F0 movements plus the intensity patterns.

2.4 Data visual analysis

Since the audiovisual material of our work recorded the front view of the speaker, the FACS (Facial Action Coding System) manual [6] was used to describe the momentary changes in each speakers' face while he produced the wh-question and wh-exclamation intonational contours. Based on a prior visual analysis [13], the following 20 Action Units (AU, henceforth) were selected for a restricted annotation of the facial expressions: 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 puller (AU 12), lip corner depressor (AU 15), chin raiser (AU 17), lips part (AU 25), 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), head forward (AU 57), head back (AU 58) and, finally, up and down head movement (AU 85).

Two of the authors annotated the AUs of forty videos (two repetitions of the two sentence types produced by 10 speakers) during the production of the speech acts. In order to measure the agreement between the two authors, Cohen's kappa was measured and showed a significant agreement (k=0.56). The analysis revealed that the visual cues that distinguish questions from exclamations are the combination of the eyebrow lowerer (AU 4) along with head turning right (AU 52) in the production of wh-questions, whereas raising the lip corner (AU 12), parting the lips (AU 25) and moving the head up and down (AU 85) are relevant AUs for wh-exclamations. The main difference between these two speech acts is related to the eyebrow lowering in the wh-question production and the smile observed in the wh-exclamation, as shown in Figure 2:

Figure 2: Stills of a female speaker reproducing the wh-question (above) and the wh-exclamation (below).



Based on the acoustic and visual analysis, we expect that Brazilian listeners would rely on these cues to recognize wh-questions and wh-exclamations both auditorily and visually.

3. PERCEPTUAL EXPERIMENT

The goal of the perceptual experiment was to verify the relative importance of the visual channel in relation to the auditory channel in signalling the distinction between the intonational contours of wh-questions and wh-exclamations in Brazilian Portuguese.

3.1. Participants

In the perceptual experiment, 60 Brazilian graduate and undergraduate students took the survey on the computers of the Laboratory of Acoustic Phonetics at UFRJ. For each of the three experimental conditions, a new group of listeners was recruited; hence there were 20 participants per condition. Their mean age was 21.2 years old. All participants were native BP speakers with no hearing or sight deficits. In addition,

the auditory conditions of the experiment required the use of a headset.

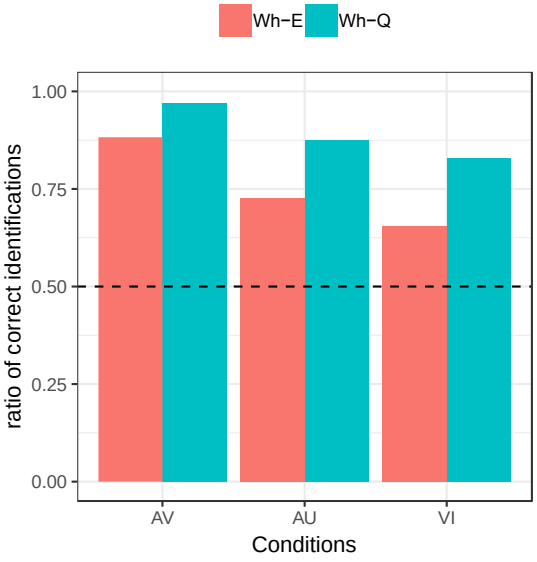
3.2. Stimuli and procedure

Taken from the visual data previously mentioned, 40 videoclips were presented in three experimental conditions in the perceptual test: audio only, video only and audiovisual, which resulted in 120 stimuli (40 stimuli per condition). The Qualtrics platform [15] was used to set up the experiment as well as to present the instructions and collect the answers. On the Qualtrics platform, the stimuli were presented one by one, randomly in an experiment run, for each participant. Participants had to decide, for each stimulus, whether the sentence *Como você sabe* was expressed as a question or as an exclamation. Subjects knew beforehand they would see, hear, or see and hear speakers producing a wh-question or a wh-exclamation. The progression was shown in a bar located at the top the screen. There was a set of 40 trials for each experimental condition. A typical experiment run lasted 10 to 15 minutes.

3.3. Results

Figure 3 gives the outcome of the perceptual experiment which shows the recognition of wh-questions and wh-exclamations above the chance level (50%) for the three experimental conditions:

Figure 3: Mean ratio of good recognition of wh-exclamations (Wh-E) and wh-questions (Wh-Q) obtained in each presentation modality: audiovisual (AV), audio only (AU) and video only (VI). The horizontal dashed bar represents the level of chance guessing (50%) between both answers.



The results were expressed as "false" or "good" answers, depending on whether the subject managed to identify the intended speech act (wh-question or wh-exclamation) based on the presented stimulus. The answers were aggregated across the ten speakers in order to obtain a success ratio, for each of the three presentation modalities (MOD: audiovisual, audio only and video only) and each speech act (ACT: wh-questions/wh-exclamations) and their interaction. This ratio served as the dependant variable and was analysed using a logistic regression with quasibinomial error in order to deal with overdispersion in residuals based on two categorical explanatory variables (MOD and ACT).

Both the modality and the speech act factors had a significant impact on the way listeners judged the speech act of the sentence (Act: $F_{(1,116)} = 36.51$, $p<0.001$; Mod: $F_{(2,117)} = 23.81$, $p<0.001$). A *post-hoc* Tukey test was applied to the modality factor to check for differences between levels. The results demonstrated that the audiovisual (AV) modality received significantly higher recognition levels than the audio (AU) and the visual (VI) presentations that do not significantly differ. The logistic regression showed that the interaction between both factors (MOD and ACT) is not significant ($F_{(2,114)} = 0.49$, $p=0.61$).

4. CONCLUSION

The results of our study indicated that wh-questions and wh-exclamations in Brazilian Portuguese can be distinguished both auditorily and visually. The acoustic description showed that, in the prenuclear region of the contours, there is a higher F0 attack in the wh-questions, and, in the nuclear region, especially in the stressed syllable, wh-questions present a steep falling F0 movement, whereas in the wh-exclamations, this steep falling F0 is slighter. In addition, specific intensity patterns were found for both intonational contours. In the visual analysis, Brazilian speakers used different Action Units to express these speech acts (lowering the eyebrow and turning the head right for wh-questions and raising the lips corner plus moving the head up and down for wh-exclamations). Finally, this work has shown that Brazilian listeners also rely on the visual channel to interpret the intonational contours of wh-questions and wh-exclamations. The results of the perceptual experiment can be summarized as follows: the audiovisual (AV) condition presented a higher overall recognition rate compared to the auditory (AU) and visual (VI) conditions. Hence, the current study confirms that adding visual information (facial gestures) improves the recognition of wh-questions and wh-exclamations intonational contours. These results provide evidence to support the fact that the distinction of utterance types in speech perception is multimodal [4].

5. REFERENCES

[1] Arantes, P. 2015. Time-Normalization of Fundamental Frequency Contours: A Hands-On Tutorial. In: Meireles, A. (Org.). *Courses on Speech Prosody*. 1ed. Newcastle upon Tyne: Cambridge Scholars Publishing, v. 1, 98-123.

[2] Barbosa, P. A. 2013. Semi-automatic and automatic tools for generating prosodic descriptors for prosody research. *Proc. TRASP* Aix-en-Provence, 86-89.

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

[4] Boersma, P., Weenink, D. 2013. *Praat*: doing phonetics by computer [Computer program]. (Version 5.1.05), http://www.praat.org/.

[5] Dohen, E., Loevenbruck, H. 2009. Interaction of audition and vision for the perception of prosodic contrastive focus. *Language and Speech* 52, 177-206.

[6] Ekman, P., Friesen, W., Hager, J. 2002. *Facial Action Coding System*. Salt Lake City, UT: A Human Face.

[7] Frota, S., Cruz, M., Svartman, F. R. F., Collischonn, G., Fonseca, A., Serra, C. R., Oliveira, P., Vigário, M. 2015. Intonational variation in Portuguese: European and Brazilian varieties. In: Frota, S., Prieto, P. (Org.). *Intonation in Romance*. 1ed. Oxford: Oxford University Press, v. 1, 235-283.

[8] House, D. 2002. Intonation and visual cues in the perception of interrogative mode in Swedish. *Proc. 7th ICSLP* Denver, 1957–1960.

[9] Kendon, A. 1980. *Gesture: Visible Action as Utterance*. Cambridge, Cambridge University Press.

[10] Krahmer, E., Ruttkay, Z., Wesselink, W., Swerts, M. 2002. Pitch, eyebrows and the perception of focus. *Proc. 1st SP* Aix-en Provence, 443-446.

[11] Miranda, L. S. 2015. Análise da entoação do português do Brasil segundo o modelo IPO. Rio de Janeiro: Federal University of Rio de Janeiro. Master thesis, UFRJ.

[12] Moraes, J. 2008. The pitch accents in Brazilian Portuguese: analysis by synthesis. *Proc. 4th SP* Campinas, 389-397.

[13] Moraes, J., Miranda, L. S., Rilliard, A. 2012. Facial gestures in the expression of prosodic attitudes of Brazilian Portuguese. *Proc. 7th GSCP* Belo Horizonte, 157-161.

[14] 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.

[15] Qualtrics Experience Management Platform. 2018. Qualtrics. Online platform: https://www.qualtrics.com/.

[16] Searle, J. R. 1969. *Speech Acts*. Cambridge: Cambridge University Press.

[17] Sony Network Entertainment International, 2014. Sony Vegas Pro. Version 14. [http://www.vegascreativesoftware.com/us/vegas-pro/]

[18] 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.

[19] Swerts, M., Krahmer, E. 2006. The importance of different facial areas for signaling visual prominence. *Proc. 9th ICSLP* Pittsburgh, 1280–1283.

[20] 't Hart, J., Collier, R., Cohen, A. 1990. *A Perceptual Study of Intonation*: *An experimental-phonetic approach to speech melody*. Cambridge: Cambridge University Press.

[21] Zendron da Cunha, K. 2016. Sentenças exclamativas em português brasileiro: um estudo experimental de interface. Florianópolis: Federal University of Santa Catarina. PhD dissertation, UFSC.