



HAL
open science

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

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

► **To cite this version:**

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

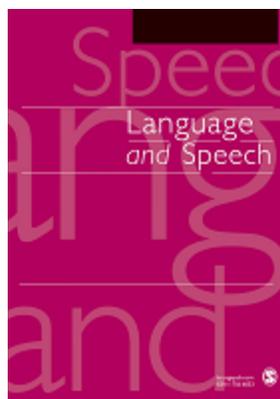
HAL Id: hal-02456308

<https://hal.science/hal-02456308>

Submitted on 27 Aug 2021

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

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



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

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

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.

- Fig. 3-Assertion production of a female speaker.mp4
- Fig. 3-Echo question production of a female speaker.mp4
- Fig. 4-Echo question production of a male speaker.mp4
- Assertion with SNR of 0dB produced by a female speaker.mp4
- Echo question with SNR of 0dB produced by a female speaker.mp4
- Assertion with SNR of -6dB produced by a female speaker.mp4
- Echo question with SNR of -6dB produced by a female speaker.mp4



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

Luma Miranda¹, Marc Swerts², João A. Moraes¹, Albert Rilliard^{3,1}

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

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

³LIMSI, CNRS, Université Paris Saclay, France

Abstract

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

1
2
3 **Keywords:** Audiovisual perception. Prosody. Statement. Echo question. Brazilian
4
5 Portuguese.
6
7
8
9

10 **Introduction**

11
12
13
14
15 When people process auditory information, like incoming speech, they tend to also
16
17 rely on visual input, such as facial expressions a speaker produces while talking. That the
18
19 visual and auditory channels can have a strong impact on speech perception has been
20
21 shown in the well-known McGurk effect (McGurk & McDonald, 1976): that is, the
22
23 presentation to participants of incongruent auditory and facial cues about specific
24
25 segments tends to lead to an illusory percept that integrates cues from the two
26
27 modalities. Visual speech (lipreading) has also been studied as a source of information to
28
29 help listeners with cochlear implants comprehend speech, while more recent studies have
30
31 demonstrated that the visual channel is beneficial for listeners with normal hearing as
32
33 well (Rosenblum, 2005).
34
35
36

37
38 Previous studies in this area of research focused on the recognition of individual
39
40 speech sounds, but there is a growing awareness that auditory and visual cues may also
41
42 interact at higher levels. For instance, the speaker's face has been documented as a
43
44 reliable indicator of emotions and attitudes in speech (Barkhuysen, Krahmer and Swerts,
45
46 2010; Crespo Sendra, Kaland, Swerts and Prieto, 2013; Ekman, Friesen and Hager, 2002;
47
48 Moraes, Miranda and Rilliard, 2012). Studies have also shown that the visual channel can
49
50 serve as a marker of communicatively relevant information that is often associated with
51
52 prosodic features. The visual channel may also signal prominence (Hadar, Steiner, Grant,
53
54 and Clifford Rose, 1983; Krahmer & Swerts, 2007; Swerts & Krahmer, 2008), that is the
55
56 marking of some words as being more salient than others in an utterance, as well as signal
57
58
59
60

1
2
3 the focus of a sentence (Dohen & Loevenbruck, 2009; Krahmer, Ruttkay, Swerts and
4 Wesselink, 2002). In this way, visual cues may signal the information status of sentence
5 elements, for example, to distinguish broad from contrastive focus (e.g., see Lambrecht,
6 1994, for a discussion). Visual cues are also used for the identification of statements
7 versus questions (Borràs-Comes & Prieto, 2011; Cruz, Swerts and Frota, 2017; House,
8 2002; Nicholson, Baum, Kilgour, Koh, Munhall and Cuddy, 2003; Srinivasan & Massaro,
9 2003). In other words, given the multimodal nature of speech production, listeners
10 perceive linguistic functions through the combination of acoustic and visual cues, both at
11 segmental and suprasegmental levels of speech (Gili Fivela, 2018).

12
13
14
15
16
17
18
19
20
21
22
23
24 In this paper, the bimodal perception of statements and echo questions in Brazilian
25 Portuguese (henceforth, BP) is investigated. The choice of these two sentence types is
26 motivated by the fact that (i) in BP the same syntactic structure is used for both sentence
27 modes, which makes it a function prominently carried out by prosodic cues and (ii) few
28 studies have analyzed the prosodic production of echo questions. This paper is based on
29 insights from previous studies that have investigated the distinction between statements
30 and questions in different languages. These studies are consistent in claiming that,
31 although the auditory signal tends to be a more informative resource to determine the
32 pragmatic status of an utterance, listeners also extract information from the visual channel
33 to identify sentence types.

34
35
36
37
38
39
40
41
42
43
44
45
46
47 A number of studies experimentally investigated audiovisual processing of sentence
48 types. House (2002) analysed the visual recognition of questions in Swedish, using an
49 animated talking head whose visual cues were systematically manipulated to convey
50 statements and questions in an interaction with humans. The experiments showed that
51 listeners relied more on intonational than on visual cues for the identification of questions.
52 For American English, Srinivasan and Massaro (2003) ran a set of experiments with both
53
54
55
56
57
58
59
60

1
2
3 natural and virtual speakers based on one sentence produced as a statement or as an echo
4
5 question; the lexico-syntactic structure of the sentence was kept constant in order to make
6
7 sure that listeners did not rely on possible syntactic or semantic cues. Similarly to the
8
9 Swedish study, the authors concluded that the intonation was more informative than the
10
11 facial expressions; yet, in the final remarks of their work, they suggested that a larger
12
13 *corpus* with natural speakers should be used to obtain additional auditory and visual cues
14
15 that were absent in their experiments.
16
17

18
19 The integration of the auditory and visual cues in speech is also supported by
20
21 experiments measuring reaction times. In European Portuguese, Cruz et al. (2017)
22
23 compared the production and perception of statements and yes-no questions. The results
24
25 showed that reaction times of the identification task were lower in the audiovisual
26
27 condition compared to the monomodal conditions, including the auditory condition.
28
29

30
31 Borràs-Comes and Prieto (2011) analysed the audiovisual recognition of echo
32
33 questions and statements with contrastive focus using a McGurk style paradigm
34
35 (presenting incongruent auditory and visual cues) and collecting reaction times from
36
37 Catalan listeners. Identification was faster when participants were exposed to congruent
38
39 stimuli, supporting the bimodal integration of speech perception. The study also
40
41 concluded that the visual channel was predominant over intonational cues for Catalan
42
43 listeners' decisions. In their discussion, the authors stress the importance of the
44
45 integration of visual and intonational cues for an accurate and rapid interpretation of these
46
47 prosodic meanings.
48
49

50
51 The exceptionally high level of recognition of visual cues in Borràs-Comes and
52
53 Prieto's (2011) study may be linked to the fact that the authors selected stimuli that "best
54
55 characterized the contrast" (Borràs-Comes & Prieto, 2011, p. 362) between the sentence
56
57 types upon which they focused. While this result provides insight into the integration of
58
59
60

1
2
3 audiovisual cues for sentence type identification, it may not really be representative as it
4 makes use of very prototypical and possibly exaggerated expressions that may not
5 generalize to more natural data. It would thus be useful to investigate to what extent
6 similar findings could be obtained with recordings of more natural and spontaneous
7 occurrences of the facial gestures accompanying the production of echo questions and
8 statements with contrastive focus.
9

10
11 Most of the previously mentioned research reveals that visual cues have some effect
12 on sentence type classification, but that their contribution is relatively small when
13 compared to auditory cues. This raises the question of how the effect would play out for
14 a language as BP, which is peculiar in that it does not mark yes-no questions by syntactic
15 means (e.g., word order), but only by intonation. Moraes (1998) showed that the
16 perceptually pertinent opposition is made by means of intonation, with a fundamental
17 frequency (F0) change on the final pre-stressed and stressed syllables of a sentence
18 nucleus used to mark assertive (F0 rise on the pre-stressed syllable followed by a fall on
19 the stressed one) and interrogative (F0 fall on the pre-stressed syllable followed by a rise
20 on the stressed one) meanings. Moraes (2008) described nuclear pitch accents for these
21 two modes, following the Autosegmental-Metric notation (Pierrehumbert, 1980), as
22 H+L*L% for statements and L+<H*L% for yes-no questions.
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43

44 As mentioned before, instead of neutral yes-no questions, which are information-
45 seeking questions, that is, the speaker asks for an information unknown to him (Frota et
46 al., 2015), we decided to analyze echo questions, because this type of yes-no question is
47 underexplored in BP. According to Frota et al. (2015), the pragmatic function of echo-
48 questions is twofold: to trigger a yes-no answer, just like the neutral yes-no question, and
49 to signal a lack of comprehension of a previous utterance. When speakers produce echo
50 questions, they repeat the previous utterance in order to verify if it was correctly
51
52
53
54
55
56
57
58
59
60

1
2
3 understood. In European and Brazilian Portuguese varieties, a similar nuclear rising F0
4 for both types of questions was found (Frota et al., 2015). So, given the attested use of
5 intonation patterns for distinguishing statements and echo questions in BP, it remains to
6 be seen whether, in addition to these prosodic cues, and to what extent BP speakers may
7 rely on facial cues for such an identification task.
8
9

10
11
12 In a pilot project, Peres, Raposo de Medeiros, Ferreira Netto and Baia (2011)
13 analysed the role of the visual channel for the identification of statements and yes-no
14 questions in BP. Their first experiment utilized a McGurk set-up whereby they created
15 stimuli by combining the audio from one source with the video from another source. A
16 second experiment proposed a video-only condition in which listeners had to assess
17 muted versions of recorded clips. The first experiment indicated a dominance of
18 intonational information over visual information to identify statements and questions,
19 while the second showed that yes-no questions and statements could be recognized by
20 Brazilian participants through the visual channel only. Still, the number of participants of
21 this study was limited to ten listeners, for all the experiments. Hence, the authors'
22 conclusions about the contribution of each channel for the perception of statements and
23 questions in BP may be considered preliminary.
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41

42 The current study extends the analysis of the prosodic and visual cues that have
43 been proposed to be valuable for the identification between statements and questions in
44 BP. In comparison with prior work, we increased the number of speakers, participants
45 and presentation conditions of audiovisual stimuli in order to obtain a better
46 understanding of the multimodal perception process involved in such a task. As it was
47 our aim to compare the relative capacities of prosodic and visual cues in conveying this
48 distinction, we choose to keep the lexical and syntactic material constant, despite a
49 possible bias, or lack of variability, this may introduce. Note that we decided to include
50
51
52
53
54
55
56
57
58
59
60

1
2
3 stimuli from multiple speakers to make sure that the perceptual results had general and
4
5 ecological validity, as opposed to previous studies that were based on evaluations of a
6
7 schematic or highly stylized animated talking head with specific auditory and visual
8
9 parameter settings. Additionally, we semi-randomly selected speakers' productions from
10
11 their ten realizations, without explicitly choosing their "best performances".
12
13

14
15 The phonetic analysis of this study served to describe relevant prosodic cues that
16
17 can be also used to identify the statements and echo questions in the initial and final
18
19 regions of the intonational contours regarding not only the F0, but also the intensity and
20
21 duration parameters. A visual analysis of the recorded data was conducted in order to
22
23 identify and categorize prototypical gestures that were related to the production of each
24
25 sentence type in BP and which listeners could recognize in perceptual tasks.
26
27

28
29 In addition, considering that the intonational cues for these contours in BP are
30
31 known to be highly intelligible, the role of the visual channel was also explored through
32
33 the identification tasks of statements and questions using degraded audio conditions. As
34
35 information from the face is known to improve speech intelligibility in noise (Benoît &
36
37 Le Goff, 1998; Ouni, Cohen, Ishak and Massaro, 2006; Sumby & Pollack, 1954), the
38
39 acoustic signal of the recorded data was degraded with two levels of Signal-to-Noise
40
41 Ratio (SNR), using a noise based on a multitalker babbling speech recording.
42
43

44
45 The main goal of this paper is to understand the contribution of the speaker's face
46
47 in the perceptual integration of auditory and visual channels for the identification of
48
49 statements and echo questions in Brazilian Portuguese. This study aims to investigate the
50
51 following research questions:
52

- 53
54 (i) To which extent the visual cues—alone or combined with auditory cues—
55
56 allow for the correct identification of statements and questions? On the
57
58 basis of the literature on BP, we may hypothesize a higher recognition
59
60

1
2
3 rate in the bimodal presentation condition as well as a lower identification
4 ratio for visual cues than for auditory ones, but still above chance.
5

6
7
8 (ii) What is the extent of the influence of the visual channel when the auditory
9 information is impoverished? Based on prior studies, we expect a higher
10 contribution of the visual channel in this condition.
11

12
13
14 (iii) Which channel will lead the subjects' interpretation when the voice and
15 the face are signaling different information? Here again, a dominant role
16 of auditory cues is foreseen.
17
18
19

20
21 The article is divided into the following sections: we first describe how we obtained
22 audiovisual recordings and how we analysed the prosodic and visual properties of these
23 data. Then, the designs of Experiments 1 and 2 are explained followed by a combined
24 statistical analysis of the results of both tests. Next, the setup of Experiment 3 is presented
25 along with the statistical analysis, combining the results of Experiments 1 and 3. Finally,
26 the outcomes of the three experiments are discussed and followed by a general
27 conclusion.
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42

43 **Methods**

44 **Structure of the *corpus* and participants**

45
46 We invited ten native speakers of BP (five male) to produce different versions of the
47 sentence “*Como você sabe*”, both as an assertion (with the meaning “*As you know.*”) or
48 as an echo question (with the meaning “*As you know?*”, which can also be rephrased as
49 “*Is it ‘as I know’ that you are asking me?*”). Note that the sentence does not contain any
50 lexical or morphosyntactic indices related to the mode; studies so far (Frota et al., 2015;
51
52
53
54
55
56
57
58
59
60

1
2
3 Moraes, 1998, 2008) have shown that this pragmatic distinction can be carried out by
4
5 intonation-only cues.
6

7
8 Speakers having a formal training in language sciences were recruited, to maximize
9
10 the chances that they would understand the communication situations linked to each
11
12 sentence mode. Although it may induce interpretative behaviors that differ from their
13
14 natural ones, we think that the fact of recording ten speakers would neutralize most of
15
16 these drawbacks. All speakers were graduate and undergraduate students at the Federal
17
18 University of Rio de Janeiro (UFRJ, henceforth), except for one speaker who is a
19
20 professor at UFRJ and one of the authors. All participants are speakers of Rio de Janeiro
21
22 variety of BP. They were aged between 19 and 65 years old at the time of the recording
23
24 session.
25
26
27
28
29
30

31 **Recording procedure**

32
33 Speakers were asked to read and sign a consent form in which the audiovisual
34
35 recording process and the future use of the data were explained. Speakers were seated
36
37 against a dark background in a sound-attenuated room at the Acoustic Phonetics
38
39 Laboratory of UFRJ. None of them wore glasses. They were recorded using a SONY
40
41 NEX-F3 video camera, placed 90 cm from the speakers. The camera was positioned to
42
43 frame the upper body and face. To increase the quality of the acoustic signal, a Zoom H4
44
45 recorder was used to jointly record each voice, with a microphone located 20 cm from the
46
47 speaker's mouth, outside the frame of the camera. Although the speakers are not actors,
48
49 they were told to express themselves as naturally as possible. The experimenter was in
50
51 the room during the whole recording session, instructing the speaker about the sentence
52
53 types. Each subject repeated the sentence ten times, alternating statements and echo
54
55 questions. This process resulted in 200 recorded utterances. After the recording sessions,
56
57
58
59
60

1
2
3 the audio waves were synchronized with the video in the software Vegas Pro 14.0 (Magix,
4
5 2014), using a handclap produced at the beginning of each recording session.
6
7
8
9

10 **Acoustic analysis**

11
12
13 The 200 recorded sentences were manually segmented at the phonemic level using the
14
15 Praat software (Boersma & Weenink, 2016). Prosodic correlates of fundamental
16
17 frequency (F0, expressed in semitones (ST) relative to 1 Hz, using Praat's default
18
19 algorithm), intensity (expressed in dB) and syllabic duration (in seconds) were estimated
20
21 from the acoustic signal. Some pre-processing of the raw measurements was made: a
22
23 Praat script produced time normalized F0 contours (Arantes, 2015) that allowed the
24
25 comparison of F0 prominences between speakers and modes. The same script was
26
27 adapted to time normalize the intensity measures. Another Praat script was used for the
28
29 duration analysis by transforming the raw phonemic duration values in a smoothed z-
30
31 score values that removed intrinsic differences in phoneme duration (Barbosa, 2013).
32
33
34

35
36 The results of these measurements indicate that the assertive and interrogative contours
37
38 bear distinct F0 movements. These distinguishing patterns are quite consistently
39
40 expressed across our recordings: to implement the declarative contour, all speakers
41
42 produce a rising F0 movement from the beginning of the contour until the nuclear pre-
43
44 stressed syllable ("cê"), followed by a fall on the stressed one ("sa"), whereas the echo
45
46 question is characterized by a low F0 from the prenuclear region of the contour until the
47
48 pre-stressed syllable ("cê"), followed by a rise on the stressed ("sa") and a fall on the
49
50 post-stressed syllable ("be"), as illustrated in Fig. 1.
51
52

53
54 [Insert Figure 1]
55
56

57 The patterns presented in Fig. 1 are consistent with the descriptions in the literature on
58
59 BP intonation (Frota et al., 2015; Miranda, 2015; Moraes, 2008). Moreover, it can be
60

1
2
3 noticed that the final pre-stressed syllable (“cê”) in the declarative contour is high, whilst
4
5 the same syllable is low in the interrogative contour, marking the contrast between the
6
7 final pre-stressed syllables (“cê”) in both contours. This behavior of the final pre-stressed
8
9 syllable (“cê”) is corroborated by Couto, Silva and Miranda (2017) who analysed the
10
11 declaratives and interrogative contours in three BP varieties—Rio de Janeiro, Fortaleza
12
13 and Salvador. This difference between the nuclear region of the contours supports
14
15 Moraes’s descriptions (1998, 2008) that can lead in representing the pitch accents of
16
17 statements as H+L* L% and those of questions as L+<H*L%, as already stated in the
18
19 introduction of this study.
20
21
22

23
24 The acoustic analysis also shows that the intonational contours of both the statement
25
26 and the echo question are accompanied by specific intensity patterns and similar duration
27
28 patterns (cf. Fig. 2).
29

30
31 [Insert Figure 2]
32

33 Regarding intensity, the main differences between statements and echo questions are
34
35 found around the tonic syllable (“sa”), with (i) a stronger pre-tonic syllable (“cê”) and
36
37 (ii) a weaker post-tonic syllable (“be”) for the assertive contour, as compared to the
38
39 interrogative one. The tonic syllable (“sa”) is the stronger and the longer in both cases,
40
41 showing comparable values across modes. In addition, based on the bottom panel of Fig.
42
43 2, it is possible to observe an increasing syllabic duration along sentences for both modes,
44
45 up to the tonic syllable (“sa”) of the nucleus.
46
47

48
49 The analysis of the prosodic parameters shows that there are clear prosodic features
50
51 discriminating statements from echo questions in BP, most clearly in terms of their F0
52
53 contours.
54
55

56 57 58 **Facial movements analysis** 59 60

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

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

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

1
2
3 head forward (AU 57), head back (AU 58) and, finally, up and down head movement
4 (AU 85). Only the presence of AUs was recorded in this study; intensity was not
5
6 measured.
7
8
9

10 Two of the authors independently analysed forty videos, corresponding to two
11 repetitions of statements and echo questions produced by ten speakers, and annotated the
12 facial movements in terms of AUs. The selected stimuli correspond to those used in the
13 perceptual experiments described hereafter. A moderate agreement between the two
14 annotators was reached, as measured by Cohen's Kappa ($\kappa = 0.488$). Although
15 significantly above 0 (i.e., no agreement above chance level), this kappa is described as
16 "moderate" by Landis and Koch (1977). Thus, to maintain the possibility of accounting
17 for possible divergences between the two annotators, the AU analysis was made keeping
18 the annotations of both authors separated. A correspondence analysis was run on the count
19 of AUs observed by speech act, speaker and annotator. The analysis extracts the main
20 dimensions that explain attribution of AUs to the stimuli. The ten first dimensions, which
21 explain about 85% of the variance, were selected to run an algorithm of hierarchical
22 clustering which was used to group together stimuli that have similar characteristics
23 (Husson, Lê, Pagès, 2017). The cluster analysis proposed a three-cluster solution.
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41

42 Cluster I is composed of stimuli showing AU 4 (brow lowerer) and AU 7 (lid
43 tightener) and expressing echo question. Cluster II contains AU 1 (inner brow raiser), AU
44 2 (outer brow raiser) and AU 55 (head tilt left) associated with the production of
45 assertions. An example of these two facial expressions can be seen in Fig. 3.
46
47
48
49
50

51 [Insert Figure 3]
52

53 Cluster III links AU 17 (chin raiser) and AU 57 (head forward) to the production
54 of echo question by speaker 7. These last two AUs may be considered part of this
55 speaker's idiosyncrasy and are illustrated in Fig. 4.
56
57
58
59
60

[Insert Figure 4]

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

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

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

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

1
2
3 to note that this difference in the facial performance can be related to the speech acts
4 explored in both studies, since echo-questions were analyzed in BP and neutral yes-no
5 questions in EP. Although an echo-question is also considered a type of yes-no question
6 (Frota et al., 2015), as stated in the introduction of this study, the visual production of
7 echo-questions may express a component of doubt or uncertainty, which was shown in
8 Fig. 4 with the facial expression of cluster III.
9

10
11
12 Summarizing the prosodic and visual descriptions, brow lowering along with
13 tightening of the eyes is associated with the echo questions that present a rising nuclear
14 pitch accent and, in terms of intensity prominence, a weaker final pre-stressed syllable
15 (“cê”) as well as a stronger final post-stressed syllable (“be”). The eyebrow raising plus
16 head movement is related to the assertion contour with a falling pitch accent and a
17 stronger final pre-stressed syllable (“cê”) plus a weaker final post-stressed syllable (“be”).
18 The final stressed syllable (“sa”) is the strongest in relation to intensity and the longest in
19 both contours.
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39

40 **Perceptual experiments**

41
42
43 In this section, the described prosodic and visual cues related respectively to statements
44 and echo questions are perceptually validated. The design of Experiments 1 and 2 are
45 presented separately and the results of both are commented together. Afterwards, the
46 setup of Experiment 3 is exposed along with the outcome that was also combined with
47 Experiment 1.
48
49
50
51
52
53
54
55
56

57 **Experiment 1: Audiovisual recognition with clear audio**

58
59
60

1
2
3 The first experiment was designed to verify the relative importance of visual cues in
4 comparison to the auditory ones as signals of the statement and echo question distinction
5 in Brazilian Portuguese.
6
7
8
9

10 11 12 *Participants*

13
14 Experiment 1 was applied to 64 participants (21 in the audiovisual, 22 in the video-
15 only and 21 in the audio-only condition) with mean age of 29,1 years old. Listeners
16 originated from various regions of Brazil, though predominantly from Rio de Janeiro.
17 Although the majority of participants were either students or professors in Linguistics at
18 UFRJ, some of them were working or studying in different scientific fields (e.g.,
19 Philosophy and Engineering). They participated in the experiment through a web
20 interface, using either their personal computer or one provided by the Acoustic Phonetics
21 Lab at UFRJ. For each presentation condition, a new group of listeners was recruited.
22 None reported hearing or sight impairments. Use of headset was required for the auditory
23 conditions implying audio presentations.
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39

40 41 *Stimuli*

42 The eighth and ninth repetitions of the assertion and echo question produced by all
43 speakers were selected. This way, we gained a randomization of the video material.
44 Unlike Borràs-Comes and Prieto (2011), we did not select the visual data qualitatively.
45 In other words, we did not select the most expressive productions from each speaker to
46 avoid potentially over-acted data. The video clips were cut in the software Vegas Pro so
47 that their length was approximately two seconds. It is worth noting that the production of
48 the whole utterance contained in the clip was often preceded or followed by a short silent
49 pause. In total, forty utterances were selected for the perceptual experiments.
50
51
52
53
54
55
56
57
58
59
60

Procedure

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

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

Experiment 2: Audiovisual recognition with degraded audio

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

Participants

1
2
3 Experiment 2 involves 43 participants (21 in the audiovisual and 22 in the audio-only
4 condition) with mean age of 24 years old. Differently from the previous experiment,
5 listeners were only from Rio de Janeiro. Participants used a computer of the Acoustic
6 Phonetics Lab at UFRJ to complete the perceptual task. They were either students or
7 professors of Linguistics at UFRJ. Since listeners were not allowed to participate in
8 multiple experiments, for each presentation condition, a new group of listeners was
9 recruited. None of the participants reported hearing or sight impairments. Again, the use
10 of headset was required for the auditory conditions of the experiment.
11
12
13
14
15
16
17
18
19
20
21
22

23 *Stimuli*

24
25
26 The video material of Experiment 2 is the same used in Experiment 1, except that the
27 auditory channel is degraded by mixing it with a multi-speaker babble noise taken from
28 Steeneken & Geurtsen (1988). Note that the speakers were not recorded in noisy
29 conditions. In this study, we are interested in noise for its capacity to degrade the given
30 prosodic cues, so to measure the relative power of visual cues. Recording the speakers in
31 noise would probably change their communication strategies as a consequence of the
32 Lombard effects, leading to an increased F0, duration and energy as well as bigger
33 motions of the face and the head (Fitzpatrick, Kim and Davis, 2015; Kim, Sironic and
34 Davis, 2011), which is not our aim here.
35
36
37
38
39
40
41
42
43
44
45

46
47 Regarding the addition of noise in the recording material, Cooke (2006) recommends
48 a fixed signal-to-noise ratio (SNR, henceforth) as being more effective when multi-
49 speaker babble noise rather than a single competing talker is employed. Furthermore, the
50 expected limits of SNR for individuals with normal hearing suggested by a study with
51 Brazilian listeners (Costa, Daniel and dos Santos, 2011) showed a range of the SNR that
52 goes from $-3,03$ dB to $-7,55$ dB with a mean of $-5,29$ dB. Based on the latter study and
53
54
55
56
57
58
59
60

1
2
3 after a pilot experiment using SNR of +3 dB, 0 dB and -6 dB, we selected the last two
4
5 levels of SNR (0 dB and -6 dB) , and applied the continuous multi-speaker babble noise
6
7 to the original utterance: a favorable SNR of 0 dB and an adverse SNR of -6 dB.
8
9

10 The degraded audio was produced in Praat and later combined again within the original
11
12 videos in Vegas Pro software. To keep the task feasible, only one repetition from each
13
14 speaker was used, thus presenting forty stimuli with degraded audio to listeners. In the
15
16 supplementary materials of this paper, four samples of the audiovisual stimuli with
17
18 degraded audio can be accessed (two statements and echo questions with SNR of 0 dB
19
20 and -6 dB produced by one speaker). It is worth remembering that the segmental material
21
22 did not change in these experiments: listeners had to decode prosody, a different type of
23
24 information than the segmental one studied in literature using such masking noise.
25
26
27
28
29

30 *Procedure*

31
32 The second perceptual experiment followed the same procedure as described in
33
34 Experiment 1. The only difference is that Experiment 2 presented forty stimuli with two
35
36 SNR levels (0 dB and -6 dB) in only two conditions: audiovisual and audio-only
37
38 conditions, each condition presented to a different group of participants as well. This
39
40 experiment also had a between-subject design, avoiding potential learning effects for the
41
42 same reason presented in Experiment 1. Each group of subjects was presented with forty
43
44 trials and, just like the previous experiment, the remaining number of trials was always
45
46 indicated at the top of the screen. Participants took fifteen minutes to complete the
47
48 perceptual task as well.
49
50
51
52
53

54 *Results of Experiments 1 and 2*

55
56
57
58
59
60

1
2
3 Results of Experiment 2 were analyzed along with Experiment 1. Subjects' responses
4 were expressed as *success* or *failure* to identify the presented mode (assertive or
5 interrogative). The proportion of success was expressed for each of the seven presentation
6 conditions in Experiments 1 and 2 (audio-only/AO; video-only/VI; audiovisual/AV;
7 audio-only in 0 dB SNR/AO-0; audio-only in -6 dB SNR/AO-6; audiovisual in 0 dB
8 SNR/AV-0; audiovisual in -6 dB SNR/AV-6—see Fig. 5 that shows this ratio in each
9 condition as well as each type of sentence, assertive or interrogative). The ratio of success
10 was used as the dependent variable for a logistic regression model (quasibinomial errors
11 were used to deal with overdispersion in binomial model residuals), using two categorical
12 explanatory factors (*Condition* and *Type*, and their interaction) as predictors.
13
14
15
16
17
18
19
20
21
22
23
24
25

26 The logistic regression shows that the interaction between both factors is not
27 significant ($F_{(6,290)} = 1.6, p = 0.15$), while both *Condition* and *Type* factors do have a
28 significant impact on the accuracy of judging the sentence mode (Type: $F_{(1,290)} = 9.28, p$
29 < 0.01 ; Condition: $F_{(6,290)} = 10.14, p < 0.01$). A *post-hoc* Tukey test was applied to the
30 *Condition* factor so to test for differences between levels of the two factors, showing that
31 three presentation conditions (video-only/VI, audio-only in 0 dB/AO-0 and audio-only in
32 -6 dB/AO-6) do significantly degrade the listener's identification performances
33 compared to the best presentation condition: audio-only (AO). A comparison of
34 participants' performance in each experimental condition of Experiments 1 and 2 is
35 shown in Fig. 5:
36
37
38
39
40
41
42
43
44
45
46
47
48

49 [Insert Fig. 5]
50

51 The results show a group of presentation conditions containing the AO, AV, AV-0 and
52 AV-6 conditions with homogeneous and high identification capabilities. The audio
53 presentation conveys identification information about sentence mode: the AO condition
54 reaches a high level of correct answers (90% of correct answers, with a confidence
55 interval that nears a perfect score)—thus there is a strong suspicion of ceiling effect in this
56
57
58
59
60

1
2
3 case. The fact that the AV, AV-0 and AV-6 conditions did not significantly differ from
4 the AO condition also advocates for a more general "ceiling effect" for all these
5 conditions: they shall thus all be regarded as presenting comparably high recognition
6 scores. Bimodal presentations (AV) achieve a comparable performance level, with non-
7 significant variations (that thus may be related to, e.g., subjects sampling). The bimodal
8 presentations with degraded audio (AV-0, AV-6) also reach a level of performance
9 comparable to the audio-only and audiovisual conditions. On the contrary, conditions
10 based on visual information alone (VI) and degraded audio alone (AO-0, AO-6) reach a
11 significantly lower identification level. This result supports the view of a cross-modal
12 fusion of information to the interpretation of sentence mode.
13
14
15
16
17
18
19
20
21
22
23
24

25
26 Regarding the factor *Type*, assertions presented an identification accuracy ratio about
27 six percent higher than echo questions, which was a significant improvement. The fact
28 that there is no interaction between the *Type* and *Condition* factors indicates a general
29 tendency of subjects to prefer the "assertion" answer in any presentation condition. One
30 may speculate that assertion functions as a default answer when no clear or interpretable
31 information is available for participants.
32
33
34
35
36
37
38
39
40
41

42 **Experiment 3: Congruent and incongruent modalities**

43
44
45 The third experiment presented congruent and incongruent audio and visual
46 modalities to evaluate which channel exerted a dominant influence on listeners'
47 identification of sentence type.
48
49
50
51

52 *Participants*

53
54
55 Experiment 3 was applied to 24 participants with mean age of 29,4 years old. Once
56 again, listeners were either students or professors of Linguistics from UFRJ. They
57
58
59
60

1
2
3 participated in the experiment through a web interface in the computer of the Acoustic
4
5 Phonetics Lab at UFRJ. The participants do not report any hearing or sight impairments
6
7 from the participants. The use of headset was also required for the audiovisual
8
9 presentation of this experiment.
10
11
12
13

14 *Stimuli*

15
16 The set of 40 videos (2 repetitions x 2 sentence types x 10 speakers) was used to mix
17
18 modalities from different stimuli. Using the Vegas Pro software, the audio utterances
19
20 were matched with video materials from other recordings. The congruent videos were
21
22 based on the audio of the first repetition synchronized with the video of the second
23
24 repetition of the same mode; that is, they consisted of both an audio and video channels
25
26 that cued the same sentence type. Incongruent videos are based on the audio of the first
27
28 repetition of one mode matched with the video from the second repetition of the other
29
30 mode, so that there was a mismatch between what the audio and video signals would cue.
31
32 We decided to construct mixed stimuli for both the congruent and incongruent cases in
33
34 order to generate a total set of data that were artificially constructed. Using this procedure,
35
36 we obtained forty artificial videos, twenty with congruent auditory and visual streams,
37
38 twenty with incongruent streams.
39
40
41
42
43
44
45
46

47 *Procedure*

48
49 The same procedure used in Experiments 1 and 2 was applied in the third experiment,
50
51 but here all participants would hear and see the videoclips. Experiment 3 presented forty
52
53 stimuli, containing either congruent or incongruent auditory and visual information
54
55 (twenty stimuli for each of these two types). The congruent audiovisual condition (AVc)
56
57 is in principle similar to the audiovisual presentation (AV) of Experiment 1, except that
58
59
60

1
2
3 the audio and visual channels were taken from two different repetitions of the same
4 speaker for the same targeted sentence mode. The incongruent condition (AVi) consisted
5 of combinations of audio and visual modalities that were retracted from distinct sentence
6 modes (e.g., the auditory information from an echo question paired with the visual
7 information from a statement). The forty stimuli were presented in audiovisual condition
8 to one group of 24 participants (within-subjects design). There were forty trials in this
9 experiment and, just like the previous perceptual tests, the remaining number of trials was
10 also indicated at the top of the screen. A typical run also lasted fifteen minutes.
11
12
13
14
15
16
17
18
19
20
21
22
23

24 *Results*

25
26 Results of Experiment 3 are analysed in relation to the outcome of Experiment 1.
27 Answers were expressed as *success* or *failure* to correctly identify the assertive or
28 interrogative mode of the acoustic part of the stimuli (because in incongruent
29 presentations, the visual information signals the alternative category). Results for the two
30 conditions of Experiment 3 are presented in Fig. 6. The proportion of success was
31 calculated for the five presentation conditions (audiovisual/AV, audio-only/AO and
32 video-only/VI from Experiment 1; congruent audiovisual/AVc and incongruent
33 audiovisual/AVi from Experiment 3) and each sentence type (statement or question).
34 Variations in the ratio of success according to *Condition* and *Type* of sentence factors
35 were again analysed by a logistic regression (with quasibinomial errors).
36
37
38
39
40
41
42
43
44
45
46
47
48

49 The logistic regression shows that the interaction between the two factors is not
50 significant ($F_{(4,218)} = 0.83, p = 0.51$), while the two main factors have a significant impact
51 on ratio of success (Type: $F_{(1,226)} = 8.08, p < 0.01$; Condition: $F_{(6,290)} = 8.85, p < 0.01$). A
52 *post-hoc* Tukey test was applied to *Condition* factor and shows that the video-only (VI)
53
54
55
56
57
58
59
60

1
2
3 and incongruent audiovisual (AVi) levels do significantly degrade the listener's
4 identification performances, compared to the audio-only (AO) presentations.
5
6

7
8 As shown in Fig. 6, stimuli in congruent audiovisual presentations were better
9 recognized than in audiovisual incongruent ones, which caused the identification rate of
10 the sentence types to decrease.
11
12
13

14 [Insert Figure 6]
15

16
17 Based on Experiment 3, we can conclude that the visual channel interferes in the
18 process of listeners' identification of statements and questions, although the auditory
19 component is dominant for the perceptual recognition of these sentence types. This
20 finding means that the speech perception of the intonation of sentence types is bimodal
21 even when the auditory channel is clear.
22
23
24
25
26
27
28
29
30
31
32

33 **General discussion** 34

35
36 The main goal of this study was to analyze the contribution of the visual channel, in
37 combination with the auditory one, as a possible resource to identify statements and echo
38 questions in BP. To this end, we made use of one carrier sentence produced repetitively
39 by ten speakers of the Rio de Janeiro variety of BP. We chose to use only one sentence
40 because our primary aim is to validate the relative role of the visual modality, controlling
41 for other potential variables (e.g., variation in articulation induced by different sequences
42 of phonemes). We did record various speakers with variable productions to make sure
43 that our findings have general validity; note also that these speakers produced prosodic
44 patterns that have been reported in the literature to be representative of the sentence
45 modes we explored.
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 First, the recorded data was described both acoustically and visually. The outcome of
4 the prosodic and visual descriptions brought to light relevant features that could function
5 as cues for this identification in perceptual tasks. The most relevant prosodic parameters
6 to distinguish both modes are the F0 contours, as described in the literature, combined
7 with intensity patterns. The most salient visual cues for questions were the brow lowerer
8 (AU 4) and lid tightener (AU 7), while chin raiser (AU 17) and head forward (AU 54)
9 were more consistently found for one of the speakers, which may be considered an
10 idiosyncratic performance or something unrelated to the distinction of modes under
11 investigation here. For statements, the eyebrow raising movements (AU 1 and AU 2) and
12 head movements (AU 55) turned out to be characteristic for this sentence type. Different
13 eyebrow and head movements seem to visually distinguish statements from echo
14 questions in BP.
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29

30 Next, we were interested in analyzing the perceptual identification of these two
31 sentence types to evaluate the relative strength of the described auditory and visual cues
32 when clear or impoverished audio signal were presented, as well as when both modalities
33 signaled conflicting information.
34
35
36
37
38
39

40 Experiment 1 investigated the capacity of participants to identify statements and echo
41 questions in clear audio conditions, in the three presentation conditions: audio-only (AO),
42 video-only (VI) and audiovisual (AV). A ceiling effect was observed in the results, that
43 implies identification performances in audiovisual condition (AV) that could not
44 outperform the audio-only condition (AO) in which the outcome was already quasi
45 perfect. Comparably high-performance levels were found for all the conditions containing
46 clear or degraded audio cues and coherent visual ones, as for clear audio cues only. It is
47 worth mentioning that this type of effect obtained in audio-only (AO) and audiovisual
48 (AV) presentation conditions of Experiment 1 is in line with previous studies that looked
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 into the way discourse items are signaled. For instance, as the contrastive focus in French
4 (Dohen & Loevenbruck, 2009) is accurately marked by prosodic cues, the visual cues in
5 bimodal presentations do not have an extra beneficial effect and so could not increase
6 performances that are already at a ceiling level.
7
8
9
10

11
12 One could hypothesize that the visual benefit in the interpretation of the intonation of
13 sentence types is higher when the analyzed intonational contours are similar (i.e., more
14 difficult to identify) compared to contours with clearly different F0 configurations, as in
15 our study. For instance, in Borràs-Comes and Prieto (2011), the difference between a
16 contrastive focus statement and an echo question in Catalan is given by pitch range, since
17 both contours have the same F0 contour (rising-falling nuclear configuration). In the
18 audiovisual experiments, the authors concluded that the visual cues had a stronger effect
19 than the audio cues in the perceptual identification of both utterance types. Also in
20 Catalan, Tubau, González-Fuente, Prieto and Spinal (2015) found that the identification
21 of confirmation and contradicting yes-answers, both with falling F0 patterns, was higher
22 in the audiovisual modality than in the audio-only modality. Similarly, in Miranda,
23 Moraes and Rilliard (2019), the distinction between wh-questions and wh-exclamations
24 in BP is characterized by the inclination of the falling pitch accent in the nuclear region
25 of both contours, which is steeper in wh-questions than in wh-exclamations. Results from
26 the perceptual identification test showed an improved recognition in the audiovisual
27 modality compared to monomodal ones (video-only and audio-only). All these studies
28 demonstrated that facial gestures enhanced the interpretation of the intonation of the
29 sentence types, without limitation due to a “ceiling effect”.
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53

54 In sum, results of Experiment 1 lend support to the conclusions of previous studies that
55 highlight that the same functions are performed by prosody and facial gestures, although
56 visual cues alone are less efficient than the prosodic ones (House, 2002; Srinivasan &
57
58
59
60

1
2
3 Massaro, 2003). Nonetheless, the speaker's face can still be considered a reliable source
4 of prosodic information at the sentence level. For instance, in Fisher's study (1969)
5 regarding the "visibility of final pitch" variation in English, listeners had to indicate in a
6 perceptual experiment the direction of terminal pitch contour in the production of
7 questions with either a rising or a falling contour, while watching the visual performances
8 of speakers without sound. Results showed that listeners were able to identify the
9 direction of the intonational contours based on the visual cues. Our findings also support
10 the view that participants are able to identify the intonational contours of different
11 utterance types by relying only on the speaker's face. Fisher attributes the listeners' visual
12 discrimination of final pitch contours to certain motor activities of the face such as the
13 movement of the lips. Still, he also suggests in the final remarks of the paper that the
14 whole face could be a source of information for the listener.

15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31 In Experiment 2, participants again had to identify statements and echo questions
32 but this time with noisy stimuli. Noisy audio-only presentations (audio-only in 0 dB
33 SNR/AO-0 and audio-only in -6 dB SNR/AO-6) show a significant decrease in
34 identification performances compared to the audio-only (AO) ones, but there still exist
35 some cues to identify mode, with performances comparable to video-only (VI)
36 presentations. The auditory (AO-0, AO-6) and visual (VI) channels share their respective
37 cues in the audiovisual in 0 dB SNR (AV-0) and audiovisual in -6 dB SNR (AV-6)
38 conditions to attain identification levels comparable to the ideal situation of a clear audio
39 (or audiovisual) presentation. The visual enhancement of the identification of intonational
40 contours in adverse auditory conditions found in our study supports findings of earlier
41 work on visual speech. For instance, Benoît & Le Goff (1998) and Nicholson et al. (2003)
42 state that there is a lower impact of the visual channel on speech perception compared to
43 a clear acoustic signal, but the visual speech can increase identification scores when the
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 auditory channel is impoverished. Therefore, our findings confirm that speech
4 intelligibility in a noisy environment is improved when participants have access to the
5 speaker's face.
6
7
8

9
10 The results of Experiment 2 also showed that babble noise can efficiently degrade
11 prosodic cues to sentence mode (and probably to other prosodic functions), opening a
12 possible type of evaluation for the contribution of multimodal cues to linguistic meaning
13 (cf. Dohen & Lovenbruck's, 2009, for an opposite prediction). However, since the
14 speakers of our study were not recorded in noisy conditions, their productions do not fit
15 those that would have been produced in a noisy environment. In such a situation, a
16 Lombard speech effect would probably have occurred. It has been found that in speakers'
17 recordings in adverse auditory conditions (Fitzpatrick et al., 2015; Kim et al., 2011), there
18 are differences regarding not only the prosody, such as increased energy, F0, duration and
19 flatter spectral tilt, but also the facial gestures, like bigger motions of the face and head.
20 These are different from the more neutral expressions we elicited in our set of data.
21 Meanwhile, it is not our goal here to evaluate the capacity of a speaker to communicate
22 in adverse conditions, but rather to measure the relative role of both modalities.
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39

40 From a perception point of view, Vatikiotis-Bateson, Barbosa, Chow, Oberg, Tan
41 and Yehia (2007) verified that both the auditory intelligibility and the visual enhancement
42 are greater when speech is recorded in noise compared to speech recorded in quiet and
43 then added noise to it. However, in the same study, when stimuli were presented to
44 listeners in an "easy listening level", the authors found results in which the recognition of
45 words in the condition of speech recorded in quiet with addition of noise was similar to
46 the condition of stimuli recorded in noise. Vatikiotis-Bateson et al. (2007) state that,
47 depending on the used listening levels of SNR, results can be substantially altered.
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 Although the results of Experiment 2 cannot be used to evaluate the degradation of
4 prosodic cues in noisy conditions, the findings remain interesting to show the role of
5 visual cues to enhance the audio decoding process in adverse auditory conditions from
6 the listeners' point of view. Apart from these caveats, Experiment 2 confirms that listeners
7 also rely on the speaker's face to perceive what type of sentence is being uttered even
8 when the speech recorded in quiet has noise added to it, giving support to the multimodal
9 perception of the intonation of sentence types.
10
11
12
13
14
15
16
17
18

19 The parallel processing of both modalities was shown in Experiment 3, as
20 incongruent presentations degraded the performance of a clear acoustic presentation,
21 introducing significant interferences. Congruent stimuli show accuracy ratio levels
22 comparable to the original stimuli (audio-only/AO and audiovisual/AV), showing that the
23 manipulation does not induce incoherencies. Incongruent presentations of auditory and
24 visual cues show that the visual channel has an impact in the listeners judgment, even if
25 auditory cues are still dominant for the perceptual recognition of sentence mode.
26
27
28
29
30
31
32
33
34

35 The outcome of Experiment 3 confirms earlier findings regarding the process of
36 information fusion from auditory and visual modalities to form a judgment over
37 pragmatic meanings of sentence types such as statements and echo questions. More
38 specifically, the results are in line with previous research using artificially constructed
39 stimuli in BP (Peres et al., 2011), as well as in other languages, such as Dutch (Swerts &
40 Krahmer, 2008), Catalan (Borràs-Comes & Prieto, 2011) and European Portuguese (Cruz
41 et al., 2017).
42
43
44
45
46
47
48
49
50

51 Based on these three perceptual experiments, results support previous audiovisual
52 studies that conclude that the auditory channel can convey all the necessary information
53 about sentence mode, while visual channel alone cannot. Yet, at the same time, the results
54 also show that visual cues are decoded by the participants even in the presence of clear
55
56
57
58
59
60

1
2
3 acoustic presentations, as the visual information degrades their decoding when
4 incoherent. When stimuli are presented in a noisy context, the results reveal that visual
5 cues also add robustness in the speech decoding process.
6
7
8

9
10 Therefore, these results are in line with Massaro and Cohen's assumptions (1983),
11 regarding not only the integration of auditory and visual cues as a fundamental process
12 of speech comprehension, but also regarding their viewpoint that we can expect a greater
13 contribution of the visual channel when either the verbal information is ambiguous or the
14 audio information is degraded.
15
16
17
18
19
20

21 22 23 **Conclusion**

24
25 In summary, this article shows that the auditory and visual channels are integrated in
26 the perception of sentence modes in BP. Brazilian listeners identify statements and
27 questions using both auditory and visual cues. Listeners rely more on the auditory cues
28 than the visual ones when presented under clear acoustic conditions. Contrary to our
29 expectations, the recognition rate of the sentences mode in the bimodal condition was not
30 higher than in the audio-only condition. Thus, the first hypothesis written in the
31 introduction section of this study was partially confirmed. However, the visual channel
32 has a beneficial effect when stimuli consists of degraded auditory cues, which confirms
33 the second hypothesis. Finally, although a dominant role of the auditory cues was verified
34 in the experiment with congruent and incongruent audio and visual modalities, the
35 auditory and visual channels are integrated in the perceptual identification process of
36 sentence types, confirming the third hypothesis as well.
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52

53
54 We intend to further explore other types of sentences in additional experiments, such
55 as the continuation statement, which is also an underexplored sentence type in BP. For
56 instance, Miranda et al. (2019) analysed the bimodal perception of BP wh-questions and
57
58
59
60

1
2
3 wh-exclamations. The results of the perceptual identification test confirm the integration
4 of the auditory and visual channels in signaling wh-questions and wh-exclamations
5 meanings in BP. These positive results suggest a further investigation of the bimodal
6 perception of the intonational contours of other sentence types.
7
8
9
10
11

12 Moreover, it may be possible to go further in the understanding of audiovisual
13 integration by monitoring a participant's reaction time, a measure that was not included
14 in these experimental settings. As shown in the introduction of this paper, previous studies
15 (Borràs-Comes & Prieto, 2011; Cruz et al., 2017) already indicated that listeners'
16 identification of sentence types is faster in audiovisual presentation conditions.
17 Experiments analyzing participant's reaction time to audiovisual recognition of sentence
18 types in Brazilian Portuguese are needed.
19
20
21
22
23
24
25
26
27
28
29
30
31
32

33 **Acknowledgments**

34 We wish to thank Dr. Katy Carlson and the three anonymous reviewers for their
35 thoughtful comments on an earlier version of the manuscript. This research has been
36 funded by the scholarship 88881.134778/2016-01 (awarded by the Brazilian Federal
37 Agency for Support and Evaluation of Graduate Education–CAPES) through a PhD
38 exchange program which allowed the first author to organize a research visit to Tilburg
39 University (NL).
40
41
42
43
44
45
46
47
48
49
50
51
52

53 **References**

54
55
56
57
58
59
60

1
2
3 Arantes, P. (2015). Time-normalization of fundamental frequency contours: A hands-on
4 tutorial. In: Meireles, A. (Org.). *Courses on Speech Prosody* (pp. 98–123). 1ed. Newcastle
5 upon Tyne, UK: Cambridge Scholars Publishing.
6
7

8
9
10 Barbosa, P. (2013). Semi-automatic and automatic tools for generating prosodic
11 descriptors for prosody research. *Proceedings of Interspeech satellite event Tools and*
12 *Resources for the Analysis of Speech Prosody*, Aix-en-Provence, France, pp. 86–89.
13
14
15

16
17
18 Barkhuysen, P., Krahmer, E. & Swerts, M. (2010). Cross-modal and incremental
19 perception of audiovisual cues to emotional speech. *Language and Speech*, 53 (1), 3–30.
20
21
22

23
24 Benoît, C. & Le Goff, B. (1998). Audiovisual speech synthesis from French text: Eight
25 years of models, designs and evaluation at the ICP. *Speech Communication*, 26, 117–129.
26
27
28

29
30 Boersma, P. & Weenink, D. (2016). *Praat: doing phonetics by computer*. Computer
31 program (Version 5.1.05) available at: <http://www.praat.org/>.
32
33

34
35 Borràs-Comes, J. & Prieto, P. (2011). Seeing tunes. The role of visual gestures in tune
36 interpretation. *Journal of Laboratory Phonology*, 2 (2), 355–380.
37
38
39

40
41 Cavé, C., Guaitella, I., Bertrand, R., Santi, S., Harlay, F. & Espesser, R. (1996). About
42 the relationship between eyebrow movements and F0 variations. *Proceedings of Fourth*
43 *International Conference on Spoken Language Processing*, Philadelphia, Pennsylvania,
44 USA. DOI: 10.1109/ICSLP.1996.607235
45
46
47
48

49
50 Cooke, M. (2006). A glimpsing model of speech perception in noise. *Journal of*
51 *Acoustical Society of America*, 119 (3), 1562–1573.
52
53
54
55
56
57
58
59
60

1
2
3 Costa, M. J., Daniel, R. C. & dos Santos, S. N. (2011). Reconhecimento de sentenças no
4
5 silêncio e no ruído em fones auriculares: valores de referência de normalidade. *CEFAG*,
6
7 13 (4), 685–691.
8
9

10
11 Couto, L. R., Silva, C. G. & Miranda, L. S. (2017). Prosódia dos enunciados declarativos
12
13 e interrogativos totais nas variedades de Salvador, Fortaleza e Rio de Janeiro. *Revista de*
14
15 *Estudos da Linguagem*, 25 (3), 1105–1142. DOI: [http://dx.doi.org/10.17851/2237-](http://dx.doi.org/10.17851/2237-2083.25.3.1105-1142)
16
17 2083.25.3.1105-1142.
18
19

20
21 Crespo Sendra, V., Kaland, C., Swerts, M. & Prieto, P. (2013). Perceiving incredulity:
22
23 the role of intonation and facial gestures. *Journal of Pragmatics*, 47, 1–13.
24
25

26
27 Cruz, M., Swerts, M. & Frota, S. (2017). The role of intonation and visual cues in the
28
29 perception of sentence types: evidence from European Portuguese varieties. *Laboratory*
30
31 *Phonology*, 8 (1), 23. DOI: <http://doi.org/10.5334/labphon.110>.
32
33

34
35 Cruz, M., Swerts, M. & Frota, S. (2015). Variation in tone and gesture within language.
36
37 *Proceedings of Eighteenth International Congress of Phonetic Sciences*, Glasgow, UK.
38
39 Retrieved from: [http://www.icphs2015.info/pdfs/ Papers/ICPHS0452.pdf](http://www.icphs2015.info/pdfs/Papers/ICPHS0452.pdf).
40
41

42
43 Debras, C. (2017). The shrug: forms and meanings of a compound enactment. *Gesture*,
44
45 16 (1), 1-34. DOI: <https://doi.org/10.1075/gest.16.1.01deb>
46
47

48
49 Dohen, M. & Loevenbruck, H. (2009). Interaction of audition and vision for the
50
51 perception of prosodic contrastive focus. *Language and Speech*, 52 (2–3), 177–206.
52
53

54
55 Ekman, P., Friesen, W. & Hager, J. (2002). *The Facial Action Coding System*. 2nd ed (CD-
56
57 ROM). Salt Lake City, Utah, USA: Research Nexus.
58
59
60

1
2
3 Fisher, C. G. (1969). The visibility of terminal pitch contour. *Journal of Speech and*
4
5 *Hearing Research*, 12 (2), 379–382.
6

7
8 Fitzpatrick, M., Kim, J. and Davis, C. (2015). The effect of seeing the interlocutor on
9
10 auditory and visual speech production in noise. *Speech Communication*, 74, 37–51.
11
12

13
14 Frota, S., Cruz, M., Svartman, F., Collischonn, G., Fonseca, A., Serra, C.,... & Vigário,
15
16 M. (2015). Intonational variation in Portuguese: European and Brazilian varieties. In:
17
18 Frota, S.; Prieto, P. (Org.). *Intonation in Romance* (pp. 235–283). 1ed. Oxford, UK:
19
20 Oxford University Press.
21
22

23
24 Gili Fivela, B. (2018). Multimodal analyses of audio-visual information: Some methods
25
26 and issues in prosody research. In: Feldhausen, I., Fliessbach, J. & Vanrell, M. M. (Eds.).
27
28 *Methods in Prosody: A Romance Language Perspective (Studies in Laboratory*
29
30 *Phonology 4)* (pp. 83–122). Berlin, Germany: Language Science Press.
31
32

33
34 Graf, H. P., Cosatto, E., Volker, S. & Huang, F. J. (2002). Visual prosody: facial
35
36 movements accompanying speech. *Proceedings of Fifth IEEE International Conference*
37
38 *on Automatic Face and Gesture Recognition*, Washington, D.C., USA. DOI:
39
40 10.1109/AFGR.2002.1004186
41
42

43
44 Granström, B., House, D. & Swerts, M. (2002). Multimodal feedback cues in human-
45
46 machine interaction. *Proceedings of First International Conference on Speech Prosody*,
47
48 Aix-en-Provence, France, pp. 347–350.
49
50

51
52 Hadar, U., Steiner, T. J., Grant, E. C. & Clifford Rose, F. (1983). Head movement
53
54 correlates of juncture and stress at sentence level. *Language and Speech*, 26, 117–129.
55
56

57
58 Hirst, D. & Espesser, R. (1993). Automatic modelling of fundamental frequency using a
59
60 quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix*, 15, 75–85.

1
2
3 House, D. (2002). Intonational and visual cues in the perception of interrogative mode in
4 Swedish. *Proceedings of Seventh International Conference on Spoken Language*
5 *Processing*, Denver, Colorado, USA, pp. 1957–1960.
6
7

8
9
10 Husson, F., Lê, S., & Pagès, J. (2017). *Exploratory multivariate analysis by example*
11 *using R*. Chapman and Hall/CRC.
12
13

14
15
16 Kim, J., Sironic, A. & Davis, C. (2011). Hearing speech in noise: seeing a loud talker is
17 better. *Perception*, 40 (7), 853–862.
18
19

20
21
22 Krahmer, E., Ruttkay, Z., Swerts, M. & Wesselink, W. (2002). Pitch, eyebrows and the
23 perception of focus. *Proceedings of First International Conference on Speech Prosody*,
24 Aix-en-Provence, France, pp. 443–446.
25
26

27
28
29 Krahmer, E. & Swerts, M. (2007). The effects of visual beats on prosodic prominence:
30 acoustic analyses, auditory perception and visual perception. *Journal of Memory and*
31 *Language*, 57 (3), 396–414.
32
33

34
35
36 Lambrecht, K. (1994). *Information structure and sentence form*. Cambridge: Cambridge
37 University Press.
38
39

40
41
42 Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for
43 categorical data. *Biometrics*, 33 (1), 159–174.
44
45

46
47
48 Massaro, D. & Cohen, M. (1983). Evaluation and integration of visual and auditory
49 information in speech perception. *Journal of Experimental Psychology Human*
50 *Perception & Performance*, 41 (5), 751–775.
51
52

53
54
55 McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746–
56 748. Retrieved from: <https://www.nature.com/articles/264746a0>.
57
58
59
60

1
2
3 Miranda, L. S. (2015). *Análise da entoação do português do Brasil segundo o modelo*
4 *IPO*. (Master thesis) – Faculdade de Letras da Universidade Federal do Rio de Janeiro,
5
6 Rio de Janeiro, 2015. Retrieved from:
7
8 <http://www.letras.ufrj.br/posverna/mestrado/MirandaLS.pdf>.
9
10
11

12
13 Miranda, L. S., Moraes, J. A. & Rilliard, A. (2019). Audiovisual perception of wh-
14 questions and wh-exclamations in Brazilian Portuguese. *Proceedings of Nineteenth*
15 *International Congress of Phonetic Sciences*, Melbourne, Australia, pp. 2941–2945.
16
17
18

19
20
21 Moraes, J. A. (1998). Intonation in Brazilian Portuguese. In: Hirst, D. & Di Cristo, A.
22 (Eds.) *Intonation Systems: A Survey of Twenty Languages* (pp. 179–194). Cambridge,
23 UK: Cambridge University Press.
24
25
26

27
28
29 Moraes, J. A. (2008). The pitch accents in Brazilian Portuguese: analysis by synthesis.
30 *Proceedings of Fourth Conference on Speech Prosody*, Campinas, Brazil, pp. 389–397.
31
32

33
34 Moraes, J. A., Miranda, L. S. & Rilliard, A. (2012). Facial gestures in the expression of
35 prosodic attitudes in Brazilian Portuguese. *Proceedings of Seventh GSCP International*
36 *Conference Speech and Corpora*, Belo Horizonte, Brazil, pp. 157–161.
37
38
39

40
41
42 Nicholson, K. G., Baum, S., Kilgour, A., Koh, C. K., Munhall, K. G. & Cuddy, L. L.
43 (2003). Impaired processing of prosodic and musical patterns after right hemisphere
44 damage. *Brain and Cognition*, 52, 382–389.
45
46
47

48
49
50 Ouni, S., Cohen, M., Ishak, H. & Massaro, D. (2006). Visual contribution to speech
51 perception: Measuring the intelligibility of animated talking heads. *EURASIP J. AUDIO*
52 *SPEECH MUSIC PROC.*, (2006) 2007: 047891. DOI:
53
54 <https://doi.org/10.1155/2007/47891>.
55
56
57
58
59
60

1
2
3 Peres, D. O., Raposo de Medeiros, B., Ferreira Netto, W. & Baia, M. F. A. (2011). The
4 role of the visual stimuli in the perception of prosody in Brazilian Portuguese.
5 *Proceedings of Fifth Conference on Laboratory Approaches to Romance Phonology*,
6 Somerville, Massachusetts, USA, pp. 136–141.
7

8
9
10
11
12
13 Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation*. Ph.D.
14 dissertation, MIT.
15

16
17
18
19 Qualtrics software, Version 2017 of Qualtrics. Copyright © [2019] *Qualtrics*. Online
20 platform available at: <https://www.qualtrics.com/>.
21

22
23
24 Rosenblum, L. D. (2005). Primacy of Multimodal Speech Perception. In: Pisoni, D. B. &
25 Remez, R. E. (Eds.). *Handbook of Speech Perception* (pp. 52–78). Oxford, UK:
26 Blackwell Publishing.
27

28
29
30
31
32 Vatikiotis-Bateson, E., Barbosa, A. V., Chow, C. Y., Oberg, M., Tan, J. and Yehia, H.
33 C. (2007). Audiovisual Lombard speech: reconciling production and perception.
34 *Proceedings of International Conference on Auditory-Visual Speech Processing*,
35 Hilvarenbeek, The Netherlands, pp. 45–50.
36

37
38
39
40
41
42 Vegas Pro software, Version 14 of Vegas Pro. Copyright © [2016] *MAGIX*. Software
43 available at: <https://www.vegascreativesoftware.com/>.
44

45
46
47
48 Srinivasan, R. & Massaro, D. (2003). Perceiving prosody from the face and voice:
49 Distinguishing statements from echoic questions in English. *Language and Speech*, 46
50 (1), 1–22.
51

52
53
54
55 Steeneken, H. & Geurtsen, F. W. M. (1988). *Description of the RSG-10 noise database*.
56 TNO Institute for Perception, report IZF, 1988–3.
57
58
59
60

1
2
3 Sumby, W. & Pollack, I. (1954). Visual contribution to speech intelligibility in noise.
4
5 *Journal of Acoustical Society of America*, 26, 212–215.
6

7
8 Swerts, M. & Krahmer, E. (2008). Facial expressions and prosodic prominence: effects
9
10 of modality and facial area. *Journal of Phonetics*, 36 (2), 219–238.
11
12

13
14 Torreira, F. & Valtersson, E. (2015). Phonetic and visual cues to questionhood in French.
15
16 *Phonetica*, 72, 20–42.
17

18
19 Tubau, S., González-Fuente, S., Prieto, P. and Espinal, M. A. (2015). Prosody and gesture
20
21 in the interpretation of yes-answers to negative yes/no-questions. *The Linguistic Review*,
22
23 32(1), 115–142.
24
25

26
27
28
29
30 **Figure 1:** Plot of the mean F_0 values (in ST) of normalized time over each modality: assertion (above) and
31 echo question (below).

32
33 **Figure 2:** Plot of the mean intensity (in dB, top panel) and duration (in z-score, bottom panel) over each
34 modality: assertion (left) and echo question (right).

35
36 **Figure 3:** Stills from a female speaker producing an assertion (above) and an echo question (below).

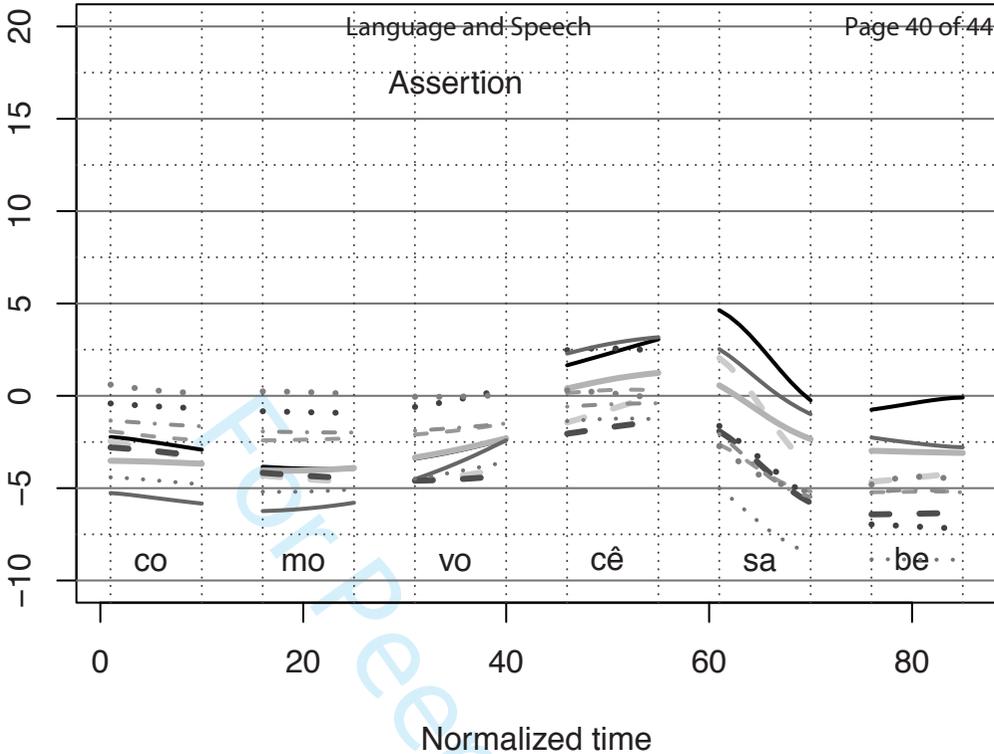
37
38 **Figure 4:** Stills from a male speaker producing an echo question.

39
40 **Figure 5:** Recognition rate of Experiments 1 and 2 with the following conditions: audio-only (AO), video-
41 only (VI) and audiovisual (AV) from Experiment 1; audiovisual in 0 dB SNR (AV-0), audiovisual in -6 dB
42 SNR (AV-6), audio-only in 0 dB SNR (AO-0) and audio-only in -6 dB SNR (AO-6) from Experiment 2.
43

44
45 **Figure 6:** Recognition rate of Experiments 1 and 3 with the following conditions: audio-only (AO), video-
46 only (VI) and audiovisual (AV) from Experiment 1; congruent audio and video (AVc) and incongruent
47 audio and video (AVi) from Experiment 3.
48
49
50
51
52
53
54
55
56
57
58
59
60

mean corrected frequency (ST)

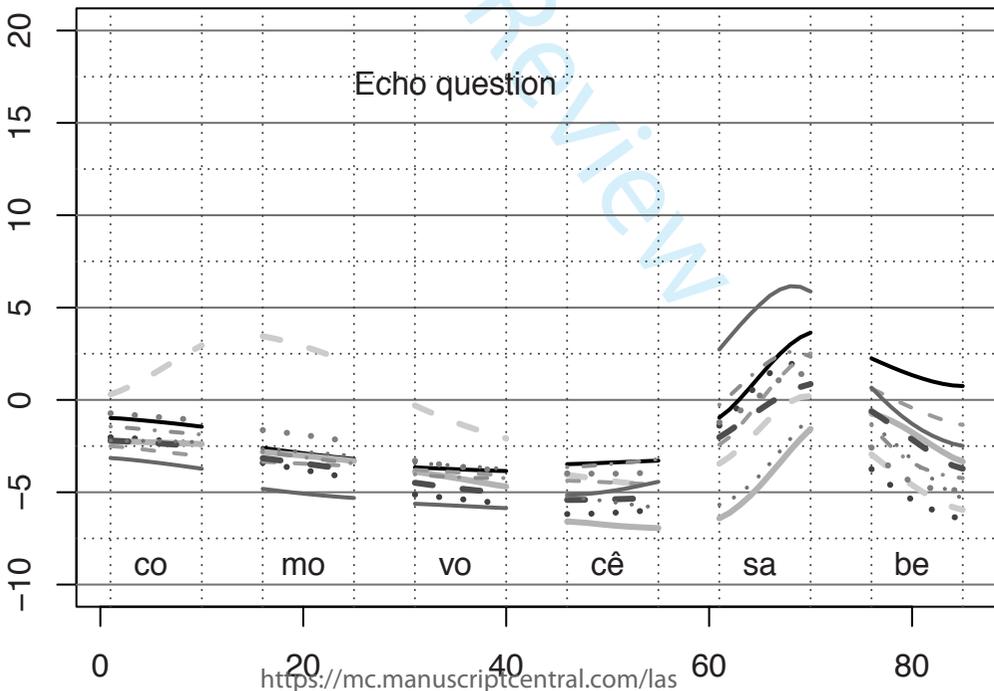
Assertion



Normalized time

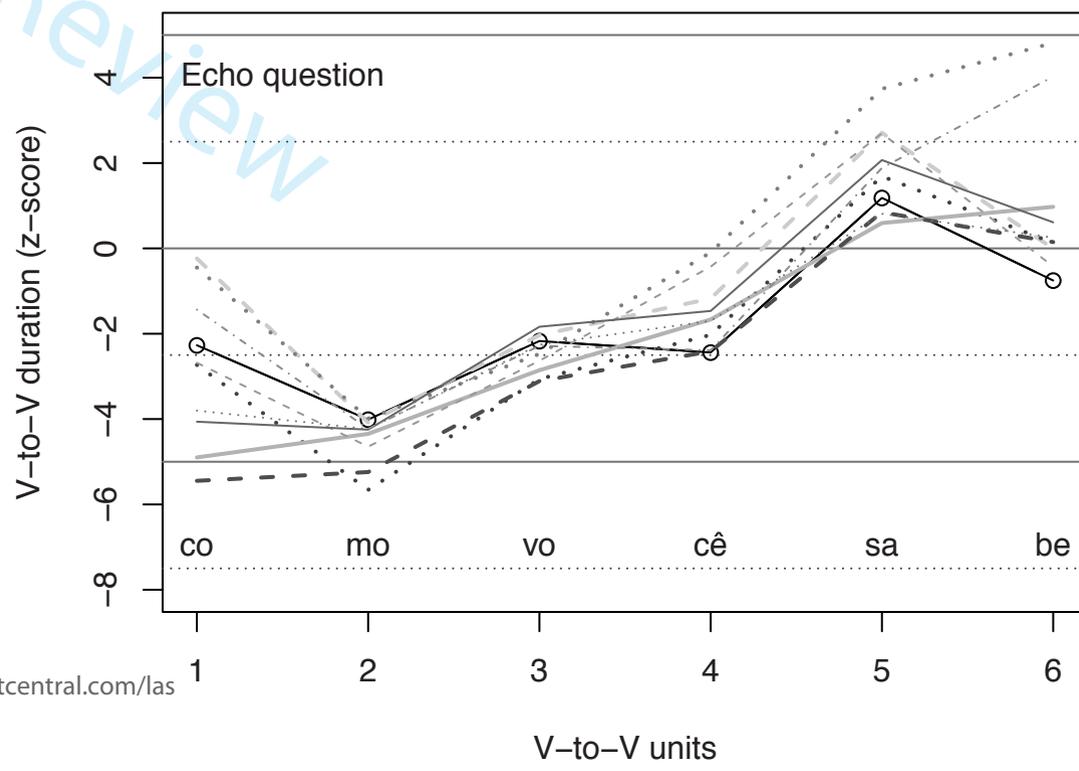
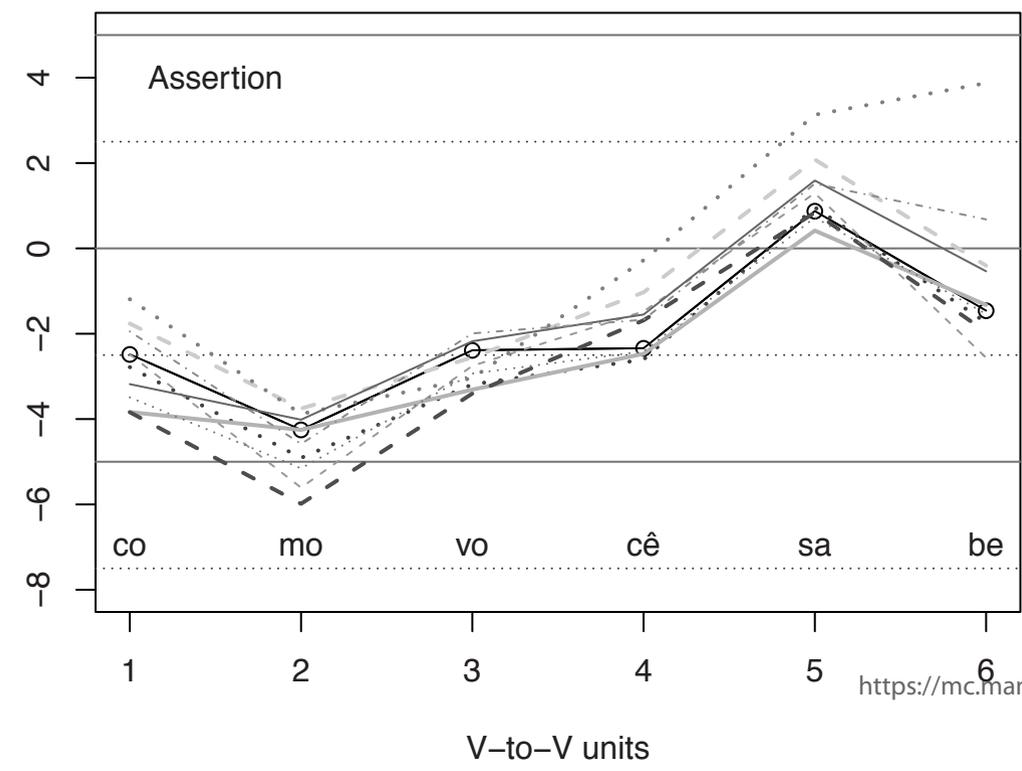
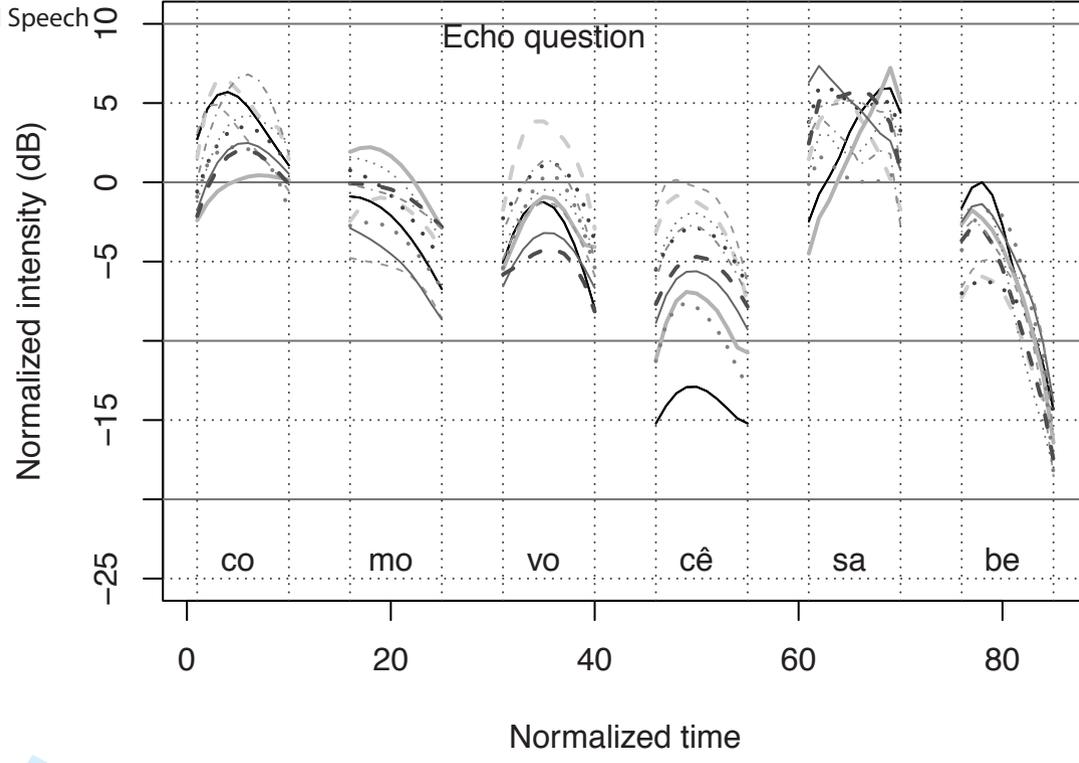
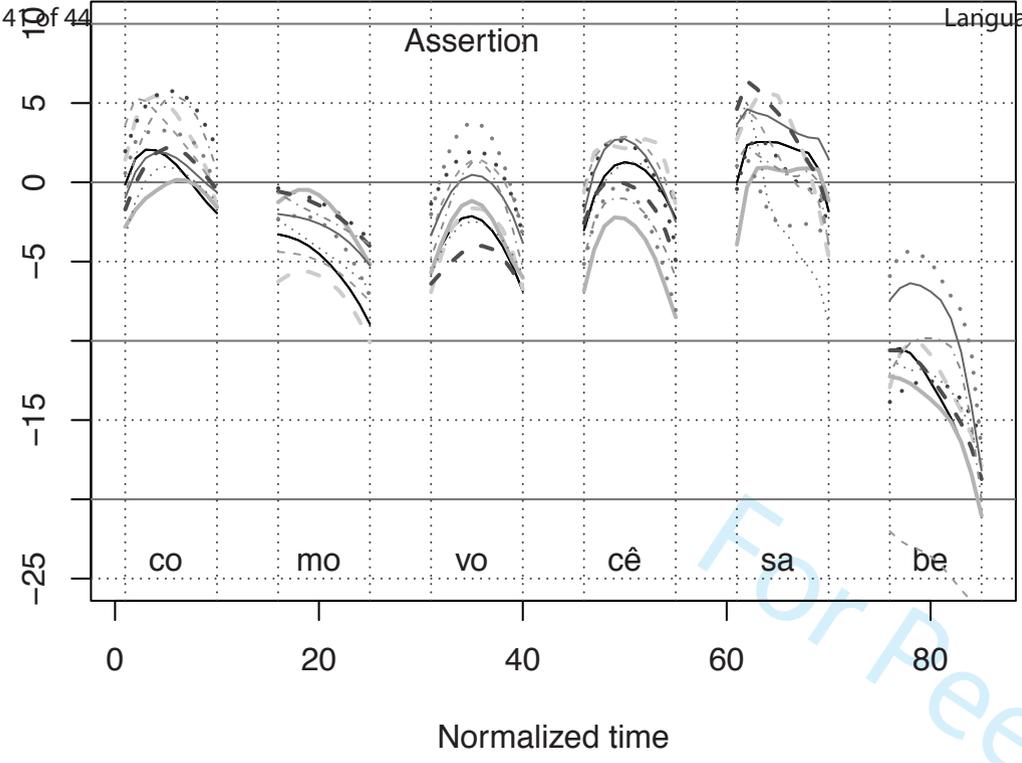
mean corrected frequency (ST)

Echo question

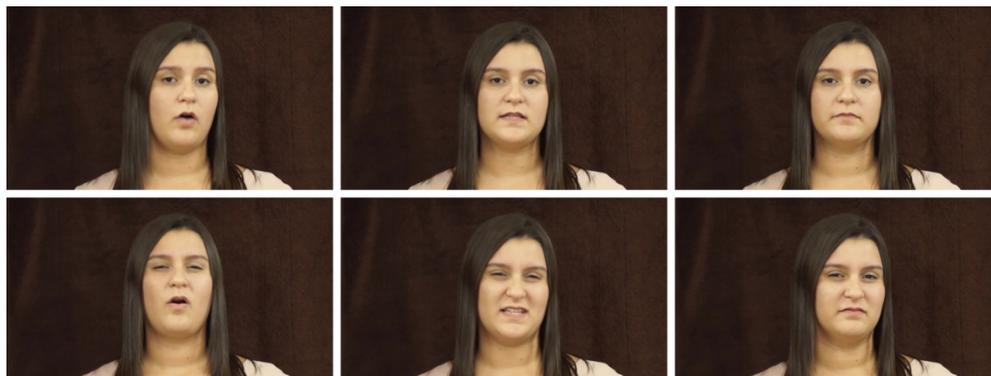


Normalized time

10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60



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

1375x518mm (72 x 72 DPI)

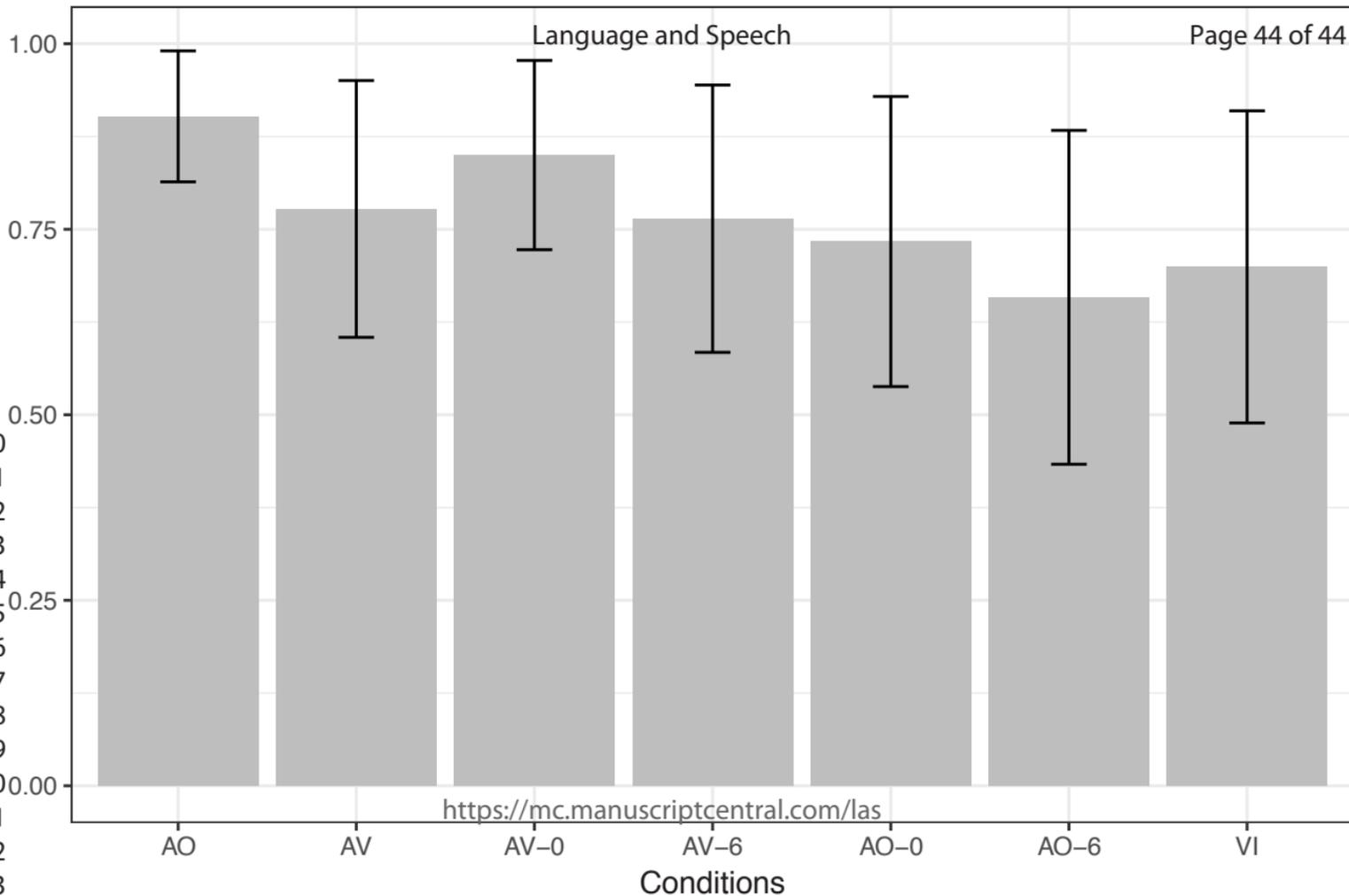
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60



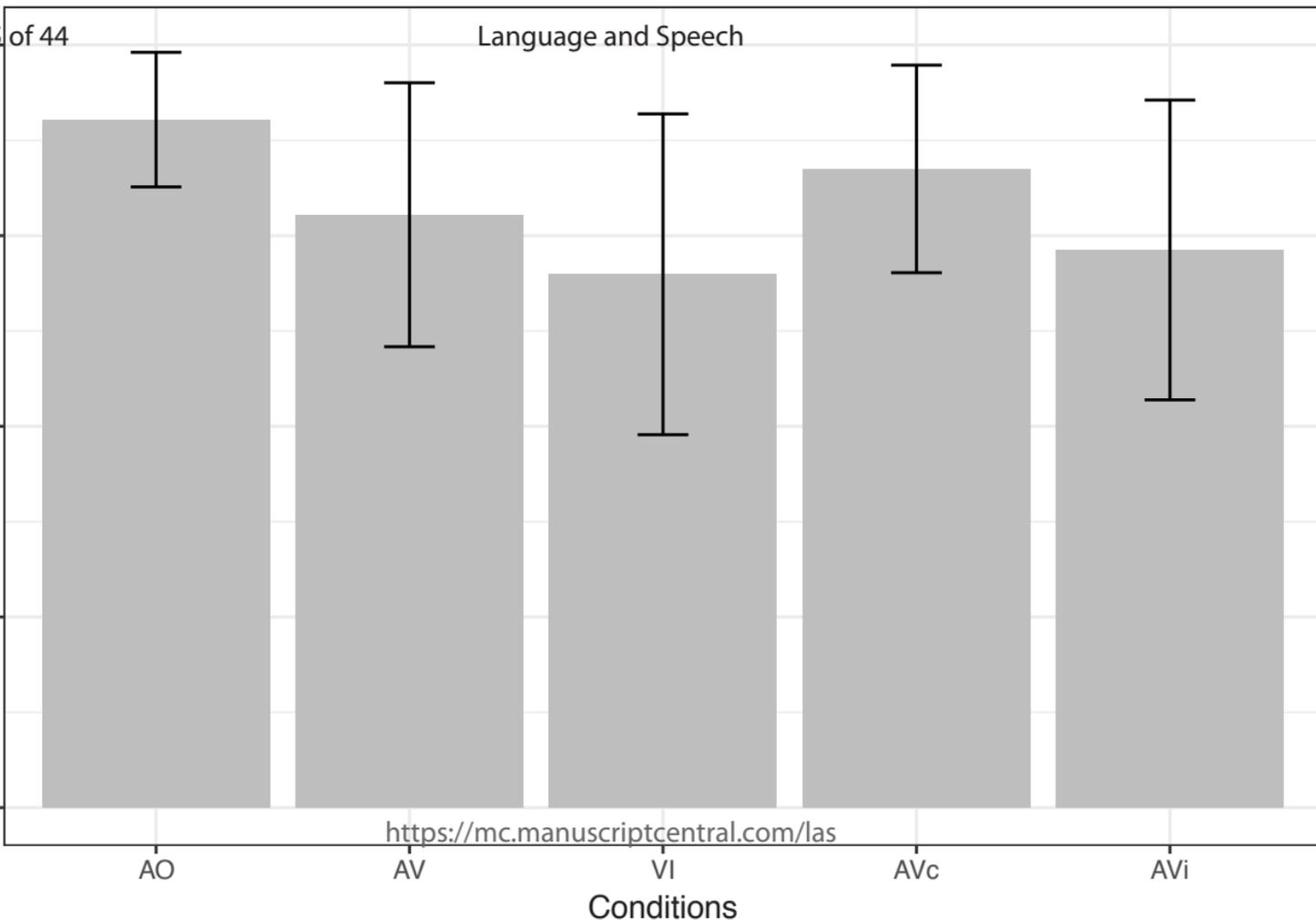
Stills from a male speaker producing an echo question.

1375x254mm (72 x 72 DPI)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23



<https://mc.manuscriptcentral.com/las>

Conditions