Laboratory Phonology
Journal of the Association for Laboratory Phonology

Cruz, M., et al 2017 The role of intonation and visual cues in the perception of sentence types: Evidence from European Portuguese varieties. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 8(1): 23, pp. 1–24, DOI: https://doi.org/10.5334/labphon.110

JOURNAL ARTICLE

The role of intonation and visual cues in the perception of sentence types: Evidence from European Portuguese varieties

Marisa Cruz[1], Marc Swerts[2] and Sónia Frota[1]

[1] Center of Linguistics, University of Lisbon, PT

[2] Tilburg University, NL

Corresponding author: Marisa Cruz (marisasousacruz@gmail.com)

In this paper, we explore the role of intonation and visual cues in the perception of statements and questions in two varieties of European Portuguese—the standard (SEP) and the insular variety of Azores, Ponta Delgada (PtD)—previously shown to convey sentence type contrasts by different uses of intonational means and/or facial gestures, namely eyebrow movements. Forty native speakers (20 from each variety) were exposed to SEP and PtD stimuli in a perception task with three conditions (audio only, video only, and audiovisual). The audiovisual condition includes congruent and incongruent (both original and manipulated) stimuli, where there is either a match or a mismatch between the auditory and visual features as potential cues for a specific sentence type. We concluded that both SEP and PtD participants rely more on intonation than on eyebrow movement to identify sentence types, even when exposed to incongruent audiovisual stimuli. In the absence of audio information, unexpectedly, participants do not interpret eyebrow raising as a question marker, not even when perceiving stimuli from their native variety. When exposed to non-native audiovisual stimuli, both SEP and PtD participants present longer reaction times (RTs), especially for incongruent stimuli. Finally, although we confirm the strength of intonation over visual cues, RTs in the audiovisual condition are significantly shorter than in the audio condition, thus pointing to the relevance of visual cues for structural/linguistic marking.

Keywords: intonation; facial gestures; multimodal perception; sentence types; varieties of European Portuguese

1. Introduction

The relation between gestures and speech has been modeled in a variety of ways. Some studies point to the hand-in-hand hypothesis (So et al., 2009; de Ruiter et al., 2012), assuming that gestures are redundant in the sense that they basically express information that can reliably be derived from the verbal content alone. Other studies propose an alternative hypothesis, based on the assumption of a trade-off relation between gestures and speech production, i.e., speech and gestures complement each other (Bangerter, 2004; Melinger & Levelt, 2004). Recently, these two hypotheses have been explored with different goals (ontogenetic, the role of gestures in social interaction, modeling, inter alia), providing an important contribution to the knowledge of different (non-)linguistic areas, prosody included. Generally speaking, the interaction between verbal prosody and visual prosody has been studied in order to understand whether these two modalities are parallel or complement each other. Within the most analyzed visual cues (eyebrows, head movements, and pointing gestures), and alongside pitch accents in the auditory

component, eyebrow movements were shown to play a strong role in the perception and distinction of specific sentence types and pragmatic meanings in several languages. Eyebrow movements may function as a question marker in French (Purson et al., 1999), even when they are not necessarily coordinated with fundamental frequency changes (Cavé et al., 1996). The same visual cue also has a significant effect on the perception of focus (Krahmer et al., 2002) and prominence (Swerts & Krahmer, 2004, 2006, 2008; Krahmer & Swerts, 2007) in Dutch, and may help to distinguish between specific sentence types and pragmatic meanings across languages, such as between Dutch and Catalan (Borràs-Comes & Prieto, 2011; Borràs-Comes et al., 2014; Crespo-Sendra et al., 2013). For instance, Crespo-Sendra et al. (2013) have shown that Catalan participants rely more on visual cues than Dutch participants in order to perceive the contrast between neutral and focused yes–no questions, and that this is related to the kind/richness of the respective intonational cues (Catalan speakers use the same intonational contour for both pragmatic meanings, with a different pitch range, whereas Dutch speakers use different intonational contours). These studies thus show that visual cues have added value compared to a speech-only condition: (i) Accuracy in the responses given by perceivers based on auditory cues is enhanced when visual information is added, and (ii) the role of visual cues is more relevant when auditory cues are ambiguous or weaker. Although complementary, visual features have been shown to have a weaker cue strength than auditory ones for the distinction between statements and questions. This was observed for instance for Swedish (House, 2002; Granström & House, 2004) or English (Srinivasan & Massaro, 2003).

In this paper, we explore the role of intonation and facial gestures in the perception of sentence types (statements and questions) across two European Portuguese (EP) varieties. In EP, the relation between intonation and visual cues has only been recently explored (Cruz & Frota, 2014; Cruz et al., 2015). It was observed that visual cues, similarly to intonational cues, may vary across varieties of the same language, and across sentence types and pragmatic meanings. Although a relation between pitch accent types and gesture types was found, this relation was mediated by sentence type and/or pragmatic meaning. As it can be observed in **Table 1**, neutral statements and neutral yes–no questions show distinct visual cues time-aligned with the nuclear pitch accent (NPA) (H +)L*, in Standard European Portuguese (SEP). However, the reverse was also found in this variety, i.e., the same facial gesture (e.g., head up-down and eyebrow raising) associated with different NPAs, conveying different sentence types or pragmatic meanings (e.g., narrow focused statements—H* + L—versus broad focused yes–no questions—H + L*). Also important, and differently from the insular variety of Portuguese spoken in Azores, Ponta Delgada (PtD), in which intonation does not seem to distinguish neutral statements and neutral yes–no questions, in SEP these two sentence types differ on the boundary tones used (L%

Table 1: Visual cues aligned with pitch accent/boundary tone types in neutral statements and neutral yes–no questions in Standard European Portuguese (SEP) and in the insular variety of Azores, Ponta Delgada (PtD). Dominant tone and gesture patterns across and within speakers are represented. Adapted from Cruz et al. (2015), Tables 1 and 2. For each variety, in the column 'Tonal,' the first row refers to the nuclear pitch accent and the second row refers to the boundary tone.

	Neutral statements		Neutral yes–no questions	
Variety	**Tonal**	**Visual**	**Tonal**	**Visual**
SEP	H+L*	head up-down	H+L*	head up-down + eyebrow raising
	L%	neutral position	LH%	neutral position
PtD	(H+)L*	head up-down	(H+)L*	head up-down + eyebrow raising
	L%	neutral position	L%	neutral position

for neutral statements; LH% for neutral yes–no questions), although time-aligned with the same visual cue (the return to the neutral position). Importantly, previous studies revealed that the interactions between (i) sentence type and pragmatic meaning and between (ii) sentence type and language variety were good predictors of visual cues time-aligned with NPA and boundary tone (BT) types. However, language variety alone was shown not to be a good predictor, unlike sentence type and pragmatic meaning (Cruz et al., 2015).

Since in production facial gestures are affected by sentence type, pragmatic meaning, and language variety, we generally hypothesize that speakers across varieties are sensitive to visual information, especially for cases where the auditory/tonal features are not very informative. The two EP varieties of SEP and PtD constitute a good test case. The nuclear contour H + L* L% is the most frequently used in SEP for neutral statements, and in PtD for both neutral statements and neutral yes–no questions (**Table 1**). Neutral statements are thus tonally and visually identical in the two varieties. However, neutral questions are tonally different but visually identical across varieties. Importantly, the contrast between neutral statements in SEP and neutral questions in PtD relies only on the visual cues: In SEP, neutral statements are predominantly produced with a head up-down movement; in PtD, neutral yes–no questions tend to be produced with an eyebrow raising movement additionally to the head up-down. Since the eyebrow raising movement is also characteristic of yes–no questions in SEP, we hypothesize that SEP participants will use this visual cue in order to identify neutral yes–no questions produced in PtD when the visual cues are available (Hypothesis 1). The same utterances, however, will be identified as statements when the visual cues are not available (audio stimuli only) (Hypothesis 2).

These predictions are addressed in the present paper, by testing the role of facial gestures in the absence of a tonal contrast across language varieties. Furthermore, this research will contribute to our understanding of the (relative) sensitivity of both SEP and PtD participants to different cues in the speech signal (intonation only, gestures only, or both) for the distinction between sentence types across varieties of the same language.

2. Methodology

2.1. Participants

Twenty-six native speakers from the standard variety of European Portuguese (SEP) and 29 native speakers from the insular variety of Azores, Ponta Delgada (PtD), participated in a perception study. The data of 20 native participants per variety were considered for analysis, both groups with a mean of 28 years of age (14 females for SEP; 9 females for PtD). None of them reported any hearing or vision problems. Fifteen subjects were excluded according to the rejection criteria explained in Section 2.3.

2.2. Materials

The materials used had been collected earlier for the construction of an audiovisual database of prosodic variation in Portuguese—the *Interactive Atlas of the Prosody of Portuguese Webplatform* (Frota & Cruz, coord., 2012–2015). We selected 2 neutral statements and 2 neutral yes–no questions produced by 3 speakers per variety (SEP and PtD) in a semi-spontaneous task elicited by means of a Discourse Completion Task (DCT) (Kasper & Dahl, 1991; Félix-Brasdefer, 2010). These utterances were previously analyzed in order to identify the most frequent patterns with respect (i) to intonational contours per sentence type/pragmatic meaning (Cruz, 2013; Crespo-Sendra et al., 2014; Frota et al., 2015a), using the P-ToBI annotation system (Frota, 2014; Frota et al., 2015b), and (ii) to facial gestures time-aligned with the nuclear pitch accents and boundary tones (Cruz & Frota, 2014; Cruz et al., 2015), using a labeling system based on FACS—Facial Action Coding System (Ekman et al., 2002).

In accordance with the most typical patterns found in those previous studies, all yes–no questions selected for this perception task present the intonational contour (i) H + L* LH% for SEP and (ii) L* L% for PtD, and (iii) the eyebrow raising movement time-aligned with the nuclear contour. As for declaratives,[1] only used in the training phase, we also selected those presenting, in both varieties, the nuclear contour H + L* L% (SEP) and L* L% (PtD), and the head up-down movement.

2.3. Procedure

To investigate the role of intonation and eyebrow raising movement for the distinction between statements and yes–no questions across varieties, we ran the same perception task for SEP and PtD participants. The task was an overt identification experiment with three conditions: Audio only (AO), video only (VO), and audiovisual (AV) (**Figure 1**). The AV condition included original and manipulated stimuli. In all conditions, participants, sitting in front of a laptop and wearing headphones, were asked to use the keyboard in order to classify each stimulus as being more declarative-like or more interrogative-like, by using a Likert scale from 1 (declarative) to 5 (interrogative). This method also allowed us to examine the degree of certainty of a given response, with extreme values of the scale meaning a high degree of certainty on the sentence type involved, the mid value meaning strong doubt, and the other two values (2 and 4) meaning a tendency to classify the stimulus as being more declarative-like (2) or as being more interrogative-like (4).

For the AO and VO conditions, the stimuli were extracted from the original audiovisual recordings using VideoPad Video Editor Professional 3.58, which allows saving the audio and video tracks separately.

In each unimodal condition (AO, VO), a total of 24 trials (2 neutral yes–no questions × 2 speakers × 2 varieties × 3 repetitions) were included; the bimodal condition (AV) included a total of 36 trials (24 trials with original stimuli—2 neutral yes–no questions × 2 speakers × 2 varieties × 3 repetitions—and 12 trials with manipulated stimuli from SEP—2 productions × 2 speakers × 3 repetitions). The manipulated stimuli in the AV condition combine the audio track of neutral statements and the video track of neutral yes–no questions, all produced in SEP. These materials were submitted to a short pretest and three evaluators agreed that it could not be seen that these AV materials had been manipulated. Note that the interpretation of AV stimuli varied as a function of

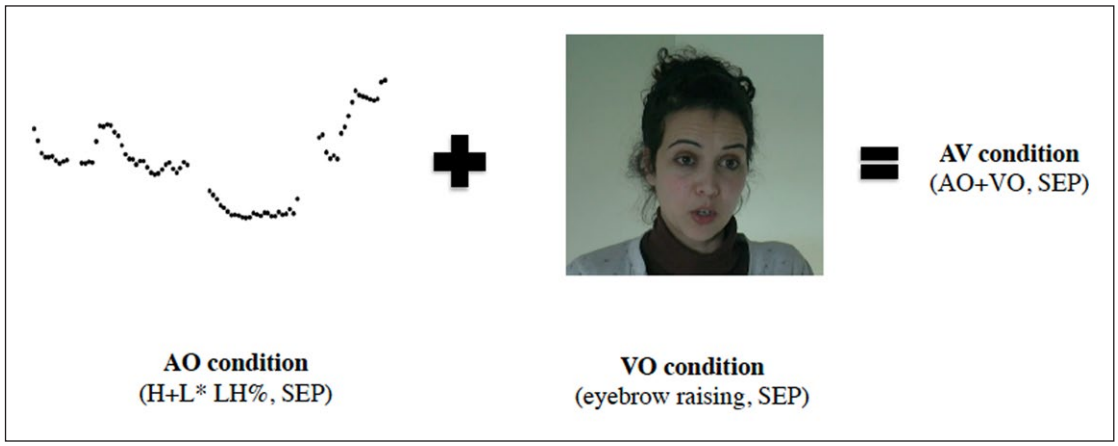

Figure 1: Example of the same stimulus – the yes–no question *Tem compota?* ('Do you have jam?') in each condition: AO, VO, and AV.

[1] The terms 'declarative' and 'statement' are interchangeably used throughout the text as synonymous.

the language background of the participants in our comparison study. Thus, AV stimuli are possibly interpreted by participants as being congruent[2] or incongruent stimuli, where there is either a match or a mismatch between the auditory and visual features as potential cues for a specific sentence type (**Table 2**). Importantly, our hypothesis is that manipulated AV stimuli from SEP represent a natural congruency for PtD participants (Hypothesis 3), as their neutral yes–no questions are produced with an all-falling intonational contour (as statements in SEP), but with eyebrow raising (as neutral yes–no questions in SEP). Inversely, we hypothesize that the original AV stimuli from PtD constitute a natural incongruency from the point of view of SEP participants (Hypothesis 4), because PtD neutral yes–no questions are visually identical to the same sentence type in SEP, with the eyebrow raising movement, but intonationally identical to SEP neutral statements, with an all-falling nuclear configuration. Overall, participants of each variety had to decide about the sentence type of a total of 84 trials (24 AO, 24 VO, 12 original AV from SEP, 12 original AV from PtD, 12 manipulated AV from SEP) and a total of 1680 responses were obtained (84 trials × 20 participants) for each variety. Reaction times (in milliseconds) were also measured in order to observe whether participants need more time to react when being exposed to stimuli that are more difficult to categorize (Chen, 2003; Schneider et al., 2011). Thus, we expect longer reaction times in the VO condition than in the AO and AV conditions (Hypothesis 5). Considering the AV stimuli and their expected interpretation depending on the participants' native variety (**Table 2**), for SEP participants, we expect longer reaction times in the manipulated AV condition from SEP, like in the original AV condition from PtD (which is naturally incongruent for SEP participants), than in the original AV condition from SEP (Hypothesis 6). For PtD participants, since we consider that the manipulated AV condition from SEP is a natural congruency for them, we expect similar reaction times between original AV stimuli from PtD and manipulated AV stimuli from SEP (Hypothesis 7).

In order to make participants acquainted with the kind of stimuli and experimental procedures used, the test phase of each condition, which included native and non-native stimuli and was the same for all participants (SEP and PtD), was preceded by a training phase, containing native sentences only. Thus, in the perception task run in SEP, the training phase included SEP sentences only, produced by a speaker not included in the test phase; the training phase of the perception task run in PtD included sentences produced by a PtD speaker (also not included in the test phase). In sum, the speaker and the materials used in the training phase were not considered in the test phase. Actually, if the same speaker appeared in both phases—training and test—we would not know whether participants' responses are influenced or not by the fact that they already heard/viewed this specific speaker before. Contrary to the test phase, which included 84 trials with neutral yes–no questions only,

Table 2: AV stimuli: Their characteristics (dark grey) and expected interpretation depending on the participants' native variety (light grey).

Stimuli	Manipulated AV from SEP	Original AV from PtD	Original AV from SEP
	intonation of statements + gesture of questions	intonation of questions + gesture of questions	
	H+L* L% + eyebrow raising		H+L* LH% + eyebrow raising
SEP	Incongruent	Naturally incongruent	Congruent
PtD	Naturally congruent	Congruent	Congruent (?)

[2] Congruent stimuli, in the present research, always refer to interrogative stimuli.

the training phase also included neutral statements, besides questions, thus summarizing 36 trials (4 productions [2 neutral statements + 2 neutral yes–no questions] × 1 speaker × 3 repetitions × 3 conditions). The test phase thus only includes two yes–no questions produced by the other two speakers, one being exactly the same sentence for both speakers (*Querem caramelos?*, 'Do you want caramels?') and the other being similar in terms of syllabic length (3 and 4 syllables: *Casaram?*, 'Did they marry?', and *Tem compota?*, 'Do you have jam?'). Participants did not receive any feedback for their responses (not even in the training phase) and they were not allowed to replay stimuli. Importantly, in the AV condition participants were not trained for manipulated combinations (see **Figure 2** for a detailed scheme of the experimental design).

All participants performed the three conditions, separated in three blocks. The duration of the break between blocks was not fixed, but it lasted the time participants needed to read the instructions preceding the subsequent training phase. The unimodal conditions (AO, VO) were presented in different orders, and the bimodal condition (AV) always followed the two unimodal ones. Thus, in each variety, for 10 participants the order of presentation was AO-VO-AV; for the other 10 the order was VO-AO-AV. The order of stimuli presentation was randomized within each condition, being different across participants. Before the AV condition, all participants were exposed to a short comic video

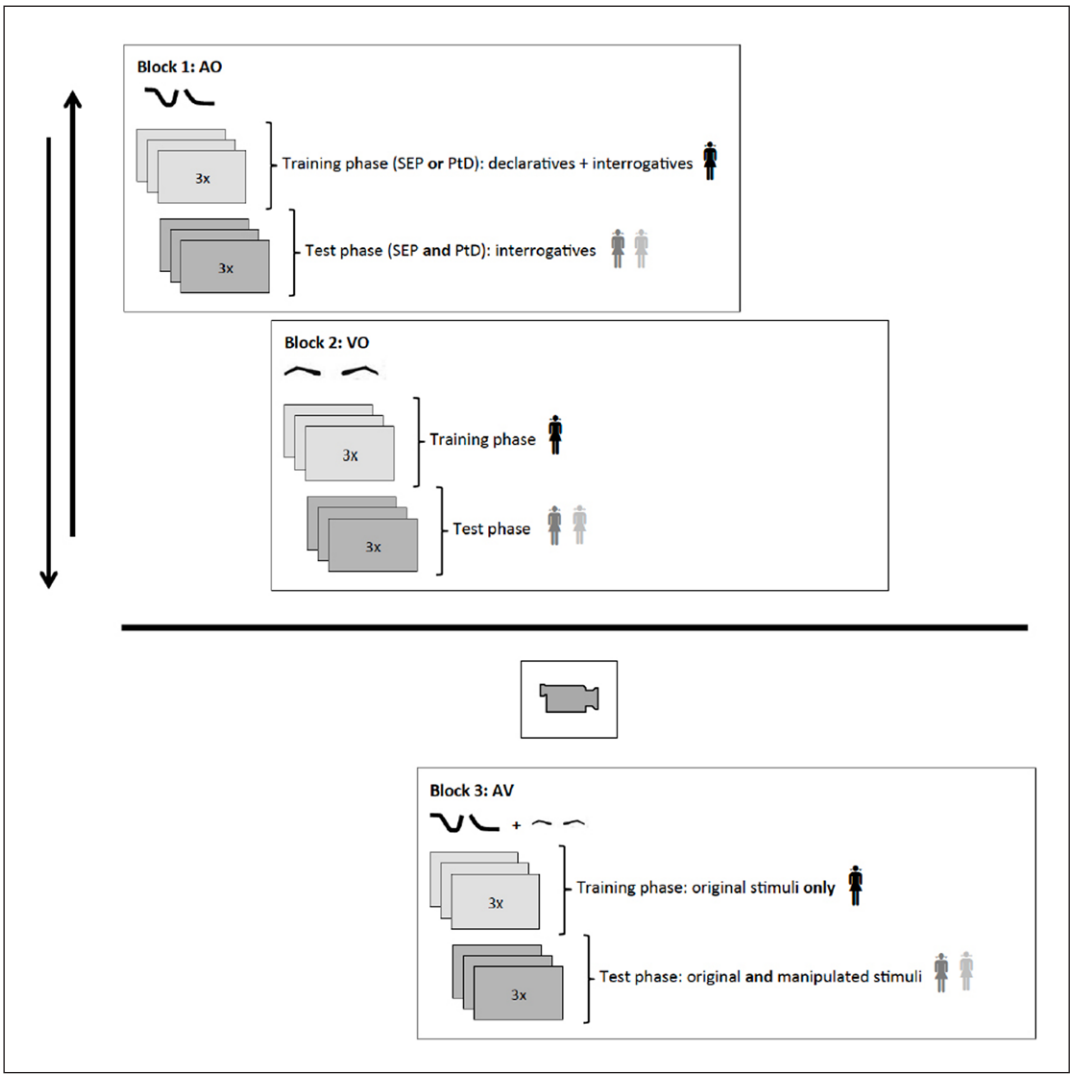

Figure 2: Detailed scheme of the experimental design.

(about 5 minutes long) to minimize potential differences in learning effects because of variable preceding order sequences (see **Figure 2**). The experiment was run in SuperLab 4.5 and lasted around 15 minutes. Participants were informed that they had to respond as fast as possible immediately after each stimulus and that the following stimulus would be played right after their response (no interstimulus interval was defined). They were also instructed to keep both hands closer to the relevant keys during the task. Each condition, independently of the order of presentation, was preceded by a display with short instructions, which also reminded participants about the Likert scale in use.

3. Results and discussion

Before the analysis of participants' responses, and in order to ensure that each participant was focused on the task, a rejection criterion was applied, following other perception studies (e.g., Ramus & Mehler, 1999). We excluded 4 participants who, in the training phase of the AO condition, were not able to distinguish between declaratives and yes–no questions of their native variety above chance level. Additionally, and in order to have a balanced data set, we also excluded 11 participants whose response files contained missing data (because they had answered before the end of a given stimulus presentation or because they had pressed the key too much, when giving their answer on the keyboard). Finally, 1 of the participants who turned out not to be native regarding the varieties under study (and only provided this information at the end of the task, while filling in the profile form) was also excluded. In total, we discarded the data of 6 participants from SEP and 9 participants from PtD, and early responses were the exclusion criterion most observed (for 10 out of 15 excluded participants). After this exclusion, the data of 20 participants per variety were considered for analysis.

In order to have a more general picture of the sentence type most frequently identified, we calculated the mean response given in each condition (AO, VO, AV). In this sense, mean responses near the extreme values of the scale (1 and 5) reveal that participants identify declaratives and interrogatives, respectively, with a high degree of certainty, whereas mean responses around values 2 and 4 just reflect a mere tendency to identify the stimuli as being more declarative-like (2) or more interrogative-like (4). Mean responses closer to value 3 indicate doubt.[3] As for the automatically recorded reaction times (RTs), we calculated mean RTs per condition and, where relevant, a deeper analysis was made, considering the mean reaction times for each level of the Likert scale.

The results for the conditions AO, VO, and AV (original stimuli) are presented in Section 3.1, where we aim at observing if the all-falling yes–no questions produced by PtD speakers are recognized as such by SEP participants, as well as identifying the kind of cues (audio only, visual only, or audiovisual) that contributes the most for this recognition (Section 3.1.1). Additionally, by analyzing the results of PtD participants perceiving SEP utterances, we aim at estimating the relative weight of each kind of cues for each variety (Section 3.1.2). In Section 3.2, we describe in detail the results for the audiovisual condition (Section 3.2.1 for SEP participants; Section 3.2.2 for PtD participants), including a comparison between original and manipulated stimuli and a discussion regarding the kind of cues (intonational or visual) that influences participants' responses. Reaction times are also analyzed in Section 3.3, in order to explore whether participants need more time to react to stimuli that are more difficult to categorize. This includes a comparison across conditions (Section 3.3.1 for SEP participants; Section 3.3.2 for PtD participants), and

[3] We converted our Likert scale into an interval scale in order to run the ANOVAs. Although we recognize the potential caveat involved in this conversion, we assume that the fact we gave clear instructions to all participants on the rates they should choose depending on the degree of certainty they had in their response ensures homogeneous intervals on the scale.

also a zoom-in analysis in the AV condition, considering the exposure of SEP participants (Section 3.3.3) and PtD participants (Section 3.3.4) to both original and manipulated AV stimuli.

3.1. Intonational cues (AO), visual cues (VO), and/or audiovisual cues (AV): Do they have the same relevance across varieties?

As described in Section 1, neutral yes–no questions in SEP and in PtD are tonally different (falling-rising in SEP, and all-falling in PtD), but visually identical (eyebrow raising movement). Thus, our hypothesis is that SEP participants will use this visual cue in order to identify yes–no questions produced in PtD when the visual cues are available (Hypothesis 1). When unavailable (audio stimuli only), and since the contrast between statements in SEP and yes–no questions in PtD relies only on the visual cues, we expect that SEP participants will identify the same utterances as statements (Hypothesis 2). Additionally, we also inspected whether the order of stimuli presentation (audio only first or video only first) influenced participants' responses or not. If an effect of task order is observed, data analysis will have to reflect a comparison between two separate groups (each one including 10 participants) within SEP or PtD, instead of considering all participants (20) per variety as a group. For this reason, we firstly observed whether there was a task order effect.

3.1.1. SEP participants

For SEP participants, a mixed factorial ANOVA with one between subject factor (task order: AO-VO-AV, VO-AO-AV), and two within subjects factors (variety heard: SEP, PtD; and condition: AO, VO, AV) yield a non-significant effect of task order ($F(1, 18) = .821$, $p > .1$), which means that SEP participants first exposed to AO stimuli did not perform differently from SEP participants first exposed to VO stimuli, thus the order of stimuli presentation does not affect participants' responses.

Additionally, a significant main effect was found for variety ($F(1, 18) = 205.737$, $p < .001$) and condition ($F(2, 36) = 18.048$, $p < .001$), and a significant interaction was observed between the two factors ($F(2, 36) = 212.862$, $p < .001$), which means that the effect of variety on responses depends on the condition (AO, VO, AV) SEP participants are exposed to. Mean responses per condition and variety perceived are shown in **Figure 3**. These findings point to two main observations. The VO condition clearly presents a different pattern of responses given by SEP participants perceiving their native variety. Differently from the AO ($M = 4.79$, $SD = .32$) and AV ($M = 4.79$, $SD = .36$) conditions, where SEP stimuli are dominantly perceived as interrogatives (clearly above 4.5), in the VO condition SEP stimuli are mostly perceived as declaratives ($M = 2.65$, $SD = .59$) and categorization seems to be more difficult, given the spread in the responses, also visible in the standard deviation for the VO condition (versus the standard deviation in the AO and AV conditions). As for the perception of PtD stimuli, the VO condition presents a similar pattern as for perception of SEP stimuli ($M = 3.06$, $SD = .68$), but in this case there is markedly more doubt (3). Actually, a paired samples t-test shows that responses given by SEP participants in the VO condition are significantly different depending on the variety they are exposed to (SEP: $M = 2.65$, $SD = .59$; PtD: $M = 3.06$, $SD = .68$; $t(19) = –2.485$, $p < .05$). We thus conclude that the visual cue presented in isolation is not enough for SEP participants to decide about the sentence type of a given production, contrary to our Hypothesis 1. What is striking about PtD stimuli is that, unlike SEP stimuli, they are perceived as more declarative-like than interrogative-like (which confirms our Hypothesis 2), even in the presence of audiovisual information.

In the AO ($M = 2.12$, $SD = .60$) and AV ($M = 1.80$, $SD = .51$) conditions, as also shown in **Figure 3**, our findings lead to the conclusion that SEP participants rely more on

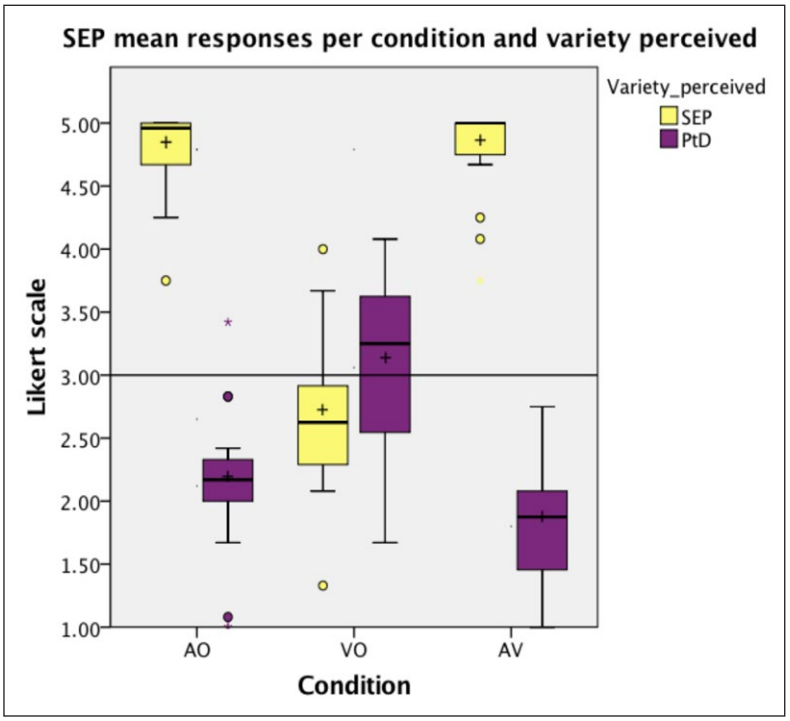

Figure 3: SEP participants: Mean responses (1 = more declarative-like; 5 = more interrogative-like) given per condition (AO, VO, AV) and variety perceived (SEP, PtD). Yellow identifies the native variety and the non-native variety is represented in purple. The '+' marks in the boxplots represent mean values.

intonation than on visual cues because they tend to classify yes–no questions produced by PtD speakers as declaratives in both conditions. Thus, although the eyebrow raising movement is present in the AV condition, the all-falling nuclear configuration (H + L* L%) that characterizes yes–no questions produced by PtD speakers seems to be the crucial cue used by SEP participants to classify the stimuli. Given this finding, we would expect that, in the absence of audio information, SEP participants would rely instead on the visual cue (the eyebrow raising movement), and thus classify the stimuli as interrogatives in the VO condition, independently of the variety perceived. However, the results do not confirm this expectation, as already mentioned above. This could be explained by the fact that the eyebrow raising movement is not exclusively a question marker in EP. Indeed, this visual cue is also used to convey narrow focused declaratives not only in SEP, but in all the varieties analyzed so far, PtD included (Cruz & Frota, 2014; Cruz et al., 2015). Another possible explanation is that the eyebrow raising movement might not be a strong cue, which thus leads participants to opt for a default response when in doubt, i.e., declarative. Thus, in the absence of audio information, SEP listeners tend to behave at chance level, or to identify the eyebrow raising movement mainly with narrow focused statements, or to opt for a default response. The latter possibility may underlie the fact that the degree of certainty of declarative answers in the AV condition is higher (mean value below 2, and lower standard deviation) than in the AO condition, for PtD stimuli.

3.1.2. PtD participants

Let us now consider the results for PtD participants. A mixed factorial ANOVA with one between subject factor (task order: AO-VO-AV, VO-AO-AV), and two within subjects factors (variety: PtD, SEP; and condition: AO, VO, AV) yield a non-significant effect of task order ($F(1, 18) = 1.213$, $p > .1$), meaning that PtD participants first exposed to AO stimuli did not perform differently from PtD participants first exposed to VO stimuli, thus

the order of stimuli presentation does not affect participants' responses. Differently from SEP participants, for PtD participants a non-significant main effect was found for variety ($F(1, 18) = 3.035, p > .05$), which means that PtD responses do not depend on the variety they are exposed to. However, as for SEP participants, a significant main effect was found for condition ($F(2, 36) = 43.516, p < .001$), with PtD responses in VO condition being significantly different from responses given in AO and AV conditions ($p < .001$), whereas a non-significant difference was found between the AO and AV conditions ($p > .1$). Mean responses per condition and variety perceived are shown in **Figure 4**. In the VO condition SEP (non-native) stimuli are perceived as more declarative-like ($M = 2.35, SD = .69$) and PtD (native) stimuli trigger a strong doubt ($M = 2.98, SD = .63$). By contrast, in the AO and AV conditions both native (AO: $M = 3.83, SD = .48$; AV: $M = 3.94, SD = .72$) and non-native stimuli (AO: $M = 4.50, SD = .55$; AV: $M = 4.36, SD = .91$) are dominantly perceived as interrogatives.

Additionally a significant interaction was observed between variety and condition ($F(2, 36) = 26.139, p < .001$), which means that the effect of variety depends on the condition (AO, VO, AV) PtD participants are exposed to, as also depicted in **Figure 4** (since no significant differences were found between task orders, responses given by the two subgroups of PtD participants were merged).

By comparing **Figures 3** and **4**, we can further observe that PtD participants differ from SEP participants when exposed to non-native stimuli: SEP participants mainly perceived PtD stimuli as declaratives (as neutral yes–no questions in PtD also exhibit a falling intonation, similarly to neutral statements in SEP), whereas PtD participants identify SEP stimuli as interrogatives. Interestingly, PtD participants classify PtD stimuli as dominantly interrogative-like, which could lead us to conclude that visual cues are important for PtD participants, helping to identify a neutral yes–no question (through the presence of the eyebrow

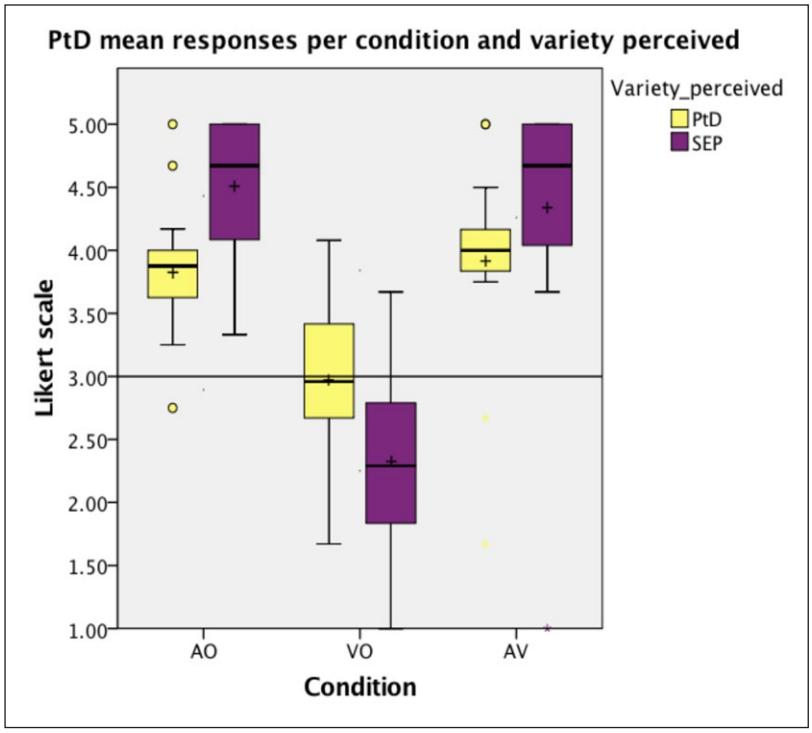

Figure 4: PtD participants: Mean responses (1 = more declarative-like; 5 = more interrogative-like) given per condition (AO, VO, AV) and variety perceived (PtD, SEP). Yellow identifies the native variety and the non-native variety is represented in purple. The '+' marks in the boxplots represent mean values.

raising accompanying the falling intonation). However, as clearly shown in **Figure 4**, in the AO condition, PtD participants also identify their native stimuli as interrogatives despite the absence of visual cues, and, furthermore, in the VO condition visual cues are not enough to identify the sentence type of the stimuli presented as interrogative (in fact, PtD participants seem to behave at chance level). These findings show that, contrary to our expectation, PtD participants, just like SEP participants, rely more on intonation than on facial gestures; otherwise, they would have identified their native VO stimuli as more interrogative-like. Actually, an independent samples t-test was run for the VO condition only with mean responses given to PtD stimuli and to SEP stimuli as dependent variables, and the group of participants as grouping variable. We observed that the performance of PtD participants in the VO condition does not significantly differ from the performance of SEP participants when faced with PtD stimuli (PtD: $M = 2.98$, $SD = .63$; SEP: $M = 3.06$, $SD = .68$; $t(38) = –.405$, $p > .1$) or with SEP stimuli (PtD: $M = 2.35$, $SD = .69$; SEP: $M = 2.65$, $SD = .59$; $t(38) = –1.496$, $p > .1$). Another argument in support of the overriding role of intonation for PtD participants is the obtained result in the training phase, in which they were exposed to both declaratives and interrogatives. Interestingly, in this phase, and based on audio cues only, they clearly distinguished between the falling intonation of declaratives and the falling intonation of interrogatives, presenting an overall correctness rate of 83%.

As with SEP participants, PtD participants in the absence of audio information tend to behave at chance level or to identify the eyebrow raising movement with narrow focused statements (used in SEP and PtD—Cruz & Frota, 2014; Cruz et al., 2015). An alternative explanation would be that when only visual cues are available, information is not enough for sentence type identification and thus participants either respond at chance level or use a default response strategy, which is declarative.

3.2. What counts most in (in)congruency? Intonational or visual cues?

As mentioned in Section 2.3, we manipulated stimuli from SEP (falling intonation from statements together with the eyebrow raising movement from yes–no questions). We hypothesize that these stimuli represent a natural congruency for PtD participants (Hypothesis 3), as their neutral yes–no questions are produced with an all-falling intonational contour (as statements in SEP), but with eyebrow raising (as neutral yes–no questions in SEP). Inversely, we hypothesize that the original AV stimuli from PtD constitute a natural incongruency from the point of view of SEP participants (Hypothesis 4), because PtD neutral yes–no questions are visually identical to the same sentence type in SEP, with the eyebrow raising movement, but intonationally identical to SEP neutral statements, with an all-falling nuclear configuration (see **Table 2**).

In order to test these hypotheses, responses given by SEP participants in the AV condition were examined by comparing: (i) Original versus manipulated AV stimuli produced by SEP speakers, and (ii) manipulated AV stimuli produced by SEP speakers versus original AV stimuli produced by PtD speakers, which form a natural mismatch for SEP participants. For PtD participants, responses given in the AV condition were examined by comparing original stimuli from their native variety and manipulated AV stimuli from SEP, which we expect to represent a natural congruency for PtD participants.

3.2.1. SEP participants

As we may observe in **Figure 5**, when exposed to manipulated stimuli from their native variety, SEP participants clearly rely on intonation and thus identify the stimuli as declaratives (< 1.5).

The behavior of SEP participants under the manipulated AV condition is similar to the one observed in the PtD AV condition, as predicted in our Hypothesis 4 (**Figure 6**).

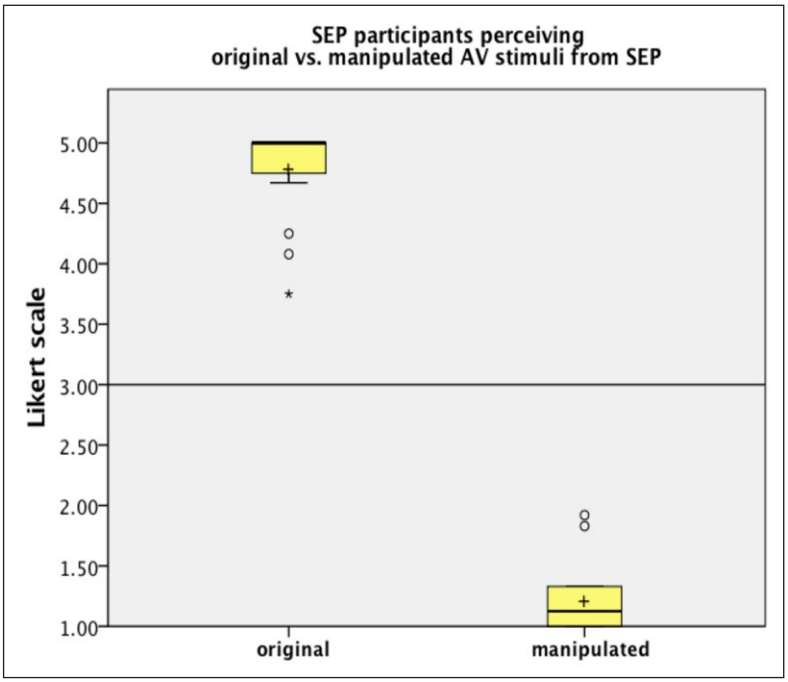

Figure 5: SEP participants: Mean responses (1 = more declarative-like; 5 = more interrogative-like) given in the AV condition—original versus manipulated stimuli from SEP. The '+' marks in the boxplots represent mean values.

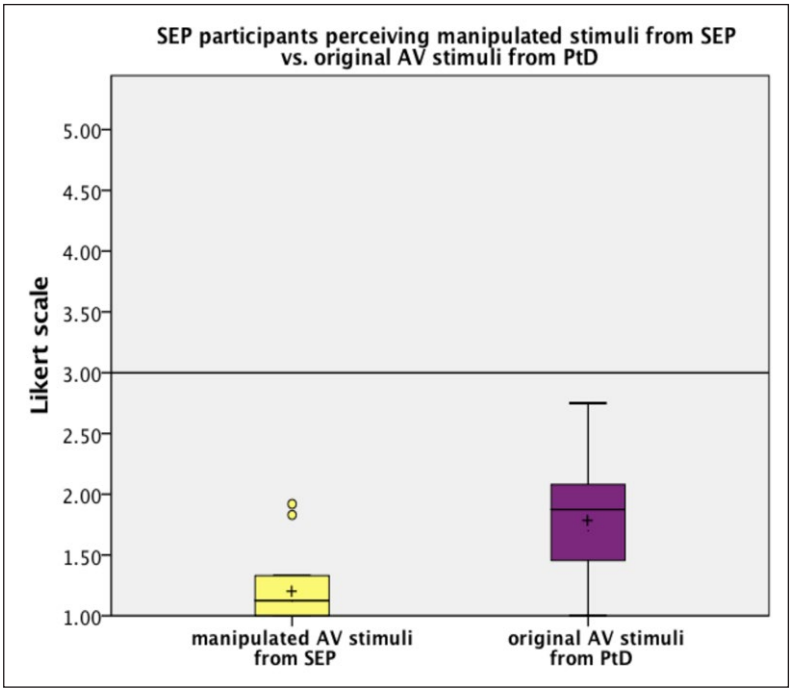

Figure 6: SEP participants: Mean responses (1 = more declarative-like; 5 = more interrogative-like) given in the AV condition—manipulated stimuli from SEP versus original stimuli from PtD (natural incongruency for SEP participants). Yellow identifies the native variety and the non-native variety is represented in purple. The '+' marks in the boxplots represent mean values.

AV stimuli produced by PtD speakers were indeed interpreted along the lines of manipulated SEP stimuli by SEP participants, i.e., as declaratives, confirming initial predictions based on the fact that the intonation of yes–no questions from PtD is the same as that of statements from SEP (H + L* L%—**Table 1**), and the visual cue (eyebrow raising) is the same for yes–no questions for both varieties. Although the tendency is the same in these

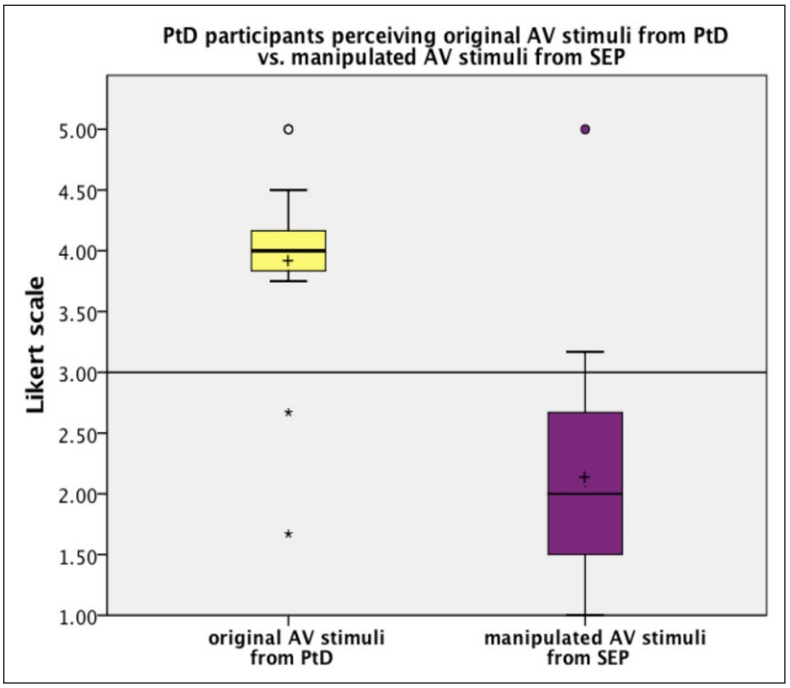

Figure 7: PtD participants: Mean responses (1 = more declarative-like; 5 = more interrogative-like) given in the AV condition—original stimuli from PtD versus manipulated stimuli from SEP (expected to be a natural congruency for PtD participants). Yellow identifies the native variety and the non-native variety is represented in purple. The '+' marks in the boxplots represent mean values.

two conditions and varieties, a paired samples t-test confirms that the responses given by SEP participants are significantly different depending on the variety they are exposed to (SEP: $M = 1.22$, $SD = .27$; PtD: $M = 1.80$, $SD = .51$; $t(19) = -5.677$, $p < .001$). By and large, responses were more spread for the stimuli from the non-native variety, which is confirmed by the highest standard deviation.

3.2.2. PtD participants

PtD participants, contrary to our Hypothesis 3, did not regard manipulated stimuli from SEP as congruent, since they were identified as declaratives regardless of the visual cue (**Figure 7**). This further supports our suggestion, put forward in Section 3.1.2, that PtD participants, as SEP participants, rely more on intonation than on facial gestures to identify the sentence type of a given stimulus.

Interestingly, although manipulated stimuli from SEP were identified as declaratives by PtD participants, a certain degree of doubt is observed, as mean responses are positioned slightly above 2 in the Likert scale. This possibly means that manipulated stimuli from SEP are difficult to categorize and that PtD participants probably took more time to respond. In the next section, a deeper analysis of reaction times is presented in order to confirm this hypothesis.

3.3. Do participants need more time to react to stimuli more difficult to categorize?

As mentioned in Section 2.3, reaction times (RTs) were also measured in order to observe whether participants need more time to react when being exposed to stimuli that are more difficult to categorize (Chen, 2003; Schneider et al., 2011). In this sense, we expect longer reaction times in the VO condition than in the AO and AV conditions for both varieties (Hypothesis 5). As for the AV condition, for SEP participants, we initially expected longer reaction times in the manipulated AV condition from SEP, like in the original AV condition from PtD (both expected to be interpreted as incongruent), than in the original

AV condition from SEP (Hypothesis 6). However, the analysis of mean responses shows that although SEP participants identify both manipulated stimuli from their native variety and original AV stimuli from PtD as declaratives, they perform significantly differently depending on the variety they are exposed to (see **Figure 6**). Indeed, although facing incongruency in both conditions, there is a higher degree of uncertainty in SEP participants' responses when perceiving non-native stimuli (closer to 2) than when exposed to native stimuli (below 1.5). For this reason, we discarded our Hypothesis 6. For PtD participants, we initially expected similar reaction times between original AV stimuli from PtD and manipulated AV stimuli from SEP (Hypothesis 7) since we considered that the manipulated AV condition from SEP was a natural congruency for PtD participants. However, the analysis of PtD participants' mean responses revealed that manipulated stimuli from SEP are regarded as incongruent instead, as they are identified as declaratives. For this reason, we also discarded our Hypothesis 7. A new approach was then followed for the analysis of reaction times, according to which we expected longer reaction times triggered by non-nativeness (and not by incongruency), for both SEP and PtD participants (Hypothesis 8).

3.3.1. Reaction times in all conditions: SEP participants

Reaction times (RTs) registered for SEP participants were analyzed with a mixed factorial ANOVA with one between subject factor (task order: AO-VO-AV, VO-AO-AV) and two within subjects factors (variety: SEP, PtD; and condition: AO, VO, AV). Similarly to the results on response type, the effect of task order was not significant, which means that reaction times were not significantly different depending on the task order performed ($F(1, 18) = .669$, $p > .1$). We found a significant effect of variety ($F(1, 18) = 10.46$, $p < .01$), with SEP participants presenting longer reaction times when facing non-native stimuli (**Figure 8**).

A significant effect of condition ($F(2, 36) = 8.23$, $p < .01$) was also found, with post-hoc tests showing that RTs in the VO condition are significantly different from those of the AO condition ($p < .05$) and of the AV condition ($p = .05$) (**Figure 9**). However, RTs in the AO condition do not significantly differ from RTs in the AV condition ($p > .1$). This result confirms our Hypothesis 5, as we expected longer reaction times in the VO condition than in the AO and AV conditions. Additionally, this result is in line with our findings for response types, as only for VO was a different pattern of responses found

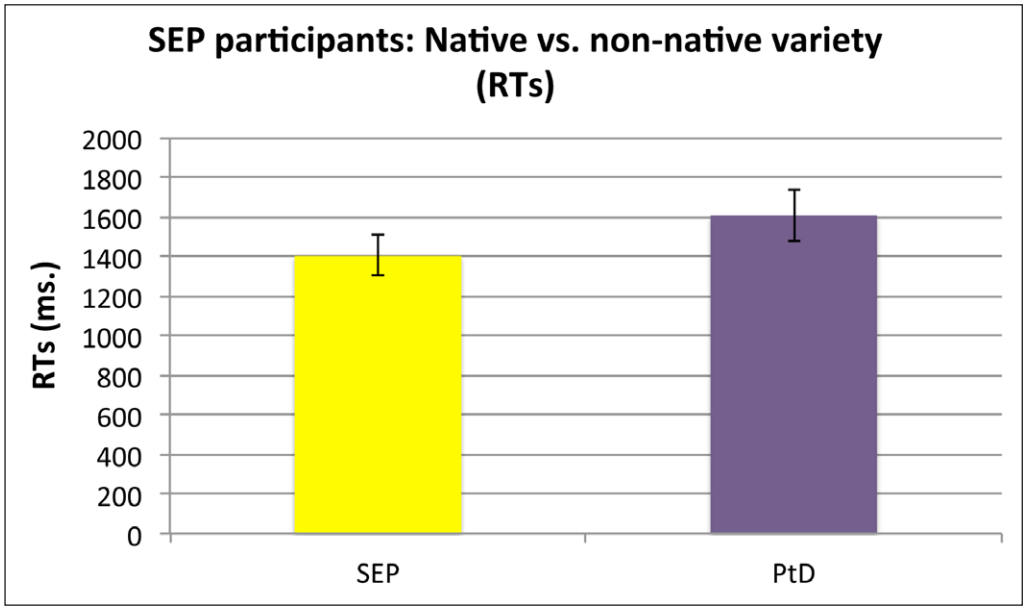

Figure 8: SEP participants: Mean reaction times (ms.) per variety perceived (SEP, PtD). Error bars represent standard error values. Yellow identifies the native variety and the non-native variety is represented in purple.

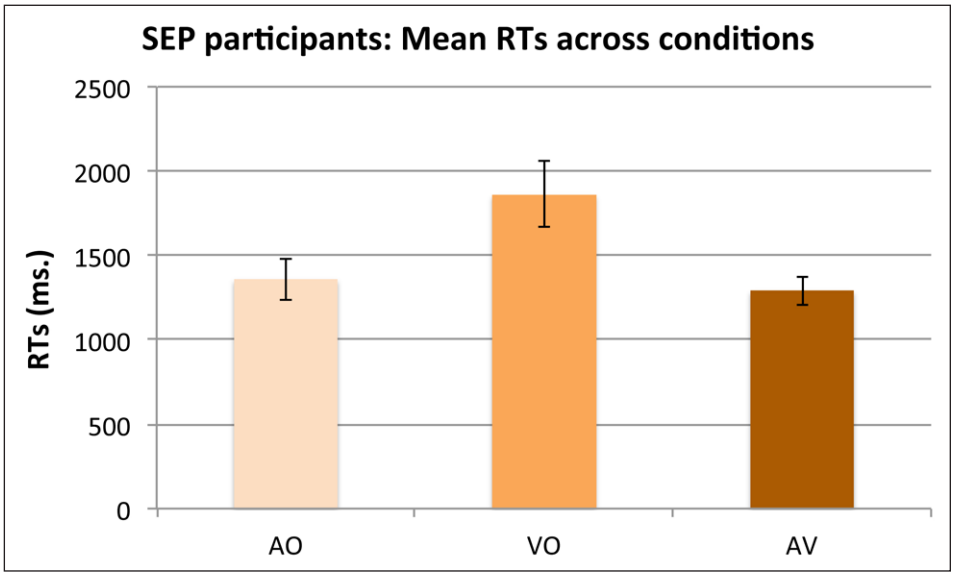

Figure 9: SEP participants: Mean reaction times (ms.) across conditions (AO, VO, AV). Error bars represent standard error values.

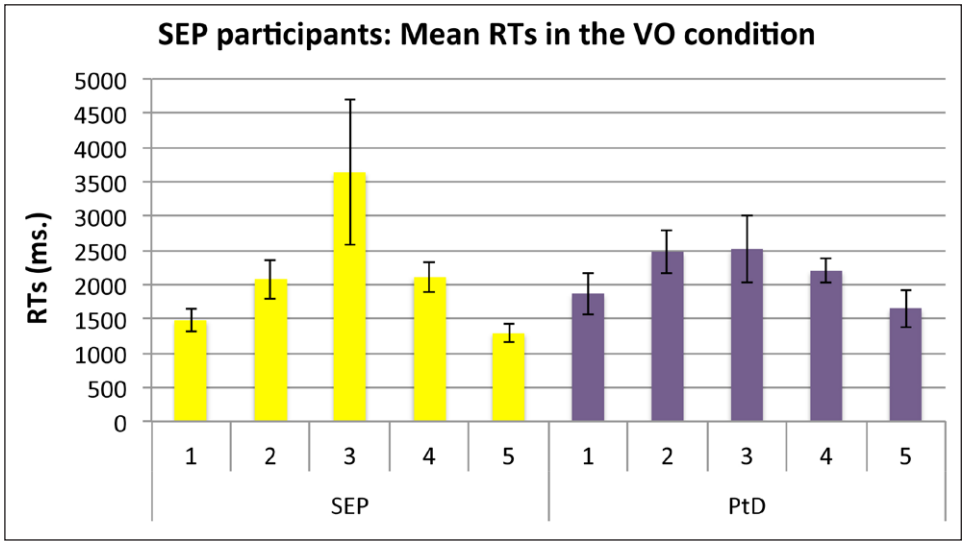

Figure 10: SEP participants: Mean reaction times (ms.) in the VO condition per each level of the Likert scale (1 = declarative, 2 = declarative-like, 3 = strong doubt, 4 = interrogative-like, 5 = interrogative) and variety perceived (SEP, PtD). Error bars represent standard error values. Yellow identifies the native variety and the non-native variety is represented in purple.

with categorization being more difficult than in the other conditions (see **Figure 3**). Also interesting, the fact that participants are exposed to a richer stimulus in the AV condition (auditory and visual cues) does not slow them down, compared to the AO condition.

Contrary to response type, reaction times do not differ across conditions depending on the variety perceived, as the interaction between these two factors (variety and condition) was not significant ($F(2, 36) = .123$, $p > .1$).

Although SEP participants are globally faster when exposed to stimuli produced by speakers from their native variety than when exposed to PtD stimuli (**Figure 8**), a closer look at the condition in which they have significantly longer reaction times (VO) revealed that reaction times registered in each level of the Likert scale are not significantly different across the varieties perceived (declarative (1): $Z = –1.138$, $p > .1$; declarative-like (2): $Z = –1.362$, $p > .1$; strong doubt (3): $Z = –.734$, $p > .1$; interrogative-like (4): $Z = –.800$, $p > .1$; interrogative (5): $Z = –1.782$, $p > .05$), as shown in **Figure 10**. This

points to response types being equally difficult to choose in this condition, in both varieties. Also interesting, we observe shorter reaction times in the extreme values of the Likert scale, i.e., when SEP participants have a high degree of certainty on their response. As the degree of certainty decreases, reaction times increase.

3.3.2. Reaction times in all conditions: PtD participants

A similar analysis was also conducted for reaction times registered for PtD participants. A mixed factorial ANOVA with one between subject factor (task order: AO-VO-AV, VO-AO-AV) and two within subjects factors (variety: PtD, SEP; and condition: AO, VO, AV) revealed a non-significant effect of task order, similarly to the results for SEP participants, indicating that the reaction times registered for the two subgroups of PtD participants were not significantly different depending on the task order they were exposed to ($F(1, 18) = .627$, $p > .1$). The only significant effect that we found was that of condition ($F(2, 36) = 9.166$, $p < .01$). Tests for contrasts show that reaction times were significantly longer in the VO condition ($F(1, 18) = 7.386$, $p < .05$), confirming our Hypothesis 5, and that reaction times do not differ between the AO and the AV condition ($F(1, 18) = 1.689$, $p > .1$), as we have observed for SEP participants as well. This is illustrated in **Figure 11**.

Mean reaction times per condition were also observed independently of the participants group (SEP and PtD). We concluded, by running a Wilcoxon test with a Bonferroni correction (significance level at $p < 0.017$), that RTs in the AV condition are not significantly lower than in the AO condition ($Z = –1.835$, $p = .033$). However, both in SEP (**Figure 9**) and in PtD (**Figure 11**) the AV condition presents the lowest RTs, thus showing that visual cues seem to play a role in the perception of sentence types, by speeding up reaction times compared to the AO condition.

Comparing results from the two groups of participants, mean reaction times across conditions were not significantly different between SEP participants and PtD participants ($F(1, 38) = .180$, $p > .1$). In contrast with SEP participants though, for PtD participants (non)nativeness did not have an impact on reaction times, i.e., reaction times registered for PtD participants were not significantly different depending on the variety they were exposed to ($F(1, 18) = .097$, $p > .1$) (**Figure 12**).

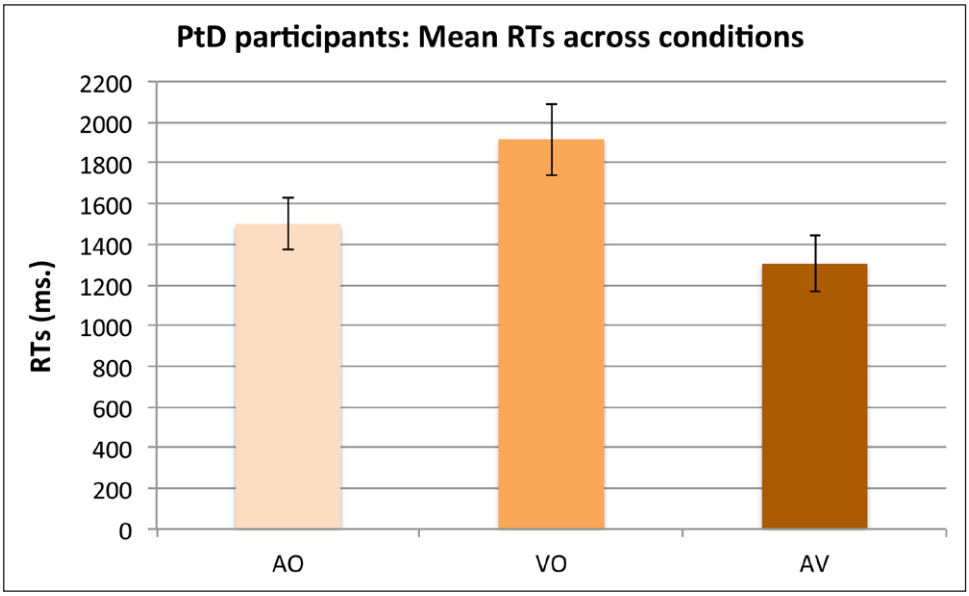

Figure 11: PtD participants: Mean reaction times (ms.) across conditions (AO, VO, AV). Error bars represent standard error values.

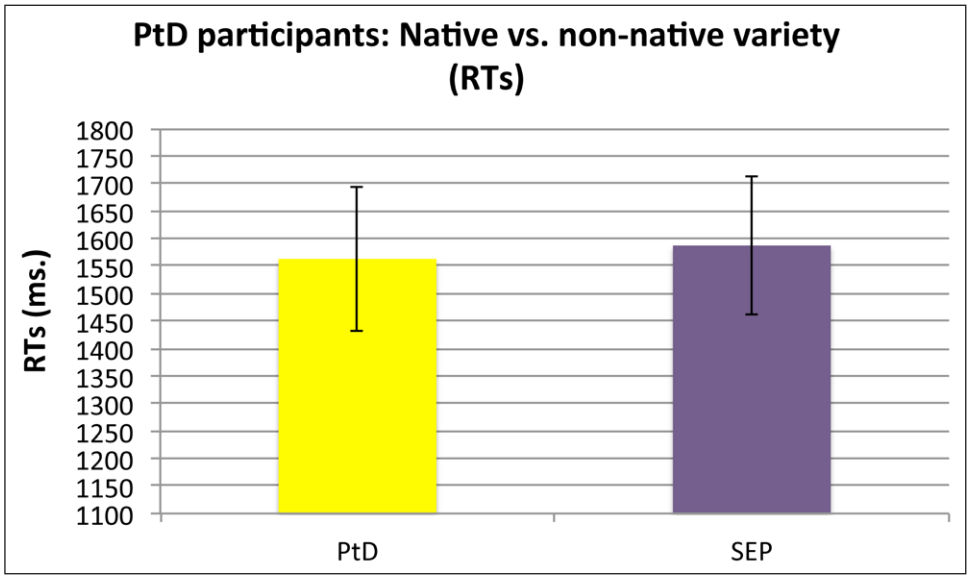

Figure 12: PtD participants: Mean reaction times (ms.) per variety perceived (PtD, SEP). Error bars represent standard error values. Yellow identifies the native variety and the non-native variety is represented in purple.

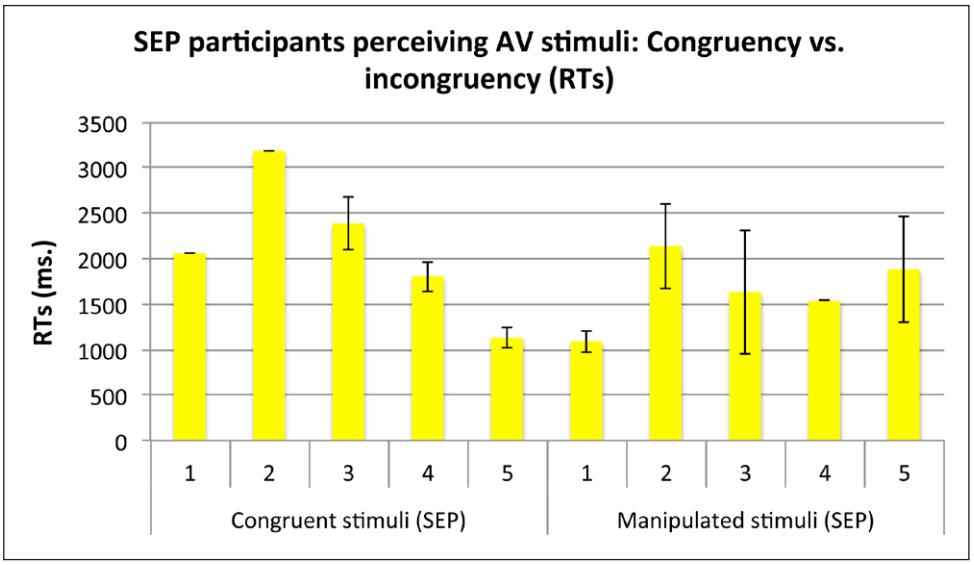

Figure 13: SEP participants: Mean reaction times (ms.) in the AV condition—original AV stimuli from SEP versus manipulated stimuli from SEP. Error bars represent standard error values.

3.3.3. Reaction times in AV condition: SEP participants

Let us now consider the hypothetical effect of incongruency in reaction times registered for SEP participants in the AV condition (**Figure 13**). In Section 3.2.1, we concluded that original AV stimuli were mainly recognized as interrogatives by SEP participants, in contrast with manipulated stimuli, mainly recognized as declaratives. By looking at reaction times in these two conditions, we now observe that SEP participants were considerably faster in identifying interrogatives (5) in the original condition (left side of the graph) and declaratives (1) in the manipulated condition (right side of the graph) when compared to the other response possibilities within each condition. This further supports the certainty of SEP listeners in their audio-guided responses (see **Figure 5**). Additionally, we observed shorter reaction times in the manipulated audiovisual condition, thus showing that, although more unsteady (which is reflected by the error bars), participants do not need more time to react even in the presence of a mismatched visual cue.

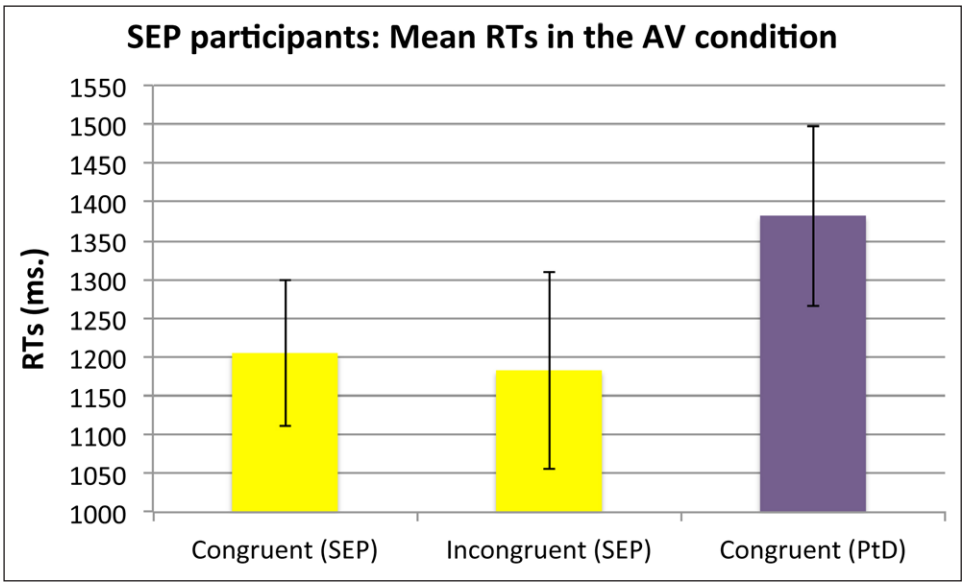

Figure 14: Mean reaction times (ms.) in the AV condition for SEP participants, by variety perceived and type of stimuli used. Error bars represent standard error values. Yellow identifies the native variety and the non-native variety is represented in purple.

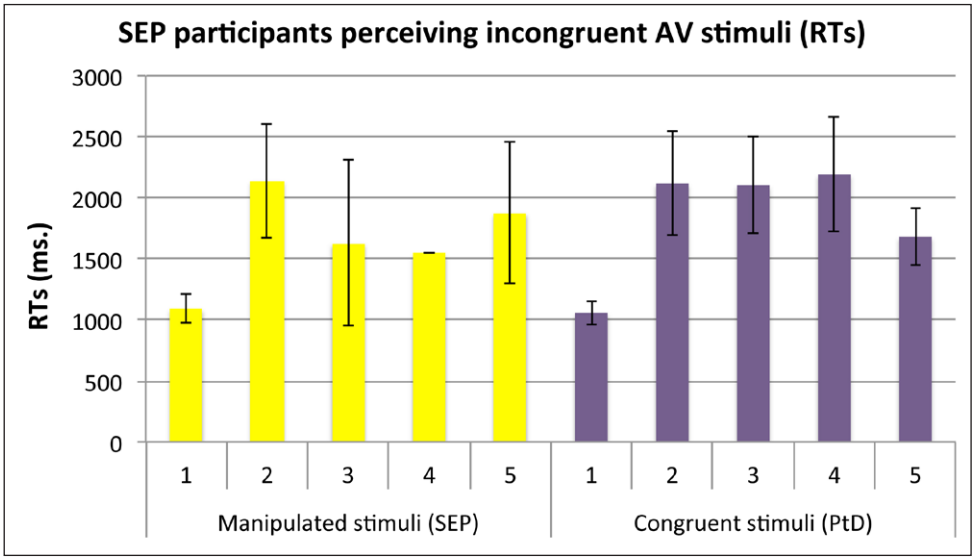

Figure 15: SEP participants: Mean reaction times (ms.) in the AV condition—manipulated stimuli from SEP versus original stimuli from PtD. Error bars represent standard error values. Yellow identifies the native variety and the non-native variety is represented in purple.

We then compared mean reaction times within the AV condition, independently of the response type (**Figure 14**). We concluded that SEP participants need more time to react to original AV stimuli produced by PtD speakers (non-native stimuli), which are naturally incongruent for them, than to manipulated AV stimuli from their native variety. This outcome thus confirms our Hypothesis 8, i.e., non-nativeness (but not incongruency) triggers longer reaction times.

Interestingly, manipulated stimuli from SEP and original stimuli from PtD were both mainly perceived as declaratives (see **Figure 6**) and reaction times are indeed shorter for the response value 1 in both conditions than for the other response possibilities within the Likert scale (**Figure 15**), thus reinforcing the high certainty that SEP participants have in their response.

3.3.4. Reaction times in AV condition: PtD participants

As for PtD participants, in Section 3.2.2, we observed that manipulated stimuli from SEP, contrary to our expectation, are recognized as declaratives (see **Figure 7**). By analyzing reaction times, a Wilcoxon test now allows us to conclude that reaction times registered for PtD participants were not significantly different when categorizing original AV stimuli from their native variety as interrogatives and manipulated stimuli from SEP as declaratives ($Z = -1.285$, $p > .1$). The fact that 'doubt' responses (3) in the manipulated AV condition from SEP present shorter reaction times than the same response type in the original AV condition from their native variety (**Figure 16**) shows that participants do not need more time to react even in the presence of a mismatched visual cue (declaratives with eyebrow raising).

When we compare mean reaction times within the AV condition independently of the response type (**Figure 17**), we conclude that our hypothesis that PtD participants need

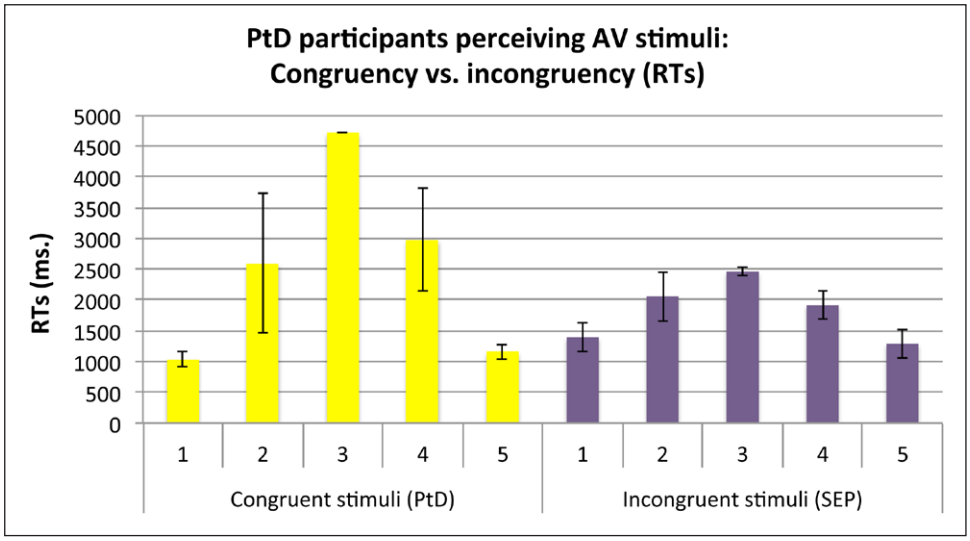

Figure 16: PtD participants: Mean reaction times (ms.) in the AV condition—original stimuli from PtD versus manipulated stimuli from SEP. Error bars represent standard error values. Yellow identifies the native variety and the non-native variety is represented in purple.

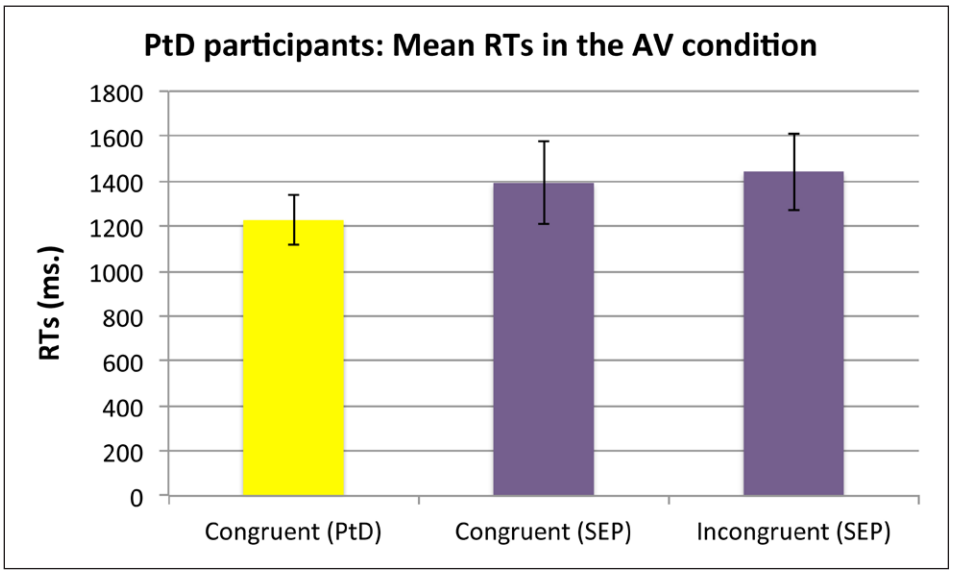

Figure 17: Mean reaction times (ms.) in the AV condition for PtD participants, by variety perceived and type of stimuli used. Error bars represent standard error values. Yellow identifies the native variety and the non-native variety is represented in purple.

more time to react to manipulated stimuli from SEP than to original stimuli from their native variety is confirmed. However, more than incongruency, we observe that non-nativeness seems to trigger longer reaction times, thus confirming our Hypothesis 8, as PtD participants, similarly to SEP participants (**Figure 14**), also need more time to react to non-native AV stimuli than to native AV stimuli.

To confirm whether (non)nativeness of stimuli significantly affected reaction times, we ran two Wilcoxon tests (respectively, for SEP participants and for PtD participants), each comparing reaction times in three pairs. A Bonferroni correction was applied, resulting in a significance level set at $p < 0.017$. For SEP participants, we concluded that reaction times in native (but manipulated, i.e., incongruent) AV stimuli are significantly shorter than those obtained in the non-native (also incongruent) condition ($Z = –2.427$, $p = .007$). This result supports our suggestion that (non)nativeness, and not (in)congruency, affected reaction times: (i) The non-native condition (AV stimuli from PtD) represents a natural incongruency for SEP and reaction times registered under this condition were significantly higher than in the condition with manipulated stimuli from SEP; (ii) although not significantly different, reaction times in the condition with manipulated stimuli from SEP are even shorter than in the condition with original AV stimuli ($Z = –.224$, $p > .1$). For PtD participants, we concluded that reaction times in the original AV condition from PtD are not significantly different from reaction times in the condition with manipulated stimuli from SEP, which is the non-native condition with the longest reaction times ($Z = –1.624$, $p > .1$). Although further data (and from other varieties) need to be studied from a visual point of view in articulation with prosody, we could probably hypothesize that this difference between SEP and PtD participants might be due to a familiarity/novelty effect. Indeed, PtD participants are constantly exposed to the SEP variety mainly through the media, and the reverse is not true for SEP participants. This probably causes an asymmetry between the two varieties, thus being a natural limitation of our study since exposure to the Standard variety by speakers of other varieties of EP, without the converse situation, is naturally driven by social and cultural conditions.

4. Conclusion

In this research, we inspected the role of intonation and visual cues in the perception of statements and questions in two varieties of European Portuguese (SEP and PtD) previously shown to convey sentence type contrasts by different uses of intonational means and/or facial gestures, in particular eyebrow movements. We concluded that both SEP and PtD participants' perception of sentence type differences depends on the variety (SEP, PtD) and on the condition (AO, VO, AV) they are exposed to. Crucially, participants from both groups rely more on intonation than on visual cues when asked to identify sentence types whether they are perceiving their native variety or the other variety. The results obtained in the VO condition also reinforces this observation: In the absence of auditory cues, participants from SEP and PtD float in their decision, by classifying VO stimuli either as declaratives or interrogatives (with a tendency for declaratives), independently of the variety (native or non-native) they are exposed to. The fact that SEP and PtD participants may consider the VO stimuli as declaratives led us to the hypothesis that the two groups of participants are identifying the eyebrow raising as a focus marker (and not as a question marker), which is used in both varieties in the production of narrow focused statements. Another interpretation is that visual cues alone are not enough for sentence type identification and thus participants either respond at chance level or use a default response strategy, which is declarative. Future research is needed, including, for instance, a labeling task, to address this question.

Intonation was also the crucial cue used in the AV condition with manipulated stimuli from SEP, and also in the AV condition with original stimuli produced by PtD speakers, which stand as a natural incongruency for SEP participants. Unexpectedly, manipulated stimuli from SEP do not represent a natural congruency for PtD participants, as they dominantly classified these stimuli as declaratives. Besides showing that PtD participants, as SEP participants, rely more on intonation than on visual cues, this finding also suggests that there is a clear difference between the falling intonation of statements from SEP and the falling intonation of neutral yes–no questions from PtD. Consequently, we could hypothesize that PtD participants are simply able to recognize that they are exposed to stimuli from non-native speakers (as differences may be found at the segmental level, as well), and that, for this reason, they discard the stimuli as being natural interrogatives in their native variety. However, the fact that PtD participants are highly successful in distinguishing declaratives and interrogatives from their native variety (both produced with a falling intonation), based on AO stimuli presented in the training phase, led us to conclude that, more than identifying non-native stimuli as such, they have a natural ability to distinguish two intonation contours apparently identical in terms of phonology and phonetic realization. Importantly, there may be subtle differences in phonetic realization that are used by native speakers of PtD to distinguish the two contours. This is a question to be explored in future research. In sum, the fact that intonation is dominant confirms previous findings for Dutch (Swerts & Krahmer, 2008): The eyebrow movement cue had an effect on accent perception in Dutch, which was comparatively much smaller than the intonational effect.

We also observed that variety and condition alone (but not in interaction) have a significant effect on reaction times in SEP participants' responses, whereas for PtD participants, only condition significantly affects reaction times. VO condition is the most difficult one, i.e., the one presenting the longest reaction times for both SEP and PtD participants, which was expected due to the type (and spread) of responses given by participants. This finding also points to the strength of intonation over visual cues. Additionally, we concluded that SEP participants do not necessarily need more time to react when being exposed to incongruent stimuli, as suggested by previous studies such as Chen (2003), and Schneider et al. (2011): Reaction times of SEP participants were shorter when exposed to manipulated stimuli from their native variety, than when exposed to original stimuli.

Both SEP and PtD participants presented longer reaction times when exposed to AV stimuli from another variety, especially when facing incongruency from the perspective of their native variety. However, in SEP, reaction times of participants' responses are significantly different when exposed to native and non-native stimuli, whereas in PtD such statistical difference was not observed. In sum, although SEP and PtD participants seem to be globally more sensitive to intonation than to visual cues for the distinction between sentence types produced by a (non)native variety, the fact that SEP participants, but not PtD, presented significantly different reaction times when exposed to non-native stimuli is explained by the familiarity effect: PtD participants are constantly in contact with SEP (through media), which is not the case of SEP participants.

Finally, although intonation overrides visual cues (along the lines of Krahmer & Swerts, 2004; Swerts & Krahmer, 2008), our findings also suggest that visual cues play a role in structural or linguistic marking (like sentence type and pragmatic meaning) as shown by the considerably lower reaction times in the AV condition comparatively to the AO condition, when running analyses over all participants.

Additional File

The additional file for this article can be found as follows:

- Examples of stimuli in the AO, VO, and AV conditions for each variety, and the AV mismatch condition for SEP. DOI: https://doi.org/10.5334/labphon.110.s1

Acknowledgements

This research was developed within the Post-Doctoral grant BPD/94695/2013 and the project grant InAPoP—Interactive Atlas of the Prosody of Portuguese (PTDC/CLE-LIN/119787/2010), funded by Fundação para a Ciência e a Tecnologia, Portugal. We thank Nuno Paulino and Pedro Oliveira for data collection in Ponta Delgada (Azores), as well as the local support of Universidade dos Açores. We are also grateful to all participants from Lisbon and Ponta Delgada (Azores) in the perception tasks, as well as to the staff in Associação Alternativa (Ponta Delgada, Azores) for helping with participants' recruitment. Last but not least, we would like to acknowledge the precious comments and suggestions of Laura de Ruiter, and the important contribution of the anonymous reviewers for this paper.

Competing Interests

The authors have no competing interests to declare.

References

Bangerter, A. 2004. Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science*, *15*, 415–419. DOI: https://doi.org/10.1111/j.0956-7976.2004.00694.x

Borràs-Comes, J., Kaland, C., Prieto, P., and Swerts, M. 2014. Audiovisual correlates of interrogativity: A comparative analysis of Catalan and Dutch. *Journal of Nonverbal Behavior, 38*, 53–66. DOI: https://doi.org/10.1007/s10919-013-0162-0

Borràs-Comes, J., and Prieto, P. 2011. 'Seeing tunes.' The role of visual gestures in tune interpretation. *Laboratory Phonology, 2*(2), 355–380. DOI: https://doi.org/10.1515/labphon.2011.013

Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., and Espesser, R. 1996. About the relationship between eyebrow movements and F0 variations. *Proceedings of the International Conference on Spoken Language Processing*, *4*, 2175–2178. Philadelphia.

Chen, A. 2003. Reaction time as an indicator of discrete intonational contrasts in English. *Proceedings of Eurospeech 2003*, 97–100.

Crespo-Sendra, V., Cruz, M., Silva, J., and Frota, S. 2014. Asking questions across Portuguese varieties. *Talk given in the 6th International Conference on Tone and Intonation in Europe (TIE)*, September 10–12, University of Utrecht, Netherlands.

Crespo-Sendra, V., Kaland, C., Swerts, M., and Prieto, P. 2013. Perceiving incredulity: The role of intonation and facial gestures. *Journal of Pragmatics, 47*, 1–13. DOI: https://doi.org/10.1016/j.pragma.2012.08.008

Cruz, M. 2013. Prosodic variation in European Portuguese: Phrasing, intonation and rhythm in central-southern varieties (Doctoral Dissertation). Lisbon, Portugal: University of Lisbon.

Cruz, M., and Frota, S. 2014. Accents on the face? Visual prosody in varieties of Portuguese. *Poster presented in the 6th International Conference on Tone and Intonation in Europe (TIE)*, September 10–12, University of Utrecht, Netherlands.

Cruz, M., Swerts, M., and Frota, S. 2015. Variation in tone and gesture within language. In: The Scottish Consortium for ICPhS 2015. *Proceedings of the 18th International*

Congress of Phonetic Sciences. Glasgow, UK: The University of Glasgow. Paper number 452 retrieved from: http://www.icphs2015.info/pdfs/Papers/ICPHS0452.pdf.

de Ruiter, J. P., Bangerter, A., and Dings, P. 2012. Interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. *Topics in Cognitive Science, 4*(2), 232–248. DOI: https://doi.org/10.1111/j.1756-8765.2012.01183.x

Ekman, P., Friesen, W. V., and Hager, J. C. 2002. *Facial Action Coding System*. Salt Lake City, UT: A Human Face.

Félix-Brasdefer, J. C. 2010. Data collection methods in speech act performance: DCTs, role plays, and verbal reports. In: Usó-Juán, E., and Martinéz-Flor, A. (eds.), *Speech Act Performance: Theoretical, Empirical, and Methodological Issues*, 41–56. Amsterdam/Philadelphia: John Benjamins Publishing.

Frota, S. 2014. The intonational phonology of European Portuguese. In: Jun, S.-A. (ed.), *Prosodic Typology II*, 6–42. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199567300.003.0002

Frota, S., and Cruz, M. (coords). (2012–2015). *Interactive Atlas of the Prosody of Portuguese Webplatform*. http://labfon.letras.ulisboa.pt/InAPoP/.

Frota, S., Cruz, M., Fernandes-Svartman, F., Collischonn, G., Fonseca, A., Serra, C., Oliveira, P., and Vigário, M. 2015a. Intonational variation in Portuguese: European and Brazilian Varieties. In: Frota, S., and Prieto, P. (eds.), *Intonation in Romance*, 235–283. Oxford: Oxford University Press.

Frota, S., Oliveira, P., Cruz, M., and Vigário, M. 2015b. *P_ToBI: Tools for the Transcription of Portuguese Prosody*. Lab. Fonética, CLUL/FLUL. ISBN: 978-989-95713-9-6. http://labfon.letras.ulisboa.pt/InAPoP/P-ToBI/.

Granström, B., and House, D. 2004. Audiovisual representation of prosody in expressive speech communication. In: Bel, B., and Marlin, I. (eds.), *Proceedings of the International Conference on Speech Prosody 2004*, 393–396.

House, D. 2002. Intonation and visual cues in the perception of interrogative mode in Swedish. *Proceedings of the International Conference on Spoken Language Processing*, 1957–1960.

Kasper, G., and Dahl, M. 1991. Research methods in interlanguage pragmatics. *Studies in Second Language Acquisition, 13*, 215–247. DOI: https://doi.org/10.1017/S0272263100009955

Krahmer, E., Ruttkay, Z., Swerts, M., and Wesselink, W. 2002. Pitch, eyebrows and the perception of focus. In: Bel, B. (ed.), *Proceedings of the Speech Prosody 2002*, 443–446.

Krahmer, E., and Swerts, M. 2004. More about brows. In: Ruttkay, Z., and Pelachaud, C. (eds.), *From Brows to Trust: Evaluating Embodied Conversational Agents*, 191–216. Dordrecht: Kluwer Academic Publishers. DOI: https://doi.org/10.1007/1-4020-2730-3_7

Krahmer, E., and Swerts, M. 2007. The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language, 57*(3), 396–414. DOI: https://doi.org/10.1016/j.jml.2007.06.005

Melinger, A., and Levelt, W. J. M. 2004. Gesture and the communicative intention of the speaker. *Gesture, 4*, 119–141. DOI: https://doi.org/10.1075/gest.4.2.02mel

Purson, A., Santi, S., Bertrand, R., Guaïtella, I., Boyer, J., and Cavé, C. 1999. The relationships between voice and gesture: Eyebrow movements and questioning. *Proceedings of European Conference of Speech Communication and Technology, 6*, 1735–1739. Budapest, Hungary.

Ramus, F., and Mehler, J. 1999. Language identification with suprasegmental cues: A study based on speech resynthesis. *JASA, 105*, 512–521. DOI: https://doi.org/10.1121/1.424522

Schneider, K., Dogil, G., and Möbius, B. 2011. Reaction time and decision difficulty in the perception of intonation. *Proceedings of Interspeech 2011*, 2221–2224. ISCA, Italy.

So, W. C., Kita, S., and Goldin-Meadow, S. 2009. Using the hands to identify who does what to whom: Gesture and speech go hand-in-hand. *Cognitive Science, 33*, 115–125. DOI: https://doi.org/10.1111/j.1551-6709.2008.01006.x

Srinivasan, R. J., and Massaro, D. W. 2003. Perceiving from the face and voice: Distinguishing statements from echoic questions in English. *Language and Speech, 46*(1), 1–22. DOI: https://doi.org/10.1177/00238309030460010201

Swerts, M., and Krahmer, E. 2004. Congruent and incongruent audiovisual cues to prominence. In: Bel, B., and Marlin, I. (eds.), *Proceedings of the International Conference on Speech Prosody 2004*, 69–72.

Swerts, M., and Krahmer, E. 2006. The importance of different facial areas for signalling visual prominence. *Proceedings of the International Conference on Spoken Language Processing* (Interspeech 2006). Pittsburgh, PA: ISCA.

Swerts, M., and Krahmer, E. 2008. Facial expression and prosodic prominence: Effects of modality and facial area. *Journal of Phonetics, 36*(2), 219–238. DOI: https://doi.org/10.1016/j.wocn.2007.05.001

How to cite this article: Cruz, M., Swerts, M., and Frota, S. 2017 The role of intonation and visual cues in the perception of sentence types: Evidence from European Portuguese varieties. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 8(1): 23, pp. 1–24, DOI: https://doi.org/10.5334/labphon.110

Submitted: 24 August 2017 **Accepted:** 24 August 2017 **Published:** 11 September 2017

]u[*Laboratory Phonology: Journal of the Association for Laboratory Phonology* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS