The perception of visual phonetic information: Eye-tracking while searching for segmental versus prosodic cues

Xizi Deng, Erin Jastrzebski, Elise McClay, H. Henny Yeung, Yue Wang

Simon Fraser University

ABSTRACT

This eye-tracking study investigated visual scanning of a face when searching for different types of phonetic information. Twenty-four native English perceivers heard two audio sentences and then saw a silent talking face. The audio sentences sometimes differed in segmental and/or prosodic information, and perceivers had to decide which sentence the silent face articulated. Several types of prosodic differences were tested (contrastive word focus, phrasal bracketing, and intonation). Results consistently showed more overall mouth looking when identifying segments than prosody, and there were distinct patterns of shifts in looking between the mouth and non-mouth areas for processing prosody versus segments, although this was not consistent across different types of prosodic differences. Overall, these results show that visual perception of phonetic cues is affected by the type of phonetic cue itself: More mouth looking for segments than prosody. Moreover, we also observed variable looking patterns for different types of visual prosody, suggesting avenues for future research.

Keywords: audio-visual processing, speech, perception, phonetic, prosody, eye-tracking

1. INTRODUCTION

Useful speech information is encoded in many sensory modalities besides audition, like visual cues from an interlocutor’s face, which can enhance auditory speech comprehension [1]–[15]. One leading reason why visual cues enhance auditory speech comprehensibility is that critical speech cues are embedded in particular areas in the face. For example, a number of studies have suggested that adults pay attention to the mouth during speech tasks when identifying segments of consonant-vowel (CV) syllables [16], or when identifying which face is the source of auditory speech [17]. For example, a classic study, Lansing & McConkie (1999), found that participants’ gaze duration at the lower part of the face was longer when making decisions about segments than intonation. These results suggest that the mouth area of a talking face provides useful cues for decoding segmental information. On the other hand, prosodic information also carries critical cues in speech communication by accentuating pragmatic or syntactic information, and facial movement of the whole head or around the eyes may be helpful for linguistic prosody [7]–[14]. Correspondingly, Swerts & Krahmer (2008) found that the upper part of a face has stronger cue value in the detection of word-level prosodic prominence in a three-word sentence. Similarly, the production of visual cues in the upper part of the face was found to be more related to distinguishing echoic questions from statements [10], as these features are linked with variations in F0 [18]. However, many others have found instead that the lower part of the face [10], including chin-opening displacement [19], lip area and jaw-opening [20], can be more facilitative of processing contrastive word focus than other facial gestures, even though head movement and eye-brow displacement exerted a smaller but still significant contribution to the correct perception of this word focus [19]. Overall, since prosody may convey information in a wide range of linguistic domains (from local word to global sentence levels), research has not been conclusive as to how a face is scanned for different types of prosodic information and how looking patterns differ for prosody and segments.

The current study expands upon the previous studies in several novel ways. First, we conducted a more comprehensive investigation on the looking patterns when processing segments versus three types of prosodic information differing in the linguistic domains that the information is conveyed: contrastive word focus (more local), sentence intonation (more global), and intermediately: phrasal bracketing, where constituents of a sentences were prosodically grouped together in different ways. Second, we examined not only how long participants looked at one particular facial area but also investigated the overall scanning dynamics on the visual face by conducting a growth curve analysis on looking patterns over time (or how eye gaze patterns changed as the distribution of critical information changed). Third, we focused on how perceivers adjusted their scanning strategies as task difficulty increased across conditions by varying the concordance of prosodic and segmental information.

In the current study, native English speakers scanned a talking face while looking for segmental and/or different prosodic information. We hypothesized more overall fixations to the mouth when identifying segmental differences and comparatively fewer fixations to the mouth when identifying prosodic differences. Second, we predicted that word-focus would show more and/or a faster shift of attention between the mouth and non-mouth areas when identifying prosodic information compared to when identifying segmental information, since visual cues from the both the lower part and upper part of the face contributed to the perception of prosody in these sentence types [9], with the upper face area being a bit less salient than the lower area [19]. However, we also predicted no such gaze shifting differences for the intonation type: Participants might fixate their eye gaze at just the mouth when searching for segmental information in this sentence type, and just the eyes when searching for intonational information [9]. Thus, we predicted less of a need to shift between the eyes and mouth in this condition [10]. Finally, we predict that the looking patterns for phrasal bracketing to be intermediate to those for word focus and intonation.

2. METHOD

2.1. Participants

Twenty-four native Canadian English speakers (5 males; 19 females), aged from 19 to 24 (mean = 21.2, SD = 2.3), were recruited. All participants had normal or corrected-to-normal vision, and none had hearing impairments or language-related pathologies.

2.2. Stimuli and Procedure

Each trial consisted of two auditory sentences and one silent video, which showed a brief clip of a face articulating an auditory sentence. A female, monolingual speaker of Canadian English recorded all auditory and visual stimuli, and these trials were arranged into *Segment*, *Prosody*, and *Both* conditions in word focus, phrasal bracketing and intonation sentences (see Figure 1 and Table 1). In experimental trials, auditory sentences differed from each other, and the silent video only matched one of them: In the *Prosody* condition, the two auditory sentences differed such that words were stressed using contrastive focus (word focus), or words were prosodically grouped together in distinct ways (phrasal bracketing), or one sentence was a question while the other was a statement (intonation), but both sentences had the same words; in the *Segment* condition, sentences differed in which words were used (having visually distinct consonants and vowels), but both sentences had the same patterns of prosody; in the *Both* condition, sentences difference in both prosody and segments. All stimuli were created in quadruplet groups, hereafter referred to as "topics." An example of stimuli from three topics are presented in Table 1.

During the experiment, each participant was calibrated in an Eyelink 1000 eye-tracker in binocular mode (SR Research) using a standard nine-point procedure on a 1280 * 1024 pixel screen. Afterwards, participants were informed that they would be presented with two auditory sentences followed by a single silent video that matched one (or both) of the audio-only stimuli in each trial. As shown in Figure 1, each trial began with presentation of both auditory sentences (with a 300 ms ISI). This was followed by a 500 ms pause, and then a fixation point turned blue before a video of a silent talking face was played. During video playback participants could select a keyboard button to indicate which of the audio recordings (or both) matched the video they saw (left arrow button indicating the first audio recording, right arrow button indicating the second and down arrow button indicating both), in which case, the video continued to play to the end, and then advanced automatically to the next trial. If no responses were made during the video, they were instructed to make a choice when the video ended, which then triggered the beginning of the next trial. Participants could thus respond anytime from the beginning to the end of stimulus presentation.

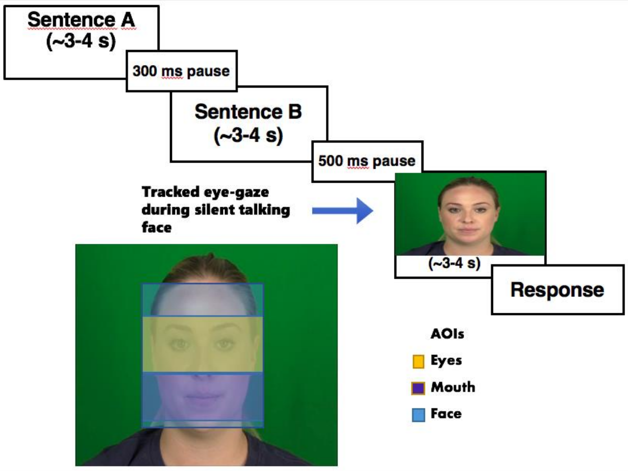

Figure 1: The structure of a trial. Participants first heard a sentence followed by another. Then, they saw a silent

talking face and began scanning the face for useful speech cues in order to make a decision about which sentence matched the video. The trial ended when the response was recorded, or when the video ended (whichever was later).

Word focus	Condition
No, **JESS** found her **PLAIN** dress for **KYLA'S** wedding.	Reference
No, Jess **FOUND** her plain **DRESS** for Kyla's **WEDDING.**	Prosody
No, **ANNE** found her **BLUE** dress for **JIMMY'S** wedding.	Segment
No, Anne **FOUND** her blue **DRESS** for Jimmy's **WEDDING.**	Both
Phrasal bracketing	
Vanessa and **[Alexander or Katrina]** will pick up the popcorn.	Reference
[Vanessa and Alexander] or Katrina will pick up the popcorn.	Prosody
Christina and **[Zachariah or Maria]** will pick up the popcorn.	Segment
[Christina and Zachariah] or Maria will pick up the popcorn.	Both
Intonation	
Sally will make a call **[before the salon closes?]**	Reference
Sally will make a call **[before the salon closes.]**	Prosody
Sally will make a call **[after she finishes lunch?]**	Segment
Sally will make a call **[after she finishes lunch.]**	Both

Table 1: Stimuli from an example topic in the three sentence types.

3. RESULTS

We first calculated the mean of raw proportion fixations to the mouth relative to the whole face (Prop) and then performed the empirical logit transformation (Elog) for the raw values through the equation below, in order to avoid the boundness issue of proportion measures, which violates the assumptions of statistical parametric models [21]:

$$\text{Elog} = \text{Log}(\text{Prop} + e/(1 - \text{Prop} + e))$$

The critical analysis was done in the time window from the onset of the video to the point when participants made their responses. This was done for only correct trials and averaged independently for all the three conditions in each sentence type.

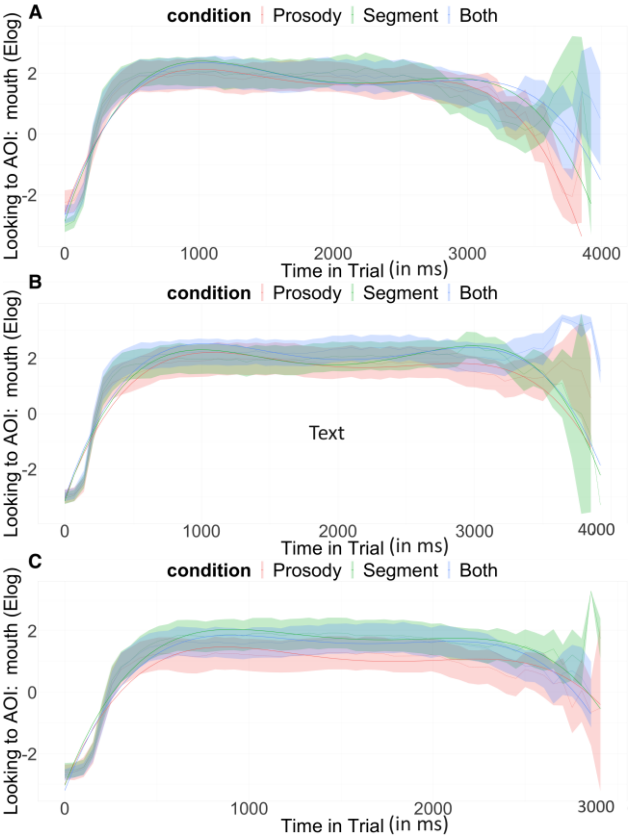

Figure 2: Elog values of fixations to the mouth (y -axis, averaged across participants and items), along the time in a trial (x-axis). Plot A, Plot B and Plot C reflect raw Elog values and model fits (solid lines) during the time window of interest in word-focus, phrasal bracketing and intonation type respectively. Note that model were calculated across 57 time bins.

Figure 2 exhibits the average Elog values of proportion fixations to the mouth during *Segment*, *Prosody* and *Both* conditions in each of the three different sentence types (word focus, phrasal bracketing, intonation). For statistical analysis, three growth curve analyses of the Elog values of proportion fixations to the mouth were performed within the above-mentioned time window for the three sentence types using R [22]. We used fourth-order orthogonal polynomial models, where Elog values were the dependent variable, and condition (*Segment*, *Prosody*, *Both*) was a fixed factor, with participant and topic (i.e. item) specified as random effects. The maximal model was performed first and if it did not converge, the random effects structure was simplified following Mirman (2014) (see https://dmirman.github.io/GCA/GCA2019.html). Models were fit using maximum-likelihood estimation [23] and estimates for the parameters used in significance tests for model fit components in the pair-wise comparisons were calculated from Satterthwaite's method.

Results for the word focus model showed that there was a significant effect of condition on the intercept (X^2 = 22.14, *p* < .001), linear (X^2 = 10.87, *p* = .004), quadratic (X^2 = 8.37, *p* = .02), cubic (X^2 = 48.09, *p* < .001) and quartic terms (X^2 = 10.07, *p* = .007). As shown in Figure 2a, there were greater overall fixations to the mouth in *Segment* condition compared to *Prosody* condition, which is also confirmed by the main effect of this difference on the intercept (estimate = .35, *p* < .001). More importantly, the significant effect of condition on the cubic term (estimate = 2.16, *p* < .001) revealed that there was a faster and more transient shift of looking to the mouth, then away from the mouth and back to the mouth again in *Prosody* condition than that in *Segment* condition, because the cubic (and quartic) term reflects the steepness of the curve around the inflection points at the extremities and a greater parameter estimate indicates a shallower curve [23]. When comparing the fixation curves in the *Both* condition, there was again a more acute change of looking in the *Prosody* condition than in the *Both* condition (quartic: estimate = 1.05, *p* = .004). However, the *Segment* condition showed the most abrupt shift of looking between the mouth and non-mouth area, relative to *Both* condition (quartic: estimate = .85, *p* = .01).

Results for the phrasal bracketing model showed that there was a significant effect of condition on the intercept (X^2 = 122.94, *p* < .001), linear (X^2 = 17.45, *p* < .001), quadratic (X^2 = 14.88, *p* < .001), and quartic terms (X^2 = 14.21, *p* < .001), indicating overall differences among these conditions. When comparing the differences between *Segment* and *Prosody* condition, again, there was a significant effect of condition on the cubic (estimate = -1.03, *p* = .03) and quartic terms (estimate = -1.27, *p* < .001), indicating that there was a more abrupt change of looking in *Prosody* condition than in *Segment* condition, between the mouth and non-mouth area. Likewise, when comparing the fixation curves in the *Both* condition versus the other two, there was a faster shift of looking in the *Both* condition than in the *Prosody* condition (quartic: estimate = -.85, *p* = .009) and a marginally faster shift of looking in *Segment* condition than *Both* condition (cubic: estimate = .87, *p* = .05).

Results for the intonation model showed that there was a significant effect of condition on the intercept (X^2 = 471.33, *p* < .001), linear (X^2 = 38.72, *p* < .001), quadratic (X^2 = 36.57, *p* < .001), and quartic terms (X^2 = 6.67, *p* = .04). When comparing the differences between *Segment* and *Prosody* condition, there was only a significant main effect of this difference on the intercept (estimate = .50, *p* < .001) and on the quadratic term (estimate = -1.58, *p* < .001), indicating more overall fixations to the mouth in *Segment* condition compared to *Prosody* condition, and an overall less steep fixation curve around the central inflection point, which reflects the participants initially attended to and eventually moved away from the mouth area more slowly in *Prosody* condition than in *Segment* condition. However, there were no significant cubic (estimate = -.32, *p* = .33) or quartic terms (estimate = -.41, *p* = .12) in the intonation type, reflecting a more parallel looking pattern between the *Segment* and *Prosody* conditions. When comparing the *Both* condition to the other two, there was also relatively more constant and stable looking, though less to the mouth, in *Prosody* condition than in *Both* condition (quartic: estimate = -.66, *p* = .01). However, there were no significant differences between the shapes of the curves around the inflection points in *Both* and *Segment* condition (quartic: estimate = -.26, *p* = .30).

4. DISCUSSION

The results showed a greater emphasis on looking to the mouth for decoding segmental information, relative to prosodic information, supporting our first hypothesis. However, this emphasis was also manifested in different ways across the three prosodic types. For word focus, there was faster shift between the mouth and non-mouth area in *Prosody* condition than in *Segment* condition, in line with our second hypothesis. For the phrasal bracketing type, differences across conditions may instead have indicated the necessity of searching the mouth for useful visual cues of critical information at different positions of the sentences as different individuals processed the syntactic ambiguity cued by the prosodic patterns, resulting in less “urgent” shift to the mouth (a curve with less steep curvatures). In the intonation sentence type, participants instead fixated facial areas where visual correlates were embedded for prosodic and segmental information respectively, without the need to shift their eye gaze back and forth, again corroborating the second hypothesis. Finally, the presence of both segmental and prosodic cues in the *Both* condition suggested that perceivers showed patterns of looking in between the *Prosody* and *Segment* conditions.

6. REFERENCES

[1] J. Navarra, H. H. Yeung, J. F. Werker, and S.

Soto-Faraco, “24 Multisensory and sensorimotor interactions in speech perception,” *The Handbook of Multisensory Processing.*, pp. 435, 2012.

[2] K. W. Grant and P.-F. Seitz, “The use of visible speech cues for improving auditory detection of spoken sentences,” *J. Acoust. Soc. Am.*, vol. 108, no. 3, p. 1197, 2000.

[3] M. E. Król, “Auditory noise increases the allocation of attention to the mouth, and the eyes pay the price: An eye-tracking study,” *PLoS One*, vol. 13, no. 3, pp. 1–15, 2018.

[4] W. H. Sumby and I. Pollack, “Visual Contribution to Speech Intelligibility in Noise,” *J. Acoust. Soc. Am.*, vol. 26, no. 2, pp. 212–215, 1954.

[5] E. Vatikiotis-Bateson, I. M. Eigsti, S. Yano, and K. G. Munhall, “Eye movement of perceivers during audiovisual speech perception,” *Percept. Psychophys.*, vol. 60, no. 6, pp. 926–940, 1998.

[6] M. Dohen and H. Lœvenbruck, “Interaction of Audition and Vision for the Perception of Prosodic Contrastive Focus,” *Lang. Speech*, vol. 52, no. 2–3, pp. 177–206, 2009.

[7] M. Swerts and E. Krahmer, “Facial expression and prosodic prominence: Effects of modality and facial area,” *J. Phon.*, vol. 36, no. 2, pp. 219–238, 2008.

[8] C. R. Lansing and G. W. McConkie, “Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences,” *Percept. Psychophys.*, vol. 65, no. 4, pp. 536–552, 2003.

[9] C. R. Lansing and G. W. McConkie, “Attention to Facial Regions in Segmental and Prosodic Visual Speech Perception Tasks,” *J. Speech, Lang. Hear. Res.*, vol. 42, pp. 526–539, 1999.

[10] E. Cvejic, J. Kim, C. Davis, and G. Gibert, “Prosody for the eyes: Quantifying visual prosody using guided principal component analysis,” *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH 2010*, no. September, pp. 1433–1436, 2010.

[11] J. M. Foxton, L. D. Riviere, and P. Barone, “Cross-modal facilitation in speech prosody,” *Cognition*, vol. 115, no. 1, pp. 71–78, 2010.

[12] K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson, “Visual Prosody and Speech Intelligibility: Head Movement Improves Auditory Speech Perception,” *Psychol. Sci.*, vol. 15, no. 2, pp. 133–137, 2004.

[13] S. Garg, G. Hamarneh, A. Jongman, J. A. Sereno, and Y. Wang, “Computer-vision analysis reveals facial movements made during Mandarin tone production align with pitch trajectories,” *Speech Commun.*, vol. 113, no. April, pp. 47–62, 2019.

[14] R. E. Ronquest, S. V. Levi, and D. B. Pisoni, “Language identification from visual-only speech signals,” *Attention, Perception and Psychophysics*, vol. 72, no. 6, pp. 1601–1613, 2010.

[15] P. Prieto, C. Puglesi, J. Borràs-Comes, E. Arroyo, and J. Blat, “Exploring the contribution of prosody and gesture to the perception of focus using an animated agent,” *J. Phon.*, vol. 49, pp. 41–54, 2015.

[16] L. G. Lusk and A. D. Mitchel, “Differential Gaze Patterns on Eyes and Mouth During Audiovisual Speech Segmentation,” *Front. Psychol.*, vol. 7, no. February, pp. 1–11, 2016.

[17] D. J. Lewkowicz, M. Schmuckler, and V. Agrawal, “The multisensory cocktail party problem in adults: Perceptual segregation of talking faces on the basis of audiovisual temporal synchrony,” *Cognition*, vol. 214, no. September, pp. 104743, 2021.

[18] C. Cavé *et al.*, “About the Relationship between Eyebrow Movements and F0 Variations,” *Proceeding Fourth Int. Conf. Spok. Lang. Process.*, vol. 4, pp. 2175–2178, 1996.

[19] R. Scarborough, P. Keating, S. L. Mattys, T. Cho, and A. Alwan, “Optical Phonetics and Visual Perception of Lexical and Phrasal Stress in English,” *Lang. Speech*, vol. 52, no. 2–3, pp. 135–175, 2009.

[20] M. Dohen, H. Lœvenbruck, M. Cathiard, and J. Schwartz, “Can we see Focus ? A Visual Perception Study of Contrastive Focus in French,” *Speech Prosody 2004, International Conference*, 2004.

[21] J. W. Dink and B. Ferguson, “Eyetracking R: An R Library for Eye-Tracking Data Analysis.,” 2015. *URL http://www.eyetrackingr.com*

[22] D. Mirman, J. A. Dixon, and J. S. Magnuson, “Statistical and computational models of the visual world paradigm: Growth curves and individual differences,” *J. Mem. Lang.*, vol. 59, no. 4, pp. 475–494, 2008.

[23] S. Kalénine, D. Mirman, E. L. Middleton, and L. J. Buxbaum, “Temporal dynamics of activation of thematic and functional knowledge during conceptual processing of manipulable artifacts,” *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 38, no. 5, pp. 1274–1295, 2012.