Visual scanning of a talking face when evaluating segmental and prosodic information

Henny Yeung, Xizi Deng, Erin Jastrzebski, Elise McClay, Lydia Castro, Yue Wang
Department of Linguistics, Simon Fraser University

Speech information has a number of visual correlates, at both segmental and prosodic levels. For example, several articulatory features (mouth aperture, lip rounding, etc.) are visible in the mouth and lips (Ronquest, Levi, & Pisoni, 2010). At the prosodic level, the mouth area can also yield information about speech duration (Navarra, Alsius, Velasco, Soto-Faraco, & Spence, 2010; Navarra, Soto-Faraco, & Spence, 2014), while other cues, like pitch and amplitude, has visual correlates in other areas of the face, including the eyebrows and other head movements (Foxton, Riviere, & Barone, 2010; Garg, Hamarneh, Jongman, Sereno, & Wang, 2019; Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004). In the present study, we ask how listeners' visual scanning of a talking face is affected by task demands targeting prosodic and segmental information in a native and an unfamiliar language. There are language-based differences in how one scans a talking face: For example, adults look more at the mouth when evaluating speech information in a non-native language than a native one (Barenholtz, Mavica, & Lewkowicz, 2016; Lewkowicz & Hansen-Tift, 2012). Prior work has not examined how scanning patterns may differ when evaluating segmental *versus* prosodic information, and this work represents an initial investigation into this question.

We adapted an audiovisual speech-matching task from Barenholtz et al. (2016). Twenty-five native English speakers heard two audio sentences, and then saw a silent video of a talking face. Their task was to judge whether the video matched either the first or second audio sentence (or whether both sentences were the same). Gaze patterns were recorded as they watched the silent video, and the behavioural responses were recorded. Trials were organized into four conditions where the two auditory sentences were a) identical, or b) differed in segments, c) prosody, or d) both. For example, if the first auditory sentence was, "***BETH wants a SILVER wrist watch for HER dresser***," (caps indicates contrastive stress), then the table below shows a second sentence in each of four conditions:

Baseline (identical sentences)	BETH wants a SILVER wrist watch for HER dresser
Prosody (differs only in stress)	Beth WANTS a silver WRIST watch for her DRESSER
Segment (differs only in segments)	ROSE wants a YELLOW wrist watch for MY dresser
Both (differs in both)	Rose WANTS a yellow WRIST watch for my DRESSER

Half of trials were in English (native language), and the other half were in Mandarin (a novel non-native language), which followed the same structure as English trials. Behavioural results (Fig. 1) show that this task was harder for Mandarin stimuli, and that the prosody condition was most difficult across both languages. Gaze was further coded as falling into two interest areas: Eyes or Mouth. An Eye-Mouth Index was generated by taking the proportion of gaze to the eyes (relative to the face) minus the proportion of gaze to the mouth (relative to the face), and was calculated for the time period that the face was visible until the behavioural response was made. Figure 2 shows results from a linear mixed-effects model predicting this index from a) contrast conditions (prosody, segment, both), b) response accuracy, and c) language. Mouth looking was generally weighted towards the mouth (β = -.61, $p < .01$), but also varied as a function of behavioural accuracy: For Mandarin trials only, correct responses

predicted increased looking to the mouth (β = -.22, p < .01). Mouth looking was also more pronounced in the Segment condition relative to the Both condition (β = -.30, p < .05). Results suggest a link between mouth-looking and the extraction of speech-relevant information, but only under high cognitive load (i.e., for Mandarin stimuli, but not for English). Future work will need to examine the effects of other types of prosodic information on visual scanning.

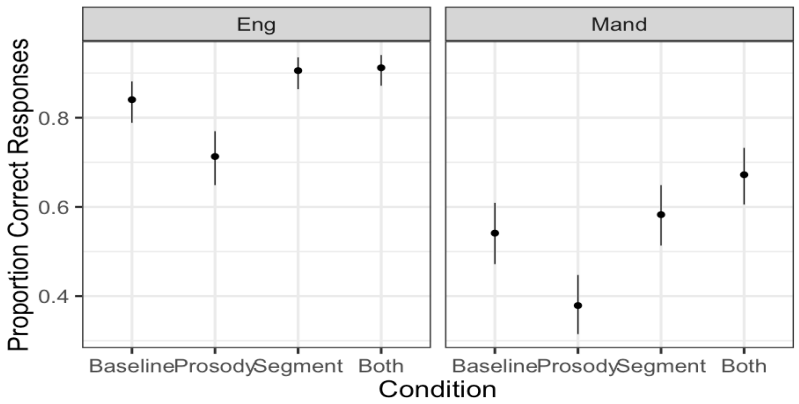
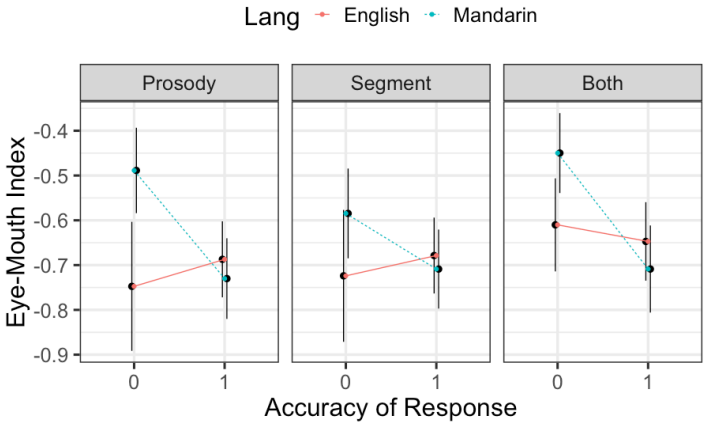


Figure 1. Behavioural results, error bars indicate 95% CIs.

Figure 2. Eye-gaze results (excluding the baseline condition), error bars are 95% CIs.



References

Barenholtz, E., Mavica, L., & Lewkowicz, D. J. (2016). Language familiarity modulates relative attention to the eyes and mouth of a talker. *Cognition*, *147*, 100–105.

Foxton, J. M., Riviere, L. D., & Barone, P. (2010). Cross-modal facilitation in speech prosody. *Cognition*, *115*(1), 71–78. https://doi.org/10.1016/j.cognition.2009.11.009

Garg, S., Hamarneh, G., Jongman, A., Sereno, J. A., & Wang, Y. (2019). Computer-vision analysis reveals facial movements made during Mandarin tone production align with pitch trajectories. *Speech Communication*, *113*(August), 47–62.

Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences*,

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual Prosody and Speech Intelligibility: Head Movement Improves Auditory Speech Perception. *Psychological Science*, *15*(2), 133–137.

Navarra, J., Alsius, A., Velasco, I., Soto-Faraco, S., & Spence, C. (2010). Perception of audiovisual speech synchrony for native and non-native language. *Brain Research*, *1323*, 84–93.

Navarra, J., Soto-Faraco, S., & Spence, C. (2014). Discriminating speech rhythms in audition, vision, and touch. *Acta Psychologica*, *151*, 197–205.

Ronquest, R. E., Levi, S. V., & Pisoni, D. B. (2010). Language identification from visual-only speech signals. *Attention, Perception, and Psychophysics*, *72*(6), 1601–1613.