Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Identifying visual prosody: where do people look?

Simone Simonetti[1], *Jeesun Kim*[1], *Chris Davis*[1]

[1] The MARCS Institute for Brain, Behaviour and Development, Western Sydney University

s.simonetti@westernsydney.edu.au, j.kim@westernsydney.edu.au,
chris.davis@westernsydney.edu.au

Abstract

Talkers produce different types of spoken prosody by varying acoustic cues (e.g., F0, duration, and amplitude), also making complementary head and face movements (visual prosody). Perceivers can categorise auditory and visual prosodic expressions at high levels of accuracy. Research using eye-tracking trained participants to recognise the visual prosody of two-word sentences and found that the upper face is more critical for determining prosody than the lower face. However, recent studies using longer sentences have shown that untrained perceivers can match lower and upper faces across modalities. Given these, we aimed to extend the eye-tracking research by examining the gaze patterns of untrained participants when judging prosody with longer utterances. Twelve participants were presented questions, narrowly focussed, or broad focussed (neutral) utterances for a 3 alternative forced-choice identification task while eye gaze was recorded. Identification accuracy was high (81-97%) and did not differ among expression types. Participants gazed at eye regions longer and more often than mouth regions for all expressions. They gazed less at the mouth region for questions than for broad and narrow focussed statements. These results are consistent with the early research indicating the importance of the upper face for determining visual prosody.

1. Introduction

When producing speech, talkers can vary acoustic properties such as F0, duration, and amplitude (prosody) to convey different information. For example, talkers typically raise F0 at the end of an utterance to indicate a question, whereas they lower it to indicate a statement [1]. Listeners usually have no difficulties in identifying the intended meaning conveyed by prosodic cues [1, 2] and can accurately discriminate between them [3].

Talkers also often move their head and face in characteristic ways (visual prosody) when they, for example, utter a question or emphasize a part of an utterance. Some movements, for instance of the mouth, have a high correlation with changes in the auditory signal. This is because there is a close link between visible speech articulation and acoustic properties. For instance, to sustain the production of a speech sound over an extended period, the talker needs to maintain the same or similar articulatory gesture [4]. Other visual cues, such as movements in the eye region, are less directly tied to the auditory signal. Nevertheless, these movements have the potential to signal changes in auditory properties. This is because such movement often occur in conjunction with acoustic properties, e.g., changes in F0 have been associated with eye squinting, eyebrow movement [5, 6], and rigid head motion [7, 8].

Although visual prosodic cues are available from both the mouth and eye regions, the results from an early study [9] suggest that people may pay more attention to cues from the eye region than to those from the mouth region. [9] used eye tracking to gauge where people looked (a proxy for where they paid attention) and asked participants to view silent videos of a talker uttering two-word phrases. After they viewed each video, they were required, at different times throughout the experiment, to perform one of three different tasks. One task, segment identification, involved deciding between two alternatives, e.g., "Ron ran" versus "We won". The two other tasks involved making a prosody judgement. One of these involved deciding whether the utterance was a question or a statement. The other one required a judgement of whether the first or second word was stressed. [9] found that people gazed more at the eye region during the question decision task than they did in the segment recognition one. The duration/number of eye gazes to the upper face did not differ between the question decision task and the stress judgment one; or between the latter and segment one. In a follow up experiment, [9] compared task performance between full face videos and videos in which the movement in the upper half of faces was frozen. They found that compared to the full face videos, performance for judging questions/statements decreased in the upper-face frozen condition; but that this manipulation did not affect performance on the stress or segment identification tasks.

Somewhat at odds with the above results, Cvejic, Kim, and Davis [3] showed that participants are able to accurately match silent videos of people uttering questions and narrow-focussed statements both within and across modalities even when presented with only the lower or the upper half of a face. These results suggest that participants can extract relevant prosody cues from the upper (eye region) as well as lower (mouth region) face areas for both focussed expressions and questions.

One reason for the different results between [9] and [3] is that these studies involved different procedures. For instance, in restricting visual cues in the eye regions, [9] froze the upper face while allowing the lower face to move, whereas [3] presented only the lower face. If, as [9] have suggested, there is a natural tendency to look at the upper face when judging whether an utterance is a question, then participants in their study may have inadvertently gazed at this region even though it was frozen. This would mean that they would have had less opportunity to extract cues from lower face regions. Of course, such a tendency would not have occurred in [3] as only the lower face was presented.

Another difference between the studies is that participants were trained in [9] but not in [3]. In [9] prior to completing the three tasks, participants were trained to perform the various judgments and were given feedback until they achieved 100% accuracy on 8 consecutive trials for segmental and stress recognition tasks and 75% accuracy on 8 consecutive trials for the question/statement task. It is possible that this training may have tended to ‘polarize” eye gaze choices to the region offering the most salient cue at the expense of other regions offering slightly less prominent information. A third different, is that [9] used two word utterances (e.g., “We won”, “Ron ran”) whereas [3] used longer sentences that included 4-5 key content words; thus in [9], gaze patterns might reflect the need

doi: 10.21437/SpeechProsody.2016-172

to obtain information as quickly as possible before the stimuli finished.

It is unclear whether some or all of these procedural differences may account for the inconsistent results. In order to address this, we conducted a modified replication of the [9] study. Our aim was to quantify perceiver's eye gaze patterns when extracting visual prosodic cues in a procedure where prosodic type was not known in advance. That is, in [9] on each block of trials, participants were first shown the alternatives they had to decide between (e.g., is the utterance a question or statement) and were then shown the silent video. In the current study, participants had a single task: they were presented with a silent video of the talker uttering a sentence and had to identify the type of prosody, i.e., an echoic question, a narrow focussed or a broad focussed statement. That is, participants were not trained and did not expect any particular prosodic expression on a given trial.

2. Method

2.1. Participants

2.1.1. Production participants

Two female native talkers of English (M_{age} = 21.5, SD = 0.7) were recruited to record the video stimuli. Both received monetary reimbursement.

2.1.2. Perception participants

Twelve students (3 males, M_{age} = 22.3, SD = 6.5) from Western Sydney University participated in this study for course credit. All learnt English at an early age and reported to have normal or corrected-to-normal vision and no hearing problems.

2.2. Stimuli

Stimuli consisted of visual only (VO) recordings of ten sentences. These sentences were produced with various prosodic expressions (broad focussed, narrow focussed, and echoic questioning) by two female talkers in their early twenties.

A large number of IEEE sentence lists [10] were rated for emotional content on a Likert scale (using a 5-point scale from -2 = very negative, 0 = neutral, and +2 = very positive). The ten sentences that received a score of 0 were selected.

The video recordings of the two talkers were captured individually in a sound attenuated booth. Each talker was seated in front of a monitor that displayed each sentence one at a time. The video camera (Sony NCCAM HXR-NX30p) was situated directly above the monitor and captured video at 1920 x 1080 full HD resolution at 50 frames per second.

Each recording session of the 10 sentences was blocked by the type of expression (broad focussed, narrow focussed, echoic questioning). The talkers were instructed as to which linguistic expression was required before each block. For broad focussed statements, talkers said aloud each sentence after first reading it silently.

A dialogue exchange task was used to elicit the focus and question expressions. In this task, the talker interacted with an interlocutor by correcting an error made by the interlocutor (a narrow corrective focussed utterance, example 1), or she questioned an item that was emphasised by the interlocutor (an echoic question, example 2).

Example 1
(a) Interlocutor: "Try to trace the **PINE**[error] lines of the painting?"
(b) Talker: "Try to trace the **FINE**[focussed] lines of the painting."

Example 2
(a) Interlocutor: "Try to trace the **FINE**[focussed] lines of the painting."
(b) Talker: "Try to trace the **FINE**[question] lines of the painting?"

The recording session was blocked by the type of expression (neutral, focussed, questioning) for the 10 sentences. Overall, each talker was recorded for a total of 30 sentences (10 sentences x 3 expression types).

Video recordings were stripped of audio, cropped to include just the head area, and segmented into each sentence using MATLAB.

2.3. Procedure

Participants were tested individually in a sound attenuated booth. They were informed that the experiment is about prosody identification from silent videos and that during the experiment their eye movements would be recorded.

Participants were told that in the experiment they would be presented with silent videos of a person uttering a sentence where an element was questioned, or was given broad (neutral) or narrow focus (examples were given). They were instructed that after watching each video, they were required to select (using the mouse) one of three response options presented on the screen ("Focus", "Question", or "Neutral"). Participants were then asked to rest their heads on a head and chin rest. They were explained that this was to limit head movements that might interfere with the recording of eye movements during the experiment.

All participants were presented VO stimuli first from one talker and then from the other (i.e., talkers were blocked). Each talker block consisted of 30 trials (3 expression conditions (focussed, question, neutral) x 10 sentences). The presentation order of the talkers was counterbalanced across participants. The presentation order of the trials within each block was randomised using the SR Research Experiment Builder display, response collection, and eye-tracking software.

Eye movements were recorded using an SR EyeLink® 1000 eye tracker. The tracker was mounted on a tower with a chin and head rest. The eye movements of the right eye were recorded for all participants.

At the beginning of each block, two neutral expressions of that talker were presented to act as a talker specific calibration. Participants were told to use these neutral expressions as a baseline to compare with subsequently presented expressions. Following this, the positions of the eye were calibrated. Three practice items were then presented, followed by the experimental trials.

To initiate the presentation of each trial, participants were required to focus on a fixation box for 1 second. The fixation box was positioned in the same location as the centre of the subsequently presented face. This was to ensure all participants had the same initial fixation point across all trials. Once the trial was triggered, participants were presented with an experimental item that lasted approximately 4 seconds. During this time, the participants' eye movements were recorded. Throughout each block, participants were given two breaks. Eye calibration was performed after each break.

Eye movement recordings were analysed using the SR Research DataViewer. Two interest areas were created for each video recording: one interest area included the eyes and the eyebrows only; and the other the mouth only (as in [11], see Figure 1). All video recordings were inspected to ensure the interest areas maintained the relevant face areas. The number of fixations and the duration of these fixations to the interest areas were analysed. The EyeLink system uses predefined algorithms to separate saccades from fixations. The system detects the velocity or acceleration of the eye. If the velocity goes above a certain threshold, a saccade is marked, otherwise, a fixation is marked.

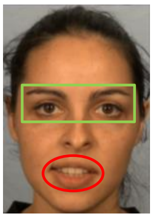

Figure 1: *Interest areas for the eye (green box) and mouth (red oval) face areas.*

3. Results

3.1. Identification task

Mean percent correct responses were collapsed across talkers and sentences within each expression type to compare identification accuracy between each of the expression types. Scores were analysed in a repeated measures ANOVA with expression type (focus, question, neutral) as a within-subjects variable. Identification scores are presented in Figure 2. Overall, there was no significant difference between the focus (91%), question (81%), or neutral (97%) expressions, $F_{(2,22)}$ = 3.18, p = .06, η_p^2 = .22.

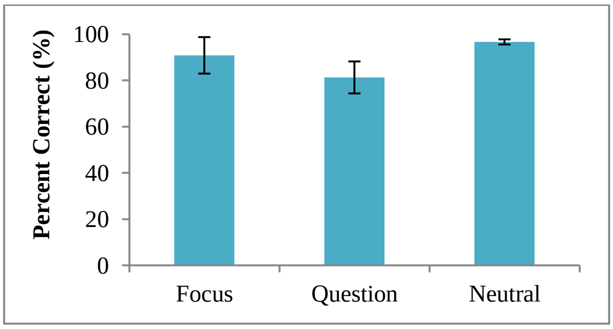

Figure 2: *Mean percent correct scores for each linguistic expression. The error bars indicate standard error.*

3.2. Eye movement analysis

The number of eye fixations and the durations (in ms) of these eye fixations for each of the interest areas (eyes, mouth) were averaged across talkers and sentences for each expression type. These scores were entered into a 3 (focus, question, neutral) by 2 (eyes, mouth) repeated measures ANOVA. Overall scores are presented in Figure 3.

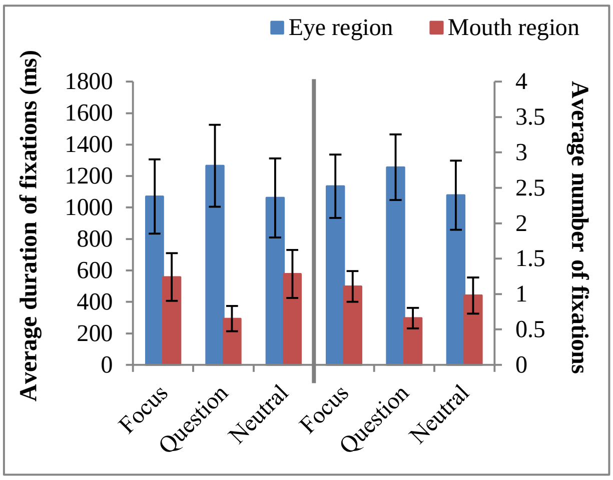

Figure 3: *Mean duration of fixations (left) and mean number of fixations (right) for the eye and mouth interest areas for each linguistic expression. The error bars indicate standard error.*

3.3. Duration of fixations

Overall, there was no significant difference in how long participants gazed at the focus (814 ms, SE = 88.8), questioning (780 ms, SE = 115.6), and neutral (819 ms, SE = 112.3) expressions, $F_{(2,22)}$ = .24, p = .79, η_p^2 = .02. Similarly, across all expression types, there was no significant difference in how long participants gazed at the eye interest area (1132 ms, SE = 246.5) compared to the mouth interest area (476 ms, SE = 122.3), $F_{(1,11)}$ = 3.86, p = .08, η_p^2 = .26. There was a significant interaction between the expression type and the area of interest, $F_{(2,22)}$ = 16.49, p < .001, η_p^2 = .60.

The above interaction (as well as all future interactions) was analysed further using a simple effect comparison with a Bonferroni adjusted alpha. For the eye region, simple comparisons reveal that the difference in scores between focussed and questioning expressions (p = .07), focussed and neutral expressions (p > .99), and questioning and neutral expressions (p = .07) did not reach significance. For the mouth region, participants showed less gazing for questions compared to focussed (p < .05) and neutral (p < .02) expressions. There was no significant difference between focussed and neutral expressions (p > .99).

Further, participants gazed longer at the eye than the mouth region for questioning expressions only (p < .01). No such differences were found for focus (p = .18) and neutral (p = .19) expressions.

3.4. Number of fixations

There were no significant differences in the number of fixations made to the focus (1.8, SE = 0.2), questioning (1.7, SE = 0.2), and neutral (1.7, SE = 0.3) faces, $F_{(2,22)} = 1.14$, p = .34, $\eta_p^2 = .09$. Participants made more fixations to the eye (2.6, SE = 0.5) compared to the mouth (0.9, SE = 0.2) area of interest, $F_{(1,11)} = 9.02$, p < .02, $\eta_p^2 = .45$. There was a significant interaction between the expression type and interest area, $F_{(2,22)} = 8.67$, p < .01, $\eta_p^2 = .44$.

Simple comparisons revealed that for the eye region, there was no significant difference between the focussed and questioning expressions (p = .14) and the focussed and neutral expressions (p > .99). More fixations were made to the eye region for questioning compared to neutral expressions (p < .05). For the mouth region, more fixations were made for focussed compared to questioning expressions (p < .01). There were no significant differences between the focussed and neutral expressions (p > .99) and the neutral and questioning expressions (p = .17).

Further, participants made more fixations to the eye than the mouth region for the focus (p < .05), question (p < .01), and neutral (p < .05) expressions.

In sum, focussed and neutral expressions were similar in the amount of gazing to the upper and also to the lower face areas; however, compared to these expressions, questioning attracted relatively more gazing at the upper face area and also less gazing at the lower face area.

4. Discussion

The current study recorded eye gaze movements (frequency and duration in eye and mouth regions) in a task in which participants were asked to identify the prosody of an utterance presented with no sound. Before each trial, participants only knew that the sentence could contain a questioned element, a narrow focussed one, or have a broad focus. It was up to them to work out which was which. Participants were able perform this task at high levels of accuracy without training (c.f., [9]). This might be because longer utterances contain more cues to prosody than shorter ones, and/or because with longer utterances participants have more time to extract relevant information.

So, how did participants perform the task? The patterns of eye gaze for the different stimulus conditions offer a clue. The results showed that, similar to [9] the pattern of eye gaze to questions was different than that for the two focussed conditions (which themselves did not differ). For questions, people looked more at the eye than the mouth region.

This pattern suggests the following: there is a reasonably clear prosodic cue that indicates that an utterance is a question and this cue is present in the eye region (e.g., squinting, brow rise, etc). This cue likely occurs at or slightly before the time of the questioned element and so may be earlier than any cue for narrow focus. Since participants do not know in advance whether the stimulus is a question, participants need to monitor the eye region for this cue (hence leading to the general pattern that participants looked more often and longer at the eye region). If, after some time has elapsed, the question cue is not perceived, then participants will concentrate on the mouth area. A possible cue for narrow focus may be relatively increased oral aperture/duration [12] in conjunction with a post focus compression. This type of cue may be clearest slightly after the focussed element has been presented. If this cue is not perceived, then the participant would respond that the utterance had a broad focus.

These results suggest that participants gaze patterns were driven by differently located face cues that contained important, task-relevant information. To examine this in more detail, a follow-up analysis could determine the temporal structure of the eye gaze trajectories. That is, given the scenario outlined above, it would be expected that gazes to the eyes would occur before those to the mouth. Furthermore, for utterances that were not questions, there should be a shift in gaze to the mouth region somewhere mid-utterance.

A further analysis could attempt to determine what the key features may have been in the different face regions. This analysis could measure overall motion in the video (e.g. using optic flow) or the motion of specific features (e.g., eyebrows). This type of analysis could determine whether there is a correlation between the amount/type of face motion and eye gaze. If eye gaze is stimulus driven (i.e., drawn to general or specific motion), then participants' eye gaze should either follow the face areas that move the most (relative to the amount of movement in the broad focussed expressions) or be tuned to the motion of particular features. Whatever is the case, these analyses only make sense with paradigms like the current one where the participant must explore and sample face regions in order to make their decisions (rather than paradigms that present participants with the type of contrast to make prior to stimulus exposure).

5. References

[1] R. J. Srinivasan and D. W. Massaro, "Perceiving prosody from the face and voice: Distinguishing statements from echoic questions in English," *Language and Speech*, vol. 46, no. 1, pp. 1-22, 2003.

[2] M. Dohen, H. Lœvenbruck, M. A. Cathiard, and J. L. Schwartz, "Visual perception of contrastive focus in reiterant French speech," *Speech Communication*, vol. 44, no. 1, pp. 155-172, 2004.

[3] E. Cvejic, J. Kim, and C. Davis, "Recognizing prosody across modalities, face areas and speakers: Examining perceivers' sensitivity to variable realizations of visual prosody," *Cognition*, vol. 122, no. 3, pp. 442-453, 2012.

[4] K. J. de Jong, "The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation," *Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 491-504, 1995.

[5] C. Cavé, I. Guaïtella, R. Bertrand, S. Santi, F. Harlay, and R. Espesser, "About the relationship between eyebrow movements and F0 variations," in *Fourth International Conference on Spoken Language Processing 1996, Proceedings*, 1996, pp. 2175-2178.

[6] B. Granström and D. House, "Audiovisual representation of prosody in expressive speech communication," *Speech Communication*, vol. 46, no. 3, pp. 473-484, 2005.

[7] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behaviour," *Speech Communication*, vol. 26, no. 1, pp. 23-43, 1998.

[8] J. Kim, E. Cvejic, and C. Davis, "Tracking eyebrows and head gestures associated with spoken prosody," *Speech Communication*, vol. 57, pp. 317-330, 2014.

[9] C. R. Lansing and G. W. McConkie, "Attention to facial regions in segmental and prosodic visual speech perception tasks," *Journal of Speech, Language, and Hearing Research*, vol. 42, no. 3, pp. 526-539, 1999.

[10] E. H. Rothauser, W. D. Chapman, N. Guttman, K. S. Nordby, H. R. Silbiger, G. E. Urbanek, and M. Weinstock, "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 227-246, 1969.

[11] S. Sullivan, T. Ruffman, and S. B. Hutton, "Age differences in emotion recognition skills and the visual scanning of emotion faces," *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, vol. 62, no. 1, pp. 53-60, 2007.

[12] R. Scarborough, P. Keating, S. L. Mattys, T. Cho, and A. Alwan, "Optical phonetics and visual perception of lexical and phrasal

stress in English," *Language and Speech*, vol. 52, no. 2-3, pp. 135-175, 2009.