Temporal relationship between auditory and visual prosodic cues

Erin Cvejic, Jeesun Kim, Chris Davis
MARCS Auditory Laboratories, University of Western Sydney, Australia
e.cvejic@uws.edu.au, j.kim@uws.edu.au, chris.davis@uws.edu.au

Abstract
It has been reported that non-articulatory visual cues to prosody tend to align with auditory cues, emphasizing auditory events that are in close alignment (visual alignment hypothesis). We investigated the temporal relationship between visual and auditory prosodic cues in a large corpus of utterances to determine the extent to which non-articulatory visual prosodic cues align with auditory ones. Six speakers saying 30 sentences in three prosodic conditions (x2 repetitions) were recorded in a dialogue exchange task, to measure how often eyebrow movements and rigid head tilts aligned with auditory prosodic cues, the temporal distribution of such movements, and the variation across prosodic conditions. The timing of brow raises and head tilts were not aligned with auditory cues, and the occurrence of visual cues was inconsistent, lending little support for the visual alignment hypothesis. Different types of visual cues may combine with auditory cues in different ways to signal prosody.

Index Terms: visual prosody, focus, phrasing, guided principal component analysis

1. Introduction
Prosody is a broad term used to describe systematic changes to the speech signal that alter the implied meaning of an utterance without modification to segmental content [1]. While prosody can be expressed in the auditory modality by modifying timing, amplitude and intonational cues, the head and face movements produced by speakers (i.e., visual speech) also provide cues that affect the interpretation of an utterance [2-6]. Visual cues to prosody appear to be distributed across the whole face including those involved in the articulatory process, as well as non-articulatory cues such as eyebrow and rigid head movements [5-8].

As prosody is perceived very well from the auditory modality alone [9], the communicative role of visual cues is unclear. Recently it has been proposed that the occurrence of a non-articulatory visual marker (eyebrow raising or a rigid tilt of the head) may align with cues in the auditory stream to indicate important information to the perceiver [10-12]. For example, in a relatively unconstrained dialogue task [10], it was found that the beginning of most eyebrow raises was effectively aligned with pitch accents (occurring approximately 60ms before). Similarly [12] examined newsreaders and found that “strong” accents (where half the raters of the auditory-only signal indicated a word as being prominent) were associated 70% of the time with eyebrow motion and 90% of the time with head movements.

The proposal that auditory-visual information is combined to mark prosody is important for models of speech production (indicating a need to take the visual modality into account) as well as for the construction of synthetic conversational agents (in order to make them more natural). So far, the studies that have examined the temporal relationship between the visual and auditory prosodic cues have typically used raters to designate auditory prominence and eyebrow/head motion from offline recordings. Furthermore, the speech data has generally come from relatively unconstrained procedures, e.g., task-oriented dialogues [10], or from newsreaders [12]. The aim of the current research was to measure eyebrow and head motion more objectively and to use an experimental procedure where the timing and production of a prosodic event can be more clearly defined. To do this, the occurrence of rigid head movements and brow raises were determined using motion capture in a relatively large corpus of speech produced in a constrained dialogue exchange task, obtaining three measures: 1) the degree of alignment of visual and auditory prosodic cues; 2) the temporal distribution of such cues, and 3) how these movements vary across prosodic conditions.

2. Method
2.1. Participants
Six male native speakers of Standard Australian English (M<sub>age</sub> = 23.2 years) participated in the data capture sessions. All participants self-reported having no known speech, hearing or communicative deficits.

2.2. Materials
The materials consisted of 30 sentences drawn from the IEEE Harvard Sentence List [13] describing mundane events with minimal emotive content. Each sentence was recorded in three prosodic conditions: as a broad focused statement, a narrow focused statement, and as an echoic question. To elicit these contrasts, a dialog exchange task involving two interlocutors was used [2, 3]. In the task, participants were required to interact with their interlocutor by either repeating what they heard the interlocutor say (broad focused statement), making a correction to an error that was made by the interlocutor (narrow focused statement, Example 1a), or questioning an emphasized item that the interlocutor produced (echoic question, Example 1b). The critical constituent within the utterance (i.e., the word that the interlocutor produced erroneously or with emphasis that was subsequently focused or questioned by the speaker) was kept consistent across all speakers and repetitions.

Example 1.

a) I. It is a band of [lead] three inches wide.
S. It is a band of [steel] three inches wide.

b) I. The brass [tube] circled the high wall.
S. The brass [tube] circled the high wall?

2.3. Apparatus
An Optotrak 3020 (Northern Digital, Inc.) was used to record the visual speech movements from 38 infrared emitting markers positioned on the head and face of the speaker (see Figure 1). These positions were chosen to reflect non-rigid movements involved in the articulation process (i.e., lip and jaw movement), as well as non-articulatory movements of the
eyebrows, and rigid rotations and translations around the centre of rotation. The three-dimensional coordinates of the markers were initially captured at 60Hz. Auditory data was synchronously captured using a Behringer C-2 condenser microphone connected to an Optotrak Data Acquisition Unit II (Northern Digital Inc.) through a Eurorack MX602A mixer, sampled at 44.1 kHz, 16-bit digitized mono.

2.4. Motion Capture Procedure

Motion capture sessions began with the placement of the movement sensors on the face of the speaker in the configuration shown in Figure 1. Each speaker was recorded individually while seated in an adjustable dentist’s chair within a double-walled, sound insulated booth. Participants engaged in the dialog exchange task outlined in Section 2.2 and were instructed to direct their speech towards the interlocutor located approximately 2.5 meters in front of them (see Figure 2). Two repetitions of each sentence were recorded in the three prosodic conditions. The total motion capture sessions lasted approximately 120 minutes (including occasional breaks). In total, 180 sentences (30 sentences x 3 prosodic contrasts x 2 tokens) were recorded for each speaker.

2.5. Data Processing

Recorded tokens were first converted to movements of a normalized head, by determining the deviation of each frame away from an individual speakers average marker configuration, and applying this movement onto the group “average” head. This data was then processed using guided principal components analysis (gPCA) [13-14] to reduce the dimensionality of the data set. Compared to “traditional” principal components analysis (PCA) which derives optimal orthogonal factors that explain the maximum amount of variance in the data using the least number of components, guided PCA uses linear decomposition to generate a set of a priori biomechanically feasible components that represent articulatory control parameters at the cost of sub-optimal explanation of variance. After estimating the centre of rotation, the rigid rotations and translations are calculated and removed, while the remaining non-rigid movements undergo gPCA (using the a priori components specified in Table 1). The recordings are then reprojected into component space, reducing the dimensionality of the data from 114 data points to 8 interpretable principal components (PCs), 3 rigid rotations and 3 rigid translations per captured frame.

Captured auditory data were subjected to semi-automatic forced phonemic alignment using the MARY text-to-speech engine [16], with the identified phoneme boundaries manually checked in Praat [17] and adjusted when necessary. The phoneme boundaries were then used to locate the critical constituent within each token. The visual movement data were interpolated from 60Hz to 100Hz using native Matlab functions to increase temporal resolution allowing for more fine-grained temporal binning in the subsequent analyses.

Table 1. A priori parameters used to guide the gPCA, percent of variance explained by each component and the extracted rigid movement parameters.

<table>
<thead>
<tr>
<th>PC</th>
<th>Movement Parameter</th>
<th>Movement Axes</th>
<th>% Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Jaw Opening</td>
<td>Y</td>
<td>38.11</td>
</tr>
<tr>
<td>2</td>
<td>Lip Opening</td>
<td>Y</td>
<td>41.52</td>
</tr>
<tr>
<td>3</td>
<td>Lower Lip Mvmt.</td>
<td>Y</td>
<td>4.03</td>
</tr>
<tr>
<td>4</td>
<td>Upper Lip Mvmt.</td>
<td>Y</td>
<td>0.81</td>
</tr>
<tr>
<td>5</td>
<td>Lip Spreading</td>
<td>X Y Z</td>
<td>4.04</td>
</tr>
<tr>
<td>6</td>
<td>Jaw Protrusion</td>
<td>Z</td>
<td>2.91</td>
</tr>
<tr>
<td>7</td>
<td>Brow Raising</td>
<td>Y</td>
<td>3.49</td>
</tr>
<tr>
<td>8</td>
<td>Brow Pinching</td>
<td>X Y</td>
<td>1.31</td>
</tr>
<tr>
<td></td>
<td>Total:</td>
<td></td>
<td>96.22</td>
</tr>
</tbody>
</table>

3. Results and Discussion

3.1. Guided Principal Component Analysis

Table 1 shows the amount of explained variance of non-rigid PC’s for each speaker. With only eight non-rigid parameters, in excess of 90% of the variance of face movement was recovered.

3.2. Brow Raises Preceding Critical Constituent

To determine how brow raises relate to an auditory prosodic event, PC7 (corresponding to brow raising in the Y-axis) was examined around the auditory onset of the critical constituent. Brow raises were chosen (as opposed to lowering) as they appear to represent the majority of brow movements [12]. The direction of brow movement at the temporal onset of the critical constituent was first determined; if the eyebrows were moving in an upward direction, the next “peak” was located, before finding the temporal onset of this upward movement (see Figure 3). Conversely, if the eyebrows were returning to a
neutral position at the time of the critical constituent, the previously occurring “peak” was temporally located, as well as the onset of that movement. A movement was considered to have occurred if the distance covered between the onset and peak of was equivalent to or greater than 3mm [7].

Of the 360 utterances recorded per prosodic condition, many showed no movements reaching the 3mm criteria. The greatest number of utterances with criteria-achieving movements occurred in the narrow focus condition (N = 213, 59.2%), followed by the echoic condition (N = 205, 57.0%) while the broad focus condition had the fewest (N = 171, 47.5%). These distributions were analysed with a one-way chi-square (with α set to .05), showing no statistically significant difference in the number of utterances accompanied by brow raises across the three prosodic conditions, $\chi^2(2, N = 589) = 5.07, p = .08$. It should be noted that although a large proportion (43%) of echoic question renditions showed no raising movements, they may have been accompanied by brow lowering (not captured in this analysis, see [12]). Indeed, echoic questions are often considered to demonstrate uncertainty, characterised by an overall smaller degree of eyebrow movements than conditions where a speaker is certain or issuing confirmation [10, 18].

Figure 4 displays the distribution of movements that occurred for each of the prosodic conditions, as a function of the time between brow movement onset and the start of the critical constituent. A two-way chi-square analyses (with α set to .025 for multiple comparisons) showed a relationship between prosodic condition and time of brow movement onset, with a differing distribution of movements occurring within each time frame across the three prosodic conditions, $\chi^2(8, N = 588) = 28.98, p < .001$. These distributions were further examined in separate one-way chi-squares (with α set to .0125) for each prosodic condition. It can be noted that the majority of movement onsets occurred more than 150ms before the start of the critical constituent for broad focus [χ²(4, N = 171) = 25.64, p < .001], narrow focus [χ²(4, N = 212) = 49.08, p < .001] and echoic question [χ²(4, N = 205) = 81.17, p < .001] conditions. Across all three prosodic conditions, the average onset time of brow raises was between 85 and 95 ms before the start of the critical word. Relative to the distribution of movements in the broad focused prosodic condition (which contains no explicit point of informational focus), both the narrow focus [χ²(4, N = 212) = 21.28, p < .001] and echoic question [χ²(4, N = 205) = 27.54, p < .001] conditions contained more criteria-achieving eyebrow movements with temporal onsets earlier than 90ms before the start of the critical word.

3.3. Rigid Head Movement and Critical Constituent

To determine how rigid head movement may relate to an auditory prosodic event, R1 was examined (i.e., pitch rotations around the X-axis, or “head nodding”, that accounted for the majority of rigid motion in [12]). Here, the displacement between the start of the critical constituent and the peak of the pitch rotation was measured (c.f. Section 2.3). A movement was only considered to have occurred if the rotation between the peak and the end of the downward rotation covered a minimum of 4° [5].

As with brow movements, only a small proportion of utterances were accompanied by pitch rotations, with the most occurring in the echoic question condition (N = 133, 36.9%), followed by the narrow focus condition (N = 122, 33.9%), with the least occurring in the broad focus condition (N = 75, 20.8%). A one-way chi-square revealed that the difference in number of utterances displaying pitch rotations between prosodic conditions was significant, $\chi^2(2, N = 330) = 17.26, p < .001$.

The distribution of pitch rotations for each of the prosodic conditions, as a function of the time between pitch rotation peak and the start of the critical constituent are displayed in Figure 5. A two-way chi-square (with α = .025 for multiple comparisons) indicated that the distribution of rigid movements peaks significantly differed as a function of the prosodic condition, $\chi^2(14, N = 329) = 68.39, p < .001$. A series of one-way chi-squares (with an adjusted α = .0125) revealed that the distributions were significantly different across the time for broad focus [$\chi^2(7, N = 75) = 51.19, p < .001$], narrow focus [$\chi^2(7, N = 121) = 166.27, p < .001$] and echoic question renditions [$\chi^2(7, N = 133) = 89.02, p < .001$]. For broad focus renditions, peaks occurred at variable times either side of the critical constituent onset, with the majority
occuring at the start of the critical word. For narrow focus renditions, pitch rotation peaks occurred most frequently more than 150ms after the start of the critical constituent, at which time the “important” part of the message has already begun, thus it is unlikely that such a movement functions to alert the perceiver, but rather may be contributing to the conveyance of suprasegmental content (i.e., reinforcing the auditory markers of focus). In echoic questions, pitch rotation peaks tended to occur after the start of the critical word, but were distributed more evenly across all time ranges. Relative to broad focused renditions, both narrow focus \( \chi^2 (7, N = 121) = 129.47, p < .001 \) and echoic questions \( \chi^2 (7, N = 133) = 82.63, p < .001 \) had a greater proportion of pitch rotation peaks occurring during or after the onset of the critical constituent within the utterance.

4. General Discussion

We examined the temporal relationship between visual prosodic cues (brow raising and rigid head tilts) and auditory speech. With regards to brow raises, the results showed that many utterances failed to show motion that reached the minimum movement criteria. This lack of motion was found even when the contrast contained an important word (i.e., narrow focus and echoic question renditions), suggesting that these movements may not be reliable cues to prosody. In instances where eyebrow raises did occur, the majority of movements began 150ms or more before the onset of the critical constituent; this finding was not consistent with that reported in [10] or with the description that such events occur “in tandem” [12] but rather suggest that such movements may function to alert perceivers to upcoming information.

For rigid head (pitch) rotations, the temporal location of the movement peak was highly variable, extending 150ms either side of the critical constituent onset. In the narrow focus and echoic question conditions, the majority of movement peaks occurred after the onset of the critical word, making it unlikely that such a cue has an alerting function. From this analysis, it may be that eyebrow and head motion cues to prosody act in different ways: with eyebrows acting to alert perceivers to, and head motion acting as confirmation of, a noteworthy event.

5. Acknowledgements

The authors thank Catherine Gasparini, Michael Fitzpatrick, Tim Paris, Guillaume Gibert and Virginie Attina for their assistance with various stages of the project, and two anonymous reviewers for their constructive feedback. Support from the Australian Research Council (TS0669874 and DP0666857) and Australasian Speech Science and Technology Association is also acknowledged.

6. References