Perceiving Visual Prosody from Point-Light Displays

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

This study examined the perception of linguistic prosody from augmented point-light displays that were derived from motion tracking six talkers producing different prosodic contrasts. In Experiment 1, we determined perceivers’ ability to use these abstract visual displays to match prosody across modalities (audio to video), when the non-matching visual display was segmentally identical and differed only in prosody. The results showed that perceivers were able to match the auditory speech to these limited face motion prosodic displays at better than chance levels; performance for the stimuli of different talkers varied greatly. A subjective perceptual rating task (Experiment 2) demonstrated that variation across talkers in the acoustic realization of prosodic contrasts may account for some of this difference; however a combination of the salience of acoustic and visual prosodic cues is likely to be driving matching performance.

Index Terms: visual prosody, focus, phrasing, point-light displays, cross-modal matching.

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]. Although prosody can be expressed in the auditory modality by modifying temporal and intonational cues, the way that an utterance is interpreted can also be affected by the talker’s visible head and face movements [2-8].

In a series of previous studies [2, 3], we demonstrated that perceivers are highly proficient in matching auditory tokens of prosody to restricted video displays (i.e., only the upper or lower halves of the face) of two talkers, when the non-matching visual token consisted of the same segmental content and differed only in prosody. In this study, we examined whether perceivers could match auditory prosody to abstract visual displays (i.e., augmented point-light), in addition to using a larger sample of talkers. Point-light stimuli have previously been used in a variety of biological motion studies, including examinations of gait, gesture and speech [9-11]. These stimuli are created by placing markers on points of interest on a subject during recording, and then processing the stimuli so that only these markers remain visible when subsequently presented to perceivers. The major benefit afforded by using such stimuli is the high degree of experimental control (which is not available when using fully-textured natural stimuli), with the ability to manipulate or control particular temporal and/or spatial properties.

In order to use point-light displays as stimuli to investigate audiovisual speech prosody, three-dimensional motion data were collected for both articulatory (e.g., lip opening, lip protrusion, jaw opening height) and non-articulatory gestures (e.g., eyebrow raises and rigid head movements), since these movements have been proposed to convey visual prosody [2-8]. As a first step, it is necessary to determine whether point-light displays are able to convey prosodic information, which has been shown to be possible with natural, non-manipulated stimuli. To do this, we used the same cross-modal prosody matching task as [2, 3] with participants asked to match auditory prosody to visual point-light displays. These results will provide the basis for future experiments that will manipulate the spatial and temporal properties of the displays (e.g., examining the effect of removing rigid head motion, or presenting only articulator movement) to understand more about how visual prosody is signaled.

2. Experiment 1

2.1. Stimuli

2.1.1. Participants

Six male native talkers of standard Australian English (M_age = 23.2 years) took part in the recording of speech materials. All were of a similar educational background, and self-reported no known vision, hearing or communicative deficits. Talkers were financially compensated for their travel expenses and time.

2.1.2. Materials

The materials consisted of 30 utterances drawn from the IEEE [12] sentence list describing mundane events with minimal emotive content. Each sentence was recorded in three prosodic conditions: a broad focus statement, a narrow focus statement, and as an echoic question, elicited in a dialogue exchange task [2, 3] with an interlocutor. Participants interacted with the interlocutor and either repeated what they heard the interlocutor say (broad focused statement), made a correction to the interlocutors erroneously produced utterance (narrow focused statement, Example 1), or questioned an emphasized constituent produced by the interlocutor (echoic question, Example 2). The critical word within the utterance (i.e., the constituent that was produced with an error or with emphasis that was subsequently focused or questioned by the talker) was kept consistent across all speakers and repetitions.

Example 1.

I: It is a band of lead three inches wide.
T: It is a band of steel three inches wide.

Example 2.

I: The pipe ran the length of the ditch.
T: The pipe ran the length of the ditch?

2.1.3. Equipment

An Optotract 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 talker (Figure 1). These positions reflected non-rigid articulatory movements (i.e., jaw and lip movement) as well as non-articulatory movements of the brows, and rigid rotations and translations of the head around the centre of rotation. The three-
dimensional coordinates of the markers were captured at 60 Hz. Auditory data was captured synchronously using a Behringer C-2 condenser microphone, connected to an Optotrack Data Acquisition Unit II (Northern Digital, Inc.) through a Eurorack MX602A mixer, sampled at 44.1 kHz, 16-bit mono.

Figure 1. Location of optical markers (with size exaggerated for clarity. 34 markers are placed directly on the face, with an additional 4 positioned on a rigid-body head band to estimate movement of the head around the centre of rotation.

2.1.4. Procedure
Movement sensors were placed on the face of the talkers as shown in Figure 1. Each talker was recorded individually while seated in an adjustable dentist’s chair within a double-walled, sound-insulated booth (Figure 2). Talkers engaged in the dialog exchange task outlined in Section 2.1.2, directing their speech towards the interlocutor located approximately 2.5 meters in front of them. Two repetitions of each sentence were recorded in each of the three prosodic conditions.

Figure 2. The experimental set-up used in the motion capture recordings.

2.2. Method

2.2.1. Materials
A subset of 10 sentences from the recorded utterances in Section 2.1 was selected for use in the perceptual experiment. These sentences were the same stimuli as used in [2] and [3], where natural stimuli were presented. The intensity of the auditory tokens was normalized to a mean relative intensity of 65 dB using Praat [13]. To convert the optical recordings to point-light displays, marker positions that were missing due to drop-out were recovered using native b-spline interpolation functions in Matlab (The MathWorks).

The 34 markers from the talkers face were represented in a three-dimensional space by solidly filled white dots on a black background. To aid participants in the interpretation of visual stimuli, the point lights were augmented with animated lips. This was achieved by first creating “phantom” marker positions below and slightly in front of the mid-superior lip marker, and above and slightly in front of the mid-inferior lip marker, before superimposing a series of color-filled triangles to join the lip markers. Eyebrows were also added by connecting the outer to mid brow, and mid brow to inner brow markers with solid lines. A “nose” was also added by joining the nose bridge marker to the nose tip marker. To elicit a three-dimensional percept, and to make movements in the z-axis (e.g., head and lip protrusions) more apparent, the point-light talkers were presented “looking” approximately 30˚ to the left. An example of the point-light talkers with visual features is shown in Figure 3 and PLTalkerExample.avi. All processing was completed in Matlab (The MathWorks).

Figure 3. Example frame of a point-light talker created for use as experimental stimuli.

2.2.2. Participants
Thirty-two undergraduate students (M_Age = 24.19 years) from the University of Western Sydney (UWS) participated for course credit. All participants self-reported normal or corrected to normal vision and hearing, and were fluent talkers of English.

2.2.3. Procedure
Participants were tested individually in a sound-attenuated booth, with visual stimuli presented on a 17” LCD computer monitor at 60 fps, and auditory stimuli presented binaurally over Senheiser HD650 stereo headphones. Stimuli were presented in a two-interval, alternate forced choice (2AFC) auditory to visual matching task. DMDX [14] was used to control stimuli presentation and response collection.

Participants were told that they would be presented two pairs of stimuli, each consisting of an auditory-only and a visual-only item, and that their task was to select the pair in which the visual display of prosody matched the auditory token (see Figure 4 and TrialExample.wmv). To avoid instance-specific strategies (i.e., reliance on absolute duration), the matching items in the correct pair were always taken from...
a different segmental content and differed only in prosody. The auditory item at the start of each pair was the same, and the standard against which the visual stimuli were to be matched. Participants indicated their response as to which pair had the same prosody via a selective button press. The order of correct response pair was counter-balanced, so it occurred equally as the first and second pair.

To avoid fatigue, and so that participants were never exposed to the same auditory stimuli more than once, two versions of the task were created, each requiring a total of 90 matching judgments, consisting of 5 sentences in each prosodic condition produced by each talker (with the tokens that participants did not hear in this Experiment presented later in Experiment 2). Participants completed only one of the versions (n = 16 completed each version). Presentation was blocked by talker, with between- and within-block randomization controlled by the presentation software. In total, the task took around 35 minutes to complete, including practice trials and several short breaks.

![Figure 4. Schematic representation of the 2AFC matching task used in Experiment 1. The same auditory item was presented first in both pairs, and was the standard against which the matching judgment was to be made. The matching item was always a different recorded token, and the non-matching item was segmentally identical but differed in prosody.](image)

### Table 1. Mean percent correct responses for the cross-modal prosody matching task, as a function of prosodic contrast, separated by talker. (df = 31), ** indicates p < .001.

<table>
<thead>
<tr>
<th>Prosodic Contrast</th>
<th>Mean Correct (%)</th>
<th>Std. Error Mean</th>
<th>t-test vs. chance (50%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Talker 1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Broad Focus</td>
<td>43.12</td>
<td>3.82</td>
<td>-1.80</td>
</tr>
<tr>
<td>Narrow Focus</td>
<td>65.62</td>
<td>3.94</td>
<td>3.97**</td>
</tr>
<tr>
<td>Echocic Question</td>
<td>68.12</td>
<td>3.79</td>
<td>4.77**</td>
</tr>
<tr>
<td><strong>Talker 2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Broad Focus</td>
<td>52.50</td>
<td>3.78</td>
<td>0.66</td>
</tr>
<tr>
<td>Narrow Focus</td>
<td>67.50</td>
<td>3.45</td>
<td>5.07**</td>
</tr>
<tr>
<td>Echocic Question</td>
<td>74.37</td>
<td>4.14</td>
<td>5.89**</td>
</tr>
<tr>
<td><strong>Talker 3</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Broad Focus</td>
<td>54.37</td>
<td>4.04</td>
<td>1.08</td>
</tr>
<tr>
<td>Narrow Focus</td>
<td>75.62</td>
<td>3.45</td>
<td>7.43**</td>
</tr>
<tr>
<td>Echocic Question</td>
<td>69.37</td>
<td>4.92</td>
<td>3.94**</td>
</tr>
<tr>
<td><strong>Talker 4</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Broad Focus</td>
<td>58.75</td>
<td>4.75</td>
<td>1.84</td>
</tr>
<tr>
<td>Narrow Focus</td>
<td>67.50</td>
<td>3.78</td>
<td>4.62**</td>
</tr>
<tr>
<td>Echocic Question</td>
<td>52.50</td>
<td>4.38</td>
<td>0.57</td>
</tr>
<tr>
<td><strong>Talker 5</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Broad Focus</td>
<td>71.25</td>
<td>4.21</td>
<td>5.05**</td>
</tr>
<tr>
<td>Narrow Focus</td>
<td>90.62</td>
<td>3.24</td>
<td>12.55**</td>
</tr>
<tr>
<td>Echocic Question</td>
<td>86.87</td>
<td>3.77</td>
<td>9.79**</td>
</tr>
<tr>
<td><strong>Talker 6</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Broad Focus</td>
<td>53.75</td>
<td>4.44</td>
<td>0.85</td>
</tr>
<tr>
<td>Narrow Focus</td>
<td>78.12</td>
<td>4.05</td>
<td>6.94**</td>
</tr>
<tr>
<td>Echocic Question</td>
<td>83.12</td>
<td>3.12</td>
<td>10.60**</td>
</tr>
</tbody>
</table>

that many of the conditions resulted in matching performance better than that expected from chance alone (i.e., 50%), particularly for narrow focus matches.

Interestingly, cross-modal matching of broad focused contrasts was performed (with the exception of Talker 5) at chance level. As broad focused auditory stimuli contained no explicit point of informational focus, it may have been difficult for perceivers to resolve the prosodic type, making them less sensitive to corresponding cues in the visual signal [see 15].

A 3×6 repeated-measures ANOVA was conducted on task performance, with prosodic contrast (broad focus statement, narrow focus statement, echoic question) and talker as between-subjects measures. A significant main effect of prosodic condition was found, F(2,62) = 30.40, p < .001, η²_p = .50, with Sidak post-hoc comparisons showing that, collapsed across talkers, auditory to visual matches for both narrow focus (M_Diff= 18.54, Sidak 95% CI: 11.60 – 25.49) and echoic questions (M_Diff= 16.77, Sidak 95% CI: 9.51 – 24.03) were performed with significantly greater accuracy than trials involving matching broad focused tokens across modalities. A significant main effect was also found for talker, F(5,155) = 16.58, p < .001, η²_p = .35.

The prosodic contrast by talker interaction was also significant, F(10,310) = 3.02, p = .001, η²_p = .09. To interpret the interaction, a series of post-hoc repeated measures ANOVA’s (with an adjusted α of .0125 for multiple comparisons) were conducted between talkers for each of the prosodic contrast types. For broad focused trials, the effect of talker was significant, F(5,155) = 4.76, p < .001, η²_p = .13. Similarly, a significant difference between talkers was found for narrow focus judgments, F(5,155) = 7.09, p < .001, η²_p = .19, and for echoic question judgments, F(5,155) = 12.21, p < .001, η²_p = .28.

Figure 4. Schematic representation of the 2AFC matching task used in Experiment 1. The same auditory item was presented first in both pairs, and was the standard against which the matching judgment was to be made. The matching item was always a different recorded token, and the non-matching item was segmentally identical but differed in prosody.

### 2.3. Results and Discussion

A series of independent samples t-tests were first conducted between matching performance for each prosodic contrast across the two versions of the task. No difference in performance accuracy was found for any of the prosodic contrasts, and thus, the following analyses are collapsed across the two task versions. The percent of correct responses for each prosodic contrast type separated by talker are shown in Table 1, collapsed across sentences. A series of t-tests showed
From these results, it is evident that prosody can be matched across modalities even when the visual signal lacks textural information and displays only minimal movement information. Of course, performance levels were somewhat lower than what was found using the same task with restricted video displays of either the upper [2] or lower face [3] (where matching accuracy exceeded 80% across all three prosodic conditions). Furthermore, the ability to cross-modally match auditory to visual point-light tokens appears to largely depend on the talker producing the stimuli; for example, Talker 5 was much easier to match cross-modally than the others.

This difference across talkers suggests several possibilities: 1) for some talkers there is a discrepancy between the auditory realization of prosody and the visual correlates that are produced, making matching across modalities more difficult for these talkers, 2) there is variation in the salience of prosodic cues produced by different talkers. It is known that the production of the acoustic features used to signal a prosodic contrast can vary [16]; with cues produced interchangeably (e.g., durational lengthening alone, or in combination with an increase in amplitude) to achieve the same type of contrast [17]. Since perceivers are initially presented with an auditory token and are asked to match this to its corresponding video, any difficulty in interpreting the prosodic content of the auditory item would make selecting the matching visual display substantially more difficult. To examine this, Experiment 2 employed a perceptual rating task (of the auditory signal) to measure the salience or perceptual strength of prosodic contrasts produced by each of the talkers. These scores can then be compared with matching accuracy in Experiment 1.

3. Experiment 2

Previous studies [18, 19] have suggested that despite acoustic variation in the realization of linguistic prosody (focus and phrasing), such contrasts can be perceptually equivalent. These studies, however, used categorical judgment tasks that required listeners to either locate the narrowly focused constituent, or categorize an utterance as a statement or question. For the current purpose, this method may be too coarse, since tokens that are somewhat ambiguous, or poorer exemplars of a particular prosodic category may still be deemed as belonging to the same category and will be given the same perceptual weighting as those tokens that are good exemplars. This makes it difficult to determine whether variable realizations convey the contrastive prosodic information with the same perceptual “strength”. To overcome this, in Experiment 2, we utilized a 7-point Likert scale and asked participants to directly evaluate the strength of the auditory prosody of each stimulus.

3.1. Method

3.1.1. Materials

The same auditory items presented in Experiment 1 were used in the perceptual rating task. However, as the broad focused statement renditions acted as the comparison for both the focus (broad vs. narrow focus) and phrasing tasks (declarative statement vs. echoic question), half of the broad focused trials were the second recorded auditory token.

3.1.2. Participants and Procedure

The same 32 participants took part as in Experiment 1. Participants completed two perceptual rating tasks in a counter-balanced order; in one task, participants rated the talkers’ production of prosodic focus, in the other they rated utterance focus. Both tasks were presented using DMDX [14], with auditory stimuli presented binaurally over Senheiser HD650 stereo headphones. Participants completed these tasks immediately after finishing Experiment 1.

For the focus rating task, participants were initially presented with the critical word printed on screen, followed by an auditory token of an utterance, and were then asked to rate how well the critical word was focused within the auditory token using a 7-point Likert scale, with a response of “1” indicating that the word received no focus, whereas a rating of “7” indicated that the word was clearly focused. In total, 60 stimulus items were presented, comprising of a single repetition of 5 sentences produced as a broad focused rendition, and 5 narrow focus renditions, from each talker. As with the first experiment, there were two versions of the task. The auditory items presented to participants were the tokens that they had not yet been exposed to in the first experiment. Presentation of items was blocked by speaker, with presentation order between- and within-blocks randomized by the presentation software.

The phrasing rating task was similar to the focus rating one, with participants using a 7-point Likert scale to evaluate the phrasal nature of a presented utterance. Participants were presented auditory tokens and asked to rate how well the utterance was produced on a continuum of “statement” to “question”, with a response of “1” indicating that the utterance was definitely phrased as a statement, while a rating of “7” indicated that the utterance was definitely phrased as a question. A total of 60 items were presented comprising of a single repetition of 5 sentences produced as a broad focused statement rendition, and 5 as an echoic question rendition from each talker, with item presentation blocked by speaker. Two versions of the task were made, so that participants never heard the same auditory token more than once across both tasks. The broad focus token used was always a different repetition from that which appeared in the focus rating task, so participants were never exposed to the same token more than once. For both tasks, participants were informed that there was no “correct” answer, and were encouraged to use the complete range of the rating scale responses.

3.2. Results and Discussion

Table 2 shows the mean ratings of each talker’s production of focus and phrasing collapsed across sentences. These ratings were subjected to mixed repeated measures ANOVAs for each perceptual task, with prosodic condition and talker as between-items measures and perceiver ratings as the within-items repeated measure.

<table>
<thead>
<tr>
<th>Talker</th>
<th>Focus Judgment Task</th>
<th>Phrasing Judgment Task</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Broad Focus</td>
<td>Narrow Focus</td>
</tr>
<tr>
<td>1</td>
<td>2.12</td>
<td>4.51</td>
</tr>
<tr>
<td>2</td>
<td>2.19</td>
<td>5.65</td>
</tr>
<tr>
<td>3</td>
<td>1.67</td>
<td>4.75</td>
</tr>
<tr>
<td>4</td>
<td>1.81</td>
<td>6.55</td>
</tr>
<tr>
<td>5</td>
<td>1.87</td>
<td>6.24</td>
</tr>
<tr>
<td>6</td>
<td>2.00</td>
<td>6.19</td>
</tr>
</tbody>
</table>
For the focus rating task, significant main effects were found for prosodic condition, $F(1,108) = 1365.98$, $p < .001$, $\eta^2_p = .93$, as well as talker, $F(5,108) = 11.88$, $p < .001$, $\eta^2_p = .36$. The talker by prosodic condition interaction was also significant, $F(5,108) = 13.11$, $p < .001$, $\eta^2_p = .38$. Sidak post hoc comparisons showed, as expected, that broad focused renditions were rated significantly lower than narrow focus renditions ($M_{diff} = 3.70$, Sidak 95% CI: 3.50 – 3.90). To determine where the differences between talkers occurred, a series of post hoc repeated measures ANOVAs were conducted separately for each prosodic condition, with talker as the between-items factor. No difference between speakers was observed for ratings of broad focused renditions, $F(5,54) = 2.05$, $p = .087$, $\eta^2_p = .16$, however significant differences occurred for ratings of narrow focus, $F(5,54) = 17.14$, $p < .001$, $\eta^2_p = .61$.

The ANOVA of the phrasing judgment task yielded similar outcomes; the main effects of prosodic condition, $F(1,105) = 1541.72$, $p < .001$, $\eta^2_p = .94$, and talker, $F(5,105) = 8.29$, $p < .001$, $\eta^2_p = .28$, were both statistically significant, as was the interaction, $F(5,105) = 10.82$, $p < .001$, $\eta^2_p = .34$. As expected, utterances phrased as statements were rated significantly lower (i.e., more statement-like) than echoic question productions ($M_{diff} = 3.92$, Sidak 95% CI: 3.72 – 4.11). Post hoc repeated measure ANOVAs revealed significant differences in ratings between talkers for realizations of echoic questions, $F(5,54) = 10.47$, $p < .001$, $\eta^2_p = .49$, but not for ratings of statements, $F(5,51) = 1.94$, $p = .10$, $\eta^2_p = .16$. Pearson product-moment correlations were conducted to determine if there was any relationship between the subjective ratings and item accuracy in Experiment 1, with items rated as having stronger focus/being more question-like being matched to the prosodically corresponding visual signal with higher accuracy. The small size of the correlations suggests that it is more likely a combination of the salience of the prosodic contrast in the auditory signal, in conjunction with the clarity of a range of different visual cues (e.g., rigid motion and brow movements) that drove performance in Experiment 1.

As ceiling effects were not apparent for the majority of talkers, these point-light stimuli provide suitable material for manipulation of the temporality and amplitude of various parameters. For example, by using a guided principal components analysis [20-22], we can characterize in excess of 95% of movement variance with eight non-rigid and six rigid movement parameters, reflecting both articulatory control and non-articulatory gestures. These parameters can then be controlled, enhanced or removed to examine the differential effect that each has on matching performance.

5. Acknowledgements

The authors thank Michael Fitzpatrick and Tim Paris for their assistance in collecting the motion capture data, and all the talkers for their time and patience during recordings. The authors also wish to thank three anonymous reviewers for their helpful feedback on the manuscript. The first author acknowledges financial support from the Australasian Speech Science and Technology Association, and the second and third authors acknowledge support from the Australian Research Council (TS0669874 and DP0666857).

6. References


