Off the top of the head: Audio-Visual Speech Perception from the nose up

Chris Davis and Jeesun Kim

Department of Psychology
The University of Melbourne, Australia
cwd@unimelb.edu.au; jeesun@unimelb.edu.au

Abstract

When someone speaks, linguistically relevant movements are produced. Most of the previous work that has examined the visual manifestation of speech has investigated regions around the oral aperture and sometimes the jaw. In the current work we examined whether people could use visual information from the top part of the talker’s face to determine which of two sentences had been uttered. In the first experiment observers saw two pairs of short video clips that showed the top part of a talker’s head as the person uttered a sentence. The audio component of the video was not played. The task was to judge the pair of videos in which the talker said the same sentence (different tokens were used). Observers were able to make this judgment at better than chance levels (on average 71% correct). In the second experiment, the same participants were presented with auditory-visual pairs and had to select in which pair the talker said the same sentence. Once again matching performance was above chance (69% correct). Taken together, the results indicate that participants are sensitive to the speech relatedness of movements from the upper part of the head and face. Implications for audio-visual speech processing as well as for the development of virtual characters are considered.

1. Introduction

Imagine two people talking to each other face to face. In such a conversation information is transmitted using many different modes. Most prominent is the exchange of spoken messages encoded by propositional symbols (i.e., words). However, accompanying the words is an extensive array of other information (facial gestures, head and body movements, etc). These forms of communication provide a framework on which day to day conversation is often supported. Many experiments have demonstrated that seeing the talker as they articulate enhances the recovery of the auditory signal when it is degraded by noise. These experiments typically concentrate on describing the visual information provided by the movements of the talker’s lips (indeed, sometimes only the lip movements are shown, see [1]). More recently, interest has moved to consider speech relevant information presented at peri-oral regions. For example, Lansing and McConkie [2] monitored the eye movements of participants when they performed in two different tasks. In one task, they were to determine if a talker (shown without sound) had either spoken a statement or asked a question (e.g., “we won” or “we won?”). In the other task, participants were asked to make a judgment about what the talker said (“we won” or “Ron won”). It was found that in the statement/question task, participants directed their gaze more to the upper part of the talker’s face compared to when making the judgment about what was said. It was also found that performance on the question statement task was poor if movements from only the lower part of the face were shown. This result, showing that people often do not direct their gaze at the mouth region of talkers, is consistent with other eye-movement studies [3] and is compatible with the idea that useful information can be extracted from the head and facial movements that accompany speech [4]. In a recent study, Munhall and colleagues [5] have confirmed that aspects of head motion and fundamental frequency are correlated [6] and also showed that participants identified more syllables of speech in noise when they could see the natural head movements of the talker compared to when these were eliminated or distorted. To further explore the extent of information available in non-mouth and jaw head and face movements the current experiment showed only movements in the upper part of a talker’s head and face and tested whether people can use these to make reliable judgments about the nature of sentences that generated these movements.

2. Experiment 1

The first experiment investigated whether observers could use the head and face movements in the upper part of a talker’s face to determine when a talker had said the same sentence (even though different stimulus tokens were used). Experiment 2 examined cross-modal matching performance and was run in conjunction with the first experiment (the order of running the experiments was counter-balanced with a gap of a few minutes between each).

3. Method

3.1. Participants

Twelve undergraduate students participated in the experiment. All were native speakers of English and had normal or corrected to normal vision and no report of hearing loss.

3.2. Materials

Sixty-four video recordings of a single male talker were selected from a pre-recorded audio-visual database. The videos were recorded in a well lit sound attenuated room with a Sony TRV 900E digital camera, video at 25 fps and audio at 48000 HZ, 16-bit stereo. In half of the videos the talker said the same sentence. These recording were made on the same occasion but were taken several minutes apart. Half of the stimulus sentences were phonetically balanced sentences drawn from the Harvard sentence list [7]. These sentences did not require affective renditions as they described mundane events (e.g., “The jacket hung on the back of the wide chair”). We will call this the “non-affect set”. The other half of the stimulus sentences consisted of short phrases that were spoken to express a particular emotion (e.g., “That is really annoying;
I have to let you know”). We will call this the “affect set”. The video component of the recording was cropped (directly under the eyes) so that only the top part of the talker’s head and face was shown (see Figure 1).

Figure 1: The amount of the face that was shown.

In the first experiment the audio portion of the video was not played. Pairs of videos were created by using Adobe Premier so that 500 ms of black frames separated each clip. The video clips were sorted into two groups depending upon the duration of the spoken sentence (the maximum difference in duration between clips within a group was 40 ms). Six different versions of the experimental list were prepared in order that every item within a group would appear with every other item as different pairs. This was done so that there would be no systematic bias in the aspect of the different video pairs.

3.3. Procedure

Participants were first randomly allocated to one of the six versions of the experiment. Each participant was tested individually in a sound attenuated booth. Stimulus presentation and response collection was controlled by computer (P311 800 MHz) using the DMDX software program [8] that can display synchronized audio and video sequences. Stimuli were presented on 15” monitor. The experiment used a 2AFC procedure (see Figure 2) in which participants were asked to decide in which pair of video clips did the talker say the same sentence (it was explained that the “same sentence” videos were not identical as different tokens were used).

Responses were made by pressing one of two numbered buttons (“1” for the first pair, “2” for the second). Each trial began with the written word “PAIR 1” that was presented for 800 ms. Following this, a pair of silent videos was presented (500 ms of black was presented between each of the videos of the pair). This was followed by the word “PAIR 2” displayed for 800 ms. Following this, another pair of silent videos was presented (again with 500 ms of black presented between each of the videos of the pair). After this, the written word “respond” appeared. It was explained to participants that the first video of the second pair was the same as the very first video clip and that this clip was the standard against which the second and fourth clips were to be judged. Half the trials (affect and non-affect each) began with the pair of videos showing the talker saying the same sentence and the other half with the talker saying the different sentences (the presentation of these was mixed and randomly ordered).

4. Results

Mean percent correct 2AFC performance as a function of the stimulus type (non-affect versus affect set) is shown in Table 1. As can be seen, the level of performance was above what would be expected by chance (50%). This indicates that the top part of the head and face provides useful visual information concerning what the talker is saying.

Table 1: Mean percent correct visual-visual matching performance (and standard error) for the non-affect and affect videos

<table>
<thead>
<tr>
<th>Stimulus Type</th>
<th>Correct %</th>
<th>Std Err</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-affect</td>
<td>60</td>
<td>3.1</td>
</tr>
<tr>
<td>Affect</td>
<td>82</td>
<td>3.7</td>
</tr>
</tbody>
</table>

It is apparent from the results that the affective stimuli were easier to match than the non-affective ones. The result of a one-sample t-test determined that the level of performance for the non-affect stimuli differed from chance (50%), t (11) = 3.0, p < 0.05. A paired t-test showed the performance on the affective stimuli was superior to that for the non-affective ones, t (11) = 5.6, p < 0.05.

5. Experiment 2

Having demonstrated that the top part of the face contained sufficient visual information to determine that a talker said the same sentence, Experiment 2 examined whether participants could reliably perform a cross-modal match between an auditory presentation of a sentence and the movements generated in the top part of the head and face when speaking that sentence.
6. Method

6.1. Participants
The same observers that took part in Experiment 1 participated in this experiment.

6.2. Materials
The same materials as used in Experiment 1 were employed in this experiment. The only difference was that instead of seeing the first video clip of the pair, participants heard the audio component of the clip (the video was not shown).

6.3. Procedure
As in Experiment 1, participants were first randomly allocated to one of the six versions of the experiment. The same 2AFC task and response procedure used in Experiment 1 was employed. In this experiment, the participants were asked to decide in which pair were the voice and the head movements from the same sentence (it was explained that the “same sentence” audio and videos were not identical as different tokens were used). Each trial began with the written word “ready” that was presented for 800 ms. Following this, a spoken sentence was presented via headphones (Sennheiser HD414) at 60 dBA. Following this a silent video was presented. This was followed by the word “PAIR 2” displayed for 800 ms after which the same spoken sentence as presented initially was presented and then followed by another silent video clip. After this, the written word “respond” appeared. The rest of the procedure was the same as Experiment 1.

7. Results
Mean percent correct 2AFC performance for the two types of stimuli (non-affect and affect) is shown in Table 2. A single sample t-test determined that performance on the non-affect stimuli (64% correct) was superior to chance (50%), t(11) = 4.2, p < 0.05. As with the visual-visual match in Experiment 1, it is clear from Table 2 that more correct audio-visual matches were identified with the affect stimuli. A paired sample t-test showed that this difference was significant, t(11) = 2.3, p = 0.045. It is interesting to note that although the pattern of correct responses across the two experiments was similar, there was variation in the size of the difference in correct performance for the two types of stimuli (non-affect and affect) across the visual-visual match (22% difference) and the audio-visual match (11% difference). This difference appeared to be due to participants being slightly more accurate with the non-affect stimuli on the audio-visual matching task (compared to visual-visual matching) and slightly worse on the affect ones. A repeated measure ANOVA determined that the interaction between task and stimulus type was significant, F(1,11) = 5.8, p < 0.05.

Table 2. Mean percent correct audio-visual performance (and standard error) for the non-affect and affect videos

<table>
<thead>
<tr>
<th>Stimulus Type</th>
<th>Correct %</th>
<th>Std Err</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-affect</td>
<td>64</td>
<td>2.9</td>
</tr>
<tr>
<td>Affect</td>
<td>75</td>
<td>3.8</td>
</tr>
</tbody>
</table>

8. Discussion
The results showed that people can reliably use the speech movement information produced in the upper part of the head and face to judge if the same sentence had been said. That this was possible even though different tokens were used suggests that matching performance was not simply based upon some one-off idiosyncratic movement but may have involved the detection of some more general pattern. This above-chance matching performance occurred for both the visual-visual matching and the audio-visual matching tasks. This finding is consistent with others that have shown that head movements are correlated with changes in pitch and amplitude [4, 5] and in a sense provide a visual analogue of prosody. The results also indicated that there were more correct matches with the affect stimuli. This was particularly the case in the visual-visual matching task and was most likely due to particular facial movements associated with emotional display. This interpretation is supported by the reduction in the difference in performance for the affect and non-affect stimuli in the audio-visual matching task.

The findings are also consistent with experiments that have shown a functional correlation between eye-brow movement and variation in the fundamental frequency of an utterance [9]. In order to determine the degree to which non-rigid movements (such as eye-brows, frowns etc) play a role in the current results we filtered the video component of the clips so that only the outline of the head was apparent (as well as the eyes to act as an anchor point). Preliminary data suggest that people’s matching performance is still above chance and suggest that head movement on its own provides an important source of speech related information.
Having shown that speech related information signals by upper head and face movements can be used when a task explicitly directs people to do so, we are now in the process of conducting studies to determine if people will use this information implicitly. The experiment involves speech identification in noise and is broadly similar to that conducted by Munhall and colleagues [5] except that only the top part of the head is shown (both in full texture and in outline only, see above). Given previous findings, the expectation is that seeing the talker will enhance the intelligibility of the noisy auditory speech even though the mouth and jaw regions are not shown. Finally, the current results showing that people are sensitive to the (non-mouth and jaw) movements of the head and face that accompany speech production suggest that the overall naturalness of virtual talking heads will be increased by the provision of realistic head dynamics.

9. Acknowledgements

The second author wishes to acknowledge support from Australia Research Council Grant number DP0209664.

10. References


