The effect of visual speech timing and form cues on the processing of speech and nonspeech

Chris Davis, Jeesun Kim

The MARCS Institute, University of Western Sydney, Australia

chris.davis@uws.edu.au, j.kim@uws.edu.au

Abstract

Auditory speech processing is facilitated when the talker's face and head movements (visual speech) can be seen. This effect occurs over a range of spoken word tasks, e.g., for spoken word identification (determining which word was presented) and for speech detection (determining whether speech was presented). This study examined the effect of providing two types of visual cue on the speed of determining whether a speech or non-speech sound was presented. Speech stimuli consisted of spoken nonwords and non-speech stimuli which were the spectrally inverted-versions of these. These stimuli were presented paired with either the talker's static or moving face. Two types of moving face stimuli were used: full-face versions where both spoken form and timing cues were available and modified face versions where only the timing cues provided by peri-oral motion were available (i.e., the mouth area was obscured). The results showed that the peri-oral timing cues facilitated response time for both speech and non-speech stimuli (compared to the static face condition). An additional facilitatory effect was found for the full-face versions (compared to the peri-oral timing cue condition) but this effect only occurred for the speech stimuli. The different roles these cues play in speech processing are discussed.

Index Terms: visual speech, auditory and visual speech processing, visual form and timing cues

1. Introduction

It is well established that seeing the talker's visual speech influences the process of speech perception. The effect of visual speech has been attributed to the information available from the oral regions of the talker's face (e.g., from mouth shapes, mouth and lip motion and some tongue positions) [1], as well information from the motion of peri-oral regions such as jaw, eyebrows and head [2; 3].

The current study focused on the effect that perceiving articulatory movements has on speech processing, and was motivated by the observation that such motion provides two types of cue. Due to the way that speech is produced, articulatory movements typically precede the production of sound. These movements potentially provide form cues as to what the upcoming speech might be and timing cues about when speech will be heard. In most studies investigating the effects of visual speech, both cues to speech form and timing are presented and so the extent to which each of these make a contribution to speech processing has rarely been examined (c.f., [4]).

One problem with attempting to determining the relative contribution of form and timing cues to auditory-visual (AV) speech processing is that the task used to measure speech processing likely has an impact on what is found. That is, tasks that present relatively intact speech and require participants to identify words or segments tend to be sensitive to form whereas tasks in which the speech signal is degraded and where participants are required to identify when (or if) a stimulus has occurred tend to show that timing is important.

For example, the McGurk effect [5] in which an auditory /ba/ (presented in conjunction with a visual "ga") is perceived as /da/ shows that the visual speech form can have a dramatic effect on speech perception. Furthermore, the McGurk effect is relatively tolerant to large asynchronies between the AV speech signals, particularly when auditory speech lags (e.g., [6]). This finding suggests that AV integration based on form cues may operate largely independently of timing cues.

The relative insensitivity to AV timing differences shown in the McGurk effect may not be too surprising given that the effect is indexed by a speech identification measure. There are two reasons why identification responses might miss AV timing effects. First, identification involves a relatively late stage of speech processing and if AV timing cues affect an early one such an influence may get washed out by the speech identification stage. Second, identification responses are typically non-speeded, so it is possible that an effect of the timing of AV cues may have dissipated by the time an identification response has been formulated. If this were the case, then AV timing effects may be more apparent with a response time measure.

The results of a recent study [4] however, suggest that using a response time measure in itself does not reveal putative AV timing effects. In this study, visual speech up to the beginning of the auditory signal was presented and the time taken by participants to discriminate whether heard speech was a /ba/ or /da/ measured. The influence of speech form and timing cues on response times was examined by presenting form cues (the full face of the talker) that did not provide reliable timing information (i.e., the auditory stimulus was presented at random between 250 and 400 ms after the visual one) or timing cues (only the talker's jaw movements) that did not provide form information about the spoken syllable. Compared to an auditory alone control, it was found that the form cue significant facilitated response times but the timing cue did not.

Where AV timing does appear to have a clear influence is in speech in noise detection task [7, 8]. In this task auditory speech is heavily masked by white noise and the participants task is to judge whether a noise only token or a speech and noise token was presented. This judgment was facilitated by simultaneously presenting the talker's visual speech. Kim and Davis [8] showed that this visual effect was affected the synchrony of the AV signals such that no influence of visual speech was found with even a relatively small auditory lag (40 ms). This result suggests that the detection task may be more sensitive to the timing of the AV signals than to form cues. Indeed, a problem with the speech detection task is that in order for the task to produce sensible (non-ceiling) data, the

auditory signal has to be so extensively masked that the extent to which an intact speech signal is actually needed is unclear (see [7]).

To address the above concerns, the current study used a task that had elements of an identification task (participants were asked to discriminate speech from non-speech that had speech-like characteristics) and a detection task (there was uncertainty as when the signal would be presented). Several factors were considered in designing this procedure. First, the task should potentially be sensitive to the influence of visual speech at an early stage of speech processing. In this regard, a task was selected that required a simple binary speeded response based on a basic analysis of the signal (but which required speech features to be determined). In this regard speech consisting of nonwords should be used to minimize the influence of lexical processing. Second, we wanted a task in which both speech and non-speech stimuli could be included since it has been proposed that early visual speech effects may not be specific to speech processing [9, 10]. Further, a method of isolating timing and form cues was required. Here we followed [4] and as a cue to speech timing presented a talker's jaw movements with the mouth area obscured by an overlaid opaque circle (see Figure 1). Note although the mouth area itself was obscured, face movements around mouth (including jaw movements) can still be visible, providing information about the temporal structure of the speech movements; however, it is hard to speech read, that is, detailed form cues are determined by mouth and lip movements (including tongue & teeth). To examine the effect of form cues, we included another visual speech condition that includes both timing and form cues, i.e., a full view of the talker's articulatory movements. So the effect of form cues was tested by examining whether seeing additional mouth movements will produce any extra effect on speech processing compared to when seeing only temporal cues.

In summary, the task consisted of the following: on each trial of the experiment, a spoken nonsense word or a non-speech sound was presented in a moderate level of noise (-5 dB) paired with either a static picture of the talker's face or a movie of the talker's face. Participants were required to respond they detect speech or non-speech signals by pressing a button of the right for speech and the left one for non-speech. So, the issues to be determined are two-fold. One is whether visual form/timing cues each would show an effect in a task that clearly taps early speech processing; the other whether these effects would differ depending on whether the cues were paired with speech or non-speech.

2. Method

2.1. Participants

Nineteen undergraduate students of University of Western Sydney participated in the experiment for course credit. All participants were native speakers of English. None of the participants reported any hearing loss and all had normal or corrected-to-normal vision.

2.2. Materials

Target items consisted of 45 speech (nonwords) and 45 non-speech stimuli (the latter were constructed from the speech stimuli, see below). Nonwords were selected from the ARC Nonword Database [11].

Auditory and visual speech stimuli of the speech items were obtained from video recording a male native Australian English talker using a Sony TRV 900E digital camera at 25 fps and 48,000 HZ audio. The male talker was positioned 1.5 m from the camera and recorded against a blank background. Illumination and the talker's distance from camera were held constant across items.

Auditory stimuli: The auditory portion of each video was processed separately so that across-token amplitudes could be normalized and a white noise masker added at a specified SNR. The average duration of the speech segments was 651 ms (SD = 140 ms) and the mean amplitude of each was normalized to an average of 56 dB SPL (SD = 4.3 dB). These speech portions occurred in the context of an initial silent period and a following one (see below). White noise was added to the entire auditory track to achieve an SNR of -5dB with respect to the speech portion.

The non-speech stimuli were constructed from the nonword speech tokens (before the white noise was added) by rotating the spectral energy of each around the midpoint of the sample's frequency range. This transformation maintained many of the characteristics of speech (e.g., similar spectral and temporal complexity) and has phonetic features (e.g., voice and voiceless sounds can be distinguished and voiceless fricatives can be identified (see [12]).

Visual stimuli: Only the lower region of the face (from the bottom of the eyes down) was presented as in [13, 14]. The video files subtended a height of 12.1 degrees of visual arc and a width of 15.7 degrees. The files were 74 frames in length and played on a black background at a screen resolution of 640 x 480 with 32-bit in grayscale at 25 frames/second. The videos began with silence and the mean onset of the auditory signal was 1764 ms (SD = 59 ms). Three types of visual condition were constructed for the Speech and Non-speech stimuli (See Figure 1): 1. Form & Timing cues (the full video, i.e., face and mouth visible). 2. Timing Cues (the moving face with the oral region obscured by superimposing a gray circle (radius 2 degrees of visual arc) to cover the mouth movements). 3. Baseline (either a static picture baseline condition that showed the full face or one where the mouth region was obscured).

The timing cue stimuli were constructed in order to present timing cues with minimal form information. To provide a rough estimate of the information available from the obscured mouth videos, five people were asked to identify aspects of the video only presentations. To make this task as easy as possible (pilot testing made it clear that open-set identification of the visual stimuli was not possible), only the broad class of the initial viseme was to be identified and a response set of five viseme classes was provided from which to choose. Participant performance was slightly better than chance ($p = 0.04$ in a one sample T-test) but was highly variable (ranging from 57% to 78% errors) and it is clear from these error rates that not much information about the spoken form was available.

In addition, 30 (15 speech, 15 non-speech) items were prepared for which video stimuli included a written message "Do not respond on this trial" was briefly presented. These items were used as catch trials to ensure that participants would watch visual stimuli throughout the experiment.

Three item blocks (each consisting of 100 experimental items, i.e., 45 speech and 45 non-speech items plus 10 catch trials) were constructed so that each target could appear with each visual (experimental) condition without being repeated in

any block. Each participant went through all the three blocks. The presentation order of the blocks and the items within the block was all random.

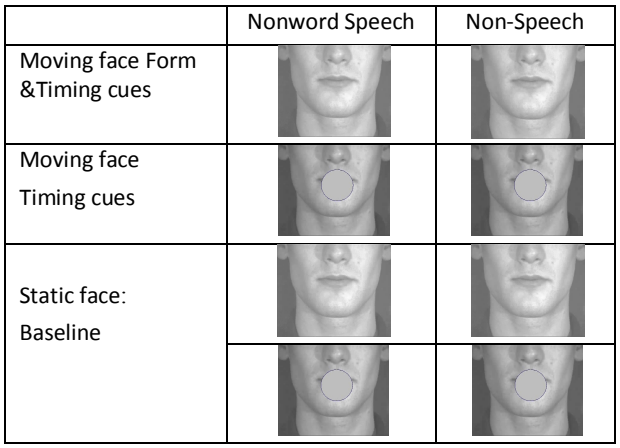

Figure 1: *A depiction of how the talker's face movements were presented in the experiment. The rows show the different stimulus conditions (Form & timing cues; Timing cues; Baseline); the columns represent the Speech and Non-Speech (spectrally inverted versions). In the Baseline, half the stimuli in the Speech and Non-speech conditions were shown with the mouth obscured.*

2.3. Procedure

Participants were tested individually and stimulus presentation and response collection was controlled by the DMDX software [15]. Note that response times were measured from the onset of the auditory speech or non-speech stimuli.

Participants were instructed that on each trial they would see a talker's static or moving face and hear (nonword) speech or non-speech (intermixed throughout the experiment). They were told to respond as soon as practical (speed and accuracy instructions) by pressing a button (with their right hand) if they detected a speech sound or another button (with their left hand) if they detected a non-speech sound. Participants were told to watch the talker's face but it was made clear that the task was to make a response to the heard stimulus. Participants were informed about filler items for which they should not respond.

For practice, 36 extra auditory and visual pairs (speech and non-speech each in moving and baseline conditions, and 8 catch trials) were presented before the experimental items. Each experiment lasted for approximately 30 mins.

3. Results & Discussion

The results of the catch-trials suggested that participants paid attention to the visual presentation. That is, the vast majority of participants (16) correctly made no response to the catch-trials; two participants incorrectly responded to one trial and one participant to two trials. The data reported here were from all of 19 participants.

Table 1 presents the mean response times for each of the experimental conditions.

Table 1. *Mean latencies (ms) for speech and non-speech as a function of visual speech type (Standard Error in parentheses).*

Visual speech type	Response Time	
	Nonword Speech	Non-Speech
Form & Timing cues	564 (23.7)	537 (18.7)
Timing cues	589 (23.3)	544 (17.4)
Static face: Baseline	691 (21.2)	650 (19.3)

The response time data were examined using two analyses of variance (ANOVAs), one for the participant data (collapsed over the different items) and one for the item data (collapsed over the different participants). Overall, the Non-speech stimuli were responded to faster than the Speech ones, $F1(1,18) = 19.56$, $p < 0.05$; $F2(1,42) = 45.72$, $p < 0.05$. There was also a main effect of presentation condition, $F(2,36) = 103.41$, $p < 0.05$; $F2(2,84) = 138.97$, $p < 0.05$. These two effects did not interact, $F(2,36) = 1.18$, $p > 0.05$; $F2(2,84) = 1.0$, $p > 0.05$.

Planned pairwise comparisons of visual speech types were conducted with a Bonferroni adjusted α of 0.0125. For the Speech stimuli there was a significant difference between the Form & Timing cue and Timing cue conditions, $F1(1,18) = 9.33$, $p < 0.012$; $F2(1,41) = 6.74$, $p = 0.0125$, and between the Timing cue condition and the Baseline conditions, $F1(1,18) = 76.11$, $p < 0.012$; $F2(1,41) = 134.50$, $p < 0.0125$. For Non-speech stimuli, the difference between the Form & Timing cue and Timing cue conditions was not significant, both F1 and F2 < 1. The difference between the Timing cue and the baseline conditions was significant, $F1(1,18) = 97.78$, $p < 0.012$; $F2(1,42) = 124.92$, $p < 0.0125$.

The percentage error rates for each of the experimental conditions are shown in Table 2.

Table 2. *Mean percent error rates for speech and non-speech as a function of visual speech type (Standard Error in parentheses).*

Visual speech type	Error Percentage	
	Speech	Non-Speech
Mouth motion visible	2.58 (.49)	3.28 (.84)
Mouth motion invisible	3.52 (.73)	2.70 (.65)
Static face: Baseline	3.29 (.52)	1.76 (.44)

There were no significant differences between Speech and Non-speech, $F1(1,18) = 1.15$, $p > 0.05$; $F2 < 1$, between the presentation types, $F1(2,36) = 1.10$, $p > 0.05$; $F2 < 1$, nor an significant interaction between the two variables, $F(2,36) = 2.024$, $p > 0.05$; $F1(2,82) = 2.19$, $p > 0.05$.

4. Discussion

The current study examined the role of speech form and timing cues in the processing of auditory speech and non-speech in a task designed to be sensitive to both. The results showed that the provision of timing cues facilitated both speech and non-speech processing. With the addition of form cues, response time to the speech stimuli was further reduced but this was not shown for the non-speech stimuli.

The effects of the visual timing cues for both speech and non-speech indicates that the information provided by this cue type is not speech-specific (the effect of presenting the visual timing cue was approximately 100 ms for both speech and non-speech). On the other hand, the effect of providing additional form cues was significant only for speech (a 25 ms facilitation), suggesting that these visual cues provide information specific to the speech signal.

One suggestion for why the form cue effect was confined to the auditory speech stimuli is that the mouth shape and motion provides information about the likely spectral composition of the up-coming auditory speech and that this information can be used to reduce uncertainty as to whether the signal is in fact speech. This proposal fits with the finding that in the speech detection task visual speech only facilitates responses if energy in the F2 region of the auditory signal is intact [7]. Moreover, Girin and colleagues [16] have demonstrated how lip shapes can be used to provide an approximate estimate of the likely spectral properties of auditory speech.

Of course in order for the visual information obtained from mouth shape/motion to influence the response task, such would have to be processed at a relatively early stage. Recent neurophysiological findings suggest that this may be the case, with presentation of visual speech affecting the timing and morphology of an early auditory response component (the N100); indeed, the strength of this visual influence was modulated by the distinctiveness of visual speech ([17; 18]).

The finding that there were no significant differences in errors across conditions suggests that response time may be the most sensitive index of visual speech effects that occur at an early stage of speech processing.

5. Conclusions

We have shown that both visual speech form and timing cues assist in deciding whether a stimulus is speech or not. That both types of information are useful makes sense since changes in the shape of the mouth during speech articulation provide information as to when to expect speech and what some of the spectral properties of speech may be. Of course, whether this information will be useful will depend on the required information needed to perform the task itself.

6. Acknowledgements

The authors thank Leo Chong for assisting in data collection and acknowledge that the study was supported by an ARC Discovery Grant (DP130104447).

7. References

[1] Summerfield, A.Q., "The use of visual information in phonetic perception", Phonetica, 36: 314-331, 1979.

[2] Davis, C. and Kim, J., "Audio-visual speech perception off the top of the head" Cognition, 100: B21-B31, 2006.

[3] Munhall, K.G., Jones, J.A., Callan, D.E., Kuratate, T. and Vatikiotis-Bateson, E., "Visual prosody and speech intelligibility head movement improves auditory speech perception", Psychological Science, 15: 133-137, 2004.

[4] Paris, T., Kim, J. and Davis, C., "The role of visual speech in the speed of auditory speech processing", submitted.

[5] McGurk, H. and MacDonald, J., "Hearing lips and seeing voices", Nature 264: 746–748, 1976.

[6] Munhall, K.G., Gribble, P., Sacco, L. and Ward, M., "Temporal constraints on the McGurk effect", Percept Psychophys, 58:351-62,1996.

[7] Grant, K.W., "The effect of speechreading on masked detection thresholds for filtered speech", Journal of the Acoustical Society of America, 109: 2272-2275, 2001.

[8] Kim, J. and Davis, C., "Investigating the audio-visual speech detection advantage", Speech Communication, 44: 19-30, 2004.

[9] Bernstein, L.E., Auer, E.T. and Takayanagi, S., "Auditory speech detection in noise enhanced by lipreading", Speech Communication, 44: 5-18, 2004.

[10] Eskelund, K., Tuomainen, J. and Andersen, T. S., "Multistage audiovisual integration of speech: dissociating identification and detection", Experimental Brain Research, 208: 447-457, 2011.

[11] Rastle, K., Harrington, J. and Coltheart, M., "358,534 nonwords: The ARC Nonword Database", Quarterly Journal of Experimental Psychology, 55A: 1339-1362, 2002.

[12] Blesser, B., "Speech perception under conditions of spectral transformation: I. Phonetic characteristics", Journal of Speech, Language and Hearing Research, 15: 5-41, 1972.

[13] Kim, J., Davis, C. and Krins, P., "Amodal processing of visual speech as revealed by priming", Cognition, 93: B39-B47, 2004.

[14] Davis, C. and Kim, J., "Repeating and remembering foreign language words: Implications for language teaching system", Artificial Intelligence Review, 16: 37-47, 2001.

[15] Forster, K.I. and Forster, J.C., "DMDX: A windows display program with millisecond accuracy", Behavior Research Methods, Instruments and Computers, 35: 116-124, 2003.

[16] Girin, L., Schwartz, J. L., and Feng, G., "Audio-visual enhancement of speech in noise", The Journal of the Acoustical Society of America, 109: 3007-3020, 2001.

[17] Van Wassenhove, V., Grant, K.W. and Poeppel, D., "Visual speech speeds up the neural processing of auditory speech", Proceedings of the National Academy of Sciences of the United States of America, 102: 1181-1186, 2005.

[18] Arnal, L. H., Wyart, V. and Giraud, A. L., "Transitions in neural oscillations reflect prediction errors generated in audiovisual speech", Nature neuroscience, 14: 797-801, 2011.