Detecting auditory-visual speech synchrony: how precise?

Chris Davis & Jeesun Kim

The MARCS Institute, University of Western Sydney, Australia

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

ABSTRACT

Previous research suggests that people are rather poor at perceiving auditory-visual (AV) speech asynchrony, especially when the visual signal occurs first. However, estimates of AV synchrony detection depend on many factors and previous measures may have underestimated its precision. Here we used a synchrony-driven search task to examine how accurately an observer could detect AV speech synchrony. In this task on each trial a participant viewed four videos (positioned at the cardinal points of a circle) that showed the lower face of a talker while hearing a spoken /ba/ syllable. One video had the original AV timing, in the others the visual speech was shifted 100 ms, 200 ms or 300 ms earlier. Participants were required to conduct a speeded visual search for the synchronized face/voice token (the position of which was randomized). The results showed that the synchrony detection window was narrow with 82% of responses selecting either the original unaltered video (29%) or the video where the visual signal led by 100 ms (53%). These results suggest that an observer is able to judge AV speech synchrony with some precision.

Index Terms: Auditory-visual speech synchrony; Synchrony search task; Inter-sensory timing

1. INTRODUCTION

The integration of multisensory information gleaned from different sensory analyses can assist us in perceiving and responding to objects and events more quickly and accurately. A problem exists in how to determine what information from the different senses should be integrated. The timing of stimulation from the different sense modalities likely provides a cue as to which inputs may belong together. To use this cue, sensory/perceptual mechanisms are required that can determine the simultaneity of inter-sensory signals (with reference to the events that gave rise to them). Attempts to assess the operating characteristics of such mechanisms (e.g., their precision) have largely relied on simply asking people to make judgments of the relationship between the timing of events presented in two different modalities.

For example, the perceived synchrony of auditory-visual (AV) speech stimuli has typically been estimated by using either a simultaneity judgment (SJ) task and/or a temporal order judgment (TOJ) one [1-3]. The SJ task simply consists of presenting an auditory and a visual stimulus to observers (with the SOA of these stimuli varied) and asking her/him to judge whether the stimuli were presented simultaneously or not. These responses are plotted with percent simultaneous responses expressed against AV SOA. The point of subjective synchrony (PSS) is the peak of this function. This response/SOA function also furnishes information about how sensitive the observer is to changes in AV SOA, here sensitivity is often reified as the width of the function at the 75% response level (a measure often glossed as the just noticeable difference, JND).

In the TOJ task, observers are asked to judge whether the A or V stimulus was presented first (again with AV SOA varied). The point PSS can be estimated by plotting a function of the percentage of (say) visual first responses and calculating the SOA where 50% of visual first responses occurred. In this paradigm, sensitivity is estimated by the slope of the function; where the JND is given as half the difference between the SOAs that correspond to the 25% and 75% response points.

Research using these tasks has suggested that the perceivers are very tolerant of asynchronies in AV speech inputs. For example, Dixon and Spitz [1] showed that audiovisual asynchrony was only detected when the visual speech signal leads the auditory speech signal by at least 250 ms. More recently, Maier and colleagues [2] showed that even when vision was presented 287 ms before the auditory signal, this pairing attracted approximately 75% synchrony responses. This results is similar to the results of [3] where there was an 80% simultaneity response rate for stimuli where the auditory signal lagged by 267 ms.

In the context of attempts to estimate the attributes of the mechanisms that determine the relative timing of inter-sensory signals, it is important to realize that measures of AV synchrony do not provide an index of some fixed ability. For one thing, it appears that the TOJ and SJ tasks may be driven by different perceptual processes [4]. Further, different studies have used stimulus materials that vary in duration from single syllables or disyllables [5; 6], single words [7], through to whole sentences [8]. Indeed, it is clear that estimates are affected by multiple stimulus and experimental factors [9; 10] and that the two tasks may have different response biases, i.e., [11; 10].

Since all behavioural estimates of AV synchrony will potentially involve biases and criterion setting it may be that any procedure will obscure how well AV synchrony can be detected. Although this may be the case, it is likely that some measures may be more sensitive than others. One candidate for a procedure that might provide a more direct measure of AV synchrony is that based on a paradigm that examined AV interaction in multi-element arrays and demonstrated that the search for a visual singleton (marked by an abrupt color change) could be greatly facilitated by the presentation of an abrupt synchronized auditory pulse [12]. Recently, Alais and colleagues ([13]) modified this basic procedure by asking four participants to indicate which flickering visual stimulus was synchronous with a sound. That is, in this new paradigm participants searched amongst 19 modulating discs that each had unique temporal phases for one that was synchronized

with a modulating auditory 1.3 Hz tone. Two types of auditory and visual modulation were tested (sinusoidal or square wave) and it was found that AV synchrony detection required transient signals, i.e., the sinusoidal AV modulations did not permit accurate detection of synchrony. The results also showed that the effectiveness of the visual search varied over the visual field, such that error distributions were more tightly tuned temporally on the right side, especially the upper-right quadrant. Importantly, this synchrony-driven visual search paradigm produced estimates of the precision of AV synchrony (the temporal integration window) that were comparatively narrow (±60 ms).

The current study used a simplified version of this AV synchrony-driven search paradigm (one that also bears some similarity with that used by [14]). The simplification was to reduce the number of visual elements that need to be searched and to use AV speech stimuli. Thus, in the current experiment, a participant was required to search among four movies showing the lower segment of the same speaker's face uttering the syllable "ba" (8 times) for the one that was synchronous with the presentation of an auditory /ba/. Note that /ba/ was used because it has a relatively rapid and well defined onset which is important [13].

The aim of the experiment was to determine whether estimates of AV speech synchrony will be narrower than those previously reported. Such a finding would not only add to the literature showing that estimates of AV synchrony can be influenced by the type of measure used and other events in the environment [15], but it would also provide useful information on the degree of precision that can be achieved in such judgments.

2. Experiment

This experiment adopted a psychophysics approach to testing where an estimate of synchrony detection was based on the data of a few experienced observers who were presented a large number of trials.

2.1. Method

2.1.1. Participants

Three observers with normal or corrected-to-normal vision and hearing participated.

2.1.2. Stimuli

Three hundred and sixty stimuli were constructed from a video of a female speaker uttering the syllable /ba/. The video was captured at 30 frames per second and audio at 44 kHz. A /ba/ syllable was selected because it had a rapid auditory rise-time (approximate 5 ms from no sound to peak amplitude in the amplitude envelope) along with a clear visual onset.

Four versions of the video were created (using a custom script in virtualdub [16]) with each version showing a segment of the lower face (see Figure 1). The upper face (particularly the eyes and eyebrows) was not shown as this can provide timing cues as to when the articulation occurred [17].

One of the versions had the original AV timing; in the other versions, the visual component was shifted forward in time (relative to the auditory component) by 100, 200 or 300 ms. These videos will be referred to as +V100; +V200 and +V300. Each of the videos were displayed at the cardinal points of a circle (see Figure 1) that had a radius that subtended 4.9^0 of visual arc for an observer who was 94 cm from the monitor (each face segment subtended 1.5^0 in width and 1.2^0 in height).

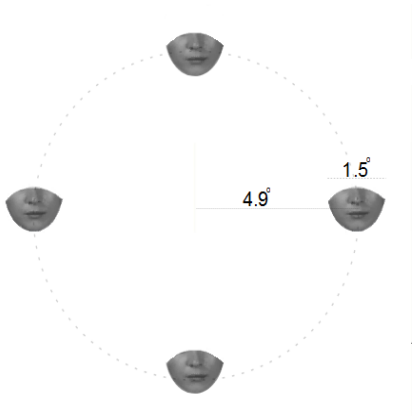

Figure 1. The figure shows a depiction of the stimulus presentation setup (the dotted lines were not presented in the experiment). The videos were displayed at the cardinal points of a circle that subtended a visual angle of 4.9^0. One face had the original AV synchrony, in the others the visual signal was shifted ahead by 100, 200 or 300 ms.

A trial consisted of eight presentations of the speaker uttering /ba/. Eight utterances were presented in order to provide the observer with several instances of the onset of the utterance.

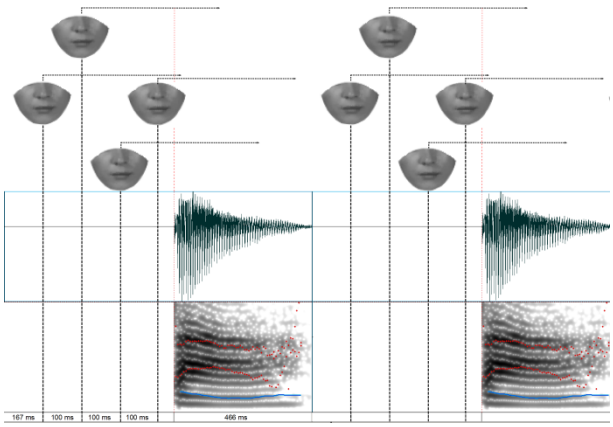

Figure 2. An illustration of the relationship between the visual and auditory components of the four displayed videos. Two cycles of the display are depicted (in all the spoken syllable was uttered 8 times in each trial). The vertical dotted lines and respective faces indicate the time at which there first was face motion. The horizontal lines show the duration of the video. The auditory component is represented by a time-amplitude waveform plot with a time-frequency spectrogram below.

The distribution of when the videos began and ended was set so that the timing of onsets and offsets would be distinct. So for example, the motion of each utterance in the +V300

video was the first to begin after the auditory signal and finished just as the auditory /ba/ component began; it did not begin again until after the auditory component had finished (see Figure 2).

2.1.3. Procedure

Each participant was tested individually in a sound attenuated ICA booth. The experimental session consisted of 360 trials and lasted for approximately 100 minutes with short rests between blocks of 30 stimuli. Each trial consisted of 8 presentations of the syllable /ba/) and lasted for approximately 8 seconds. The participant was informed that only one of the videos was the AV in synchrony and that on each trial the position of this video was to be indicated by pressing the corresponding spatial position on the number pad (i.e., the keys on the cardinal points). The participant's response was echoed on the screen and could be changed within two seconds if an incorrect key press had been made. The next trial followed automatically after this time.

Videos were played on a flat screen 48.3 cm CRT monitor and sound played (binaurally) through an EDIROL UA-25, USB audio interface over Sennheiser HD650 headphones at a comfortable level. The DMDX software package [18] was used to control the display and register responses. There were 6 practice trials at the beginning of the experiment.

2.2. Results

A summary of the results is displayed in Figure 3. The figure shows the percentage of synchrony responses that were elicited by each video. As can be seen there were many more selections of the original and V+100 ms videos compared to the other two videos.

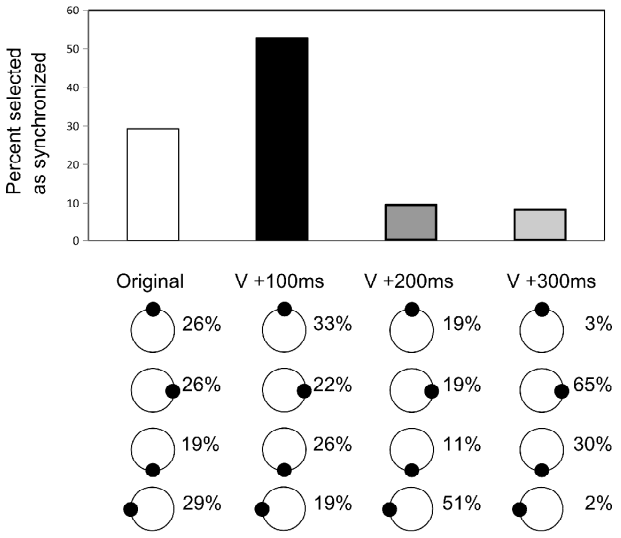

Figure 3. Overall mean percentage of the videos selected as being in sync. The circles below each column show the positions the face was displayed at when selected and the percentage that this position was selected.

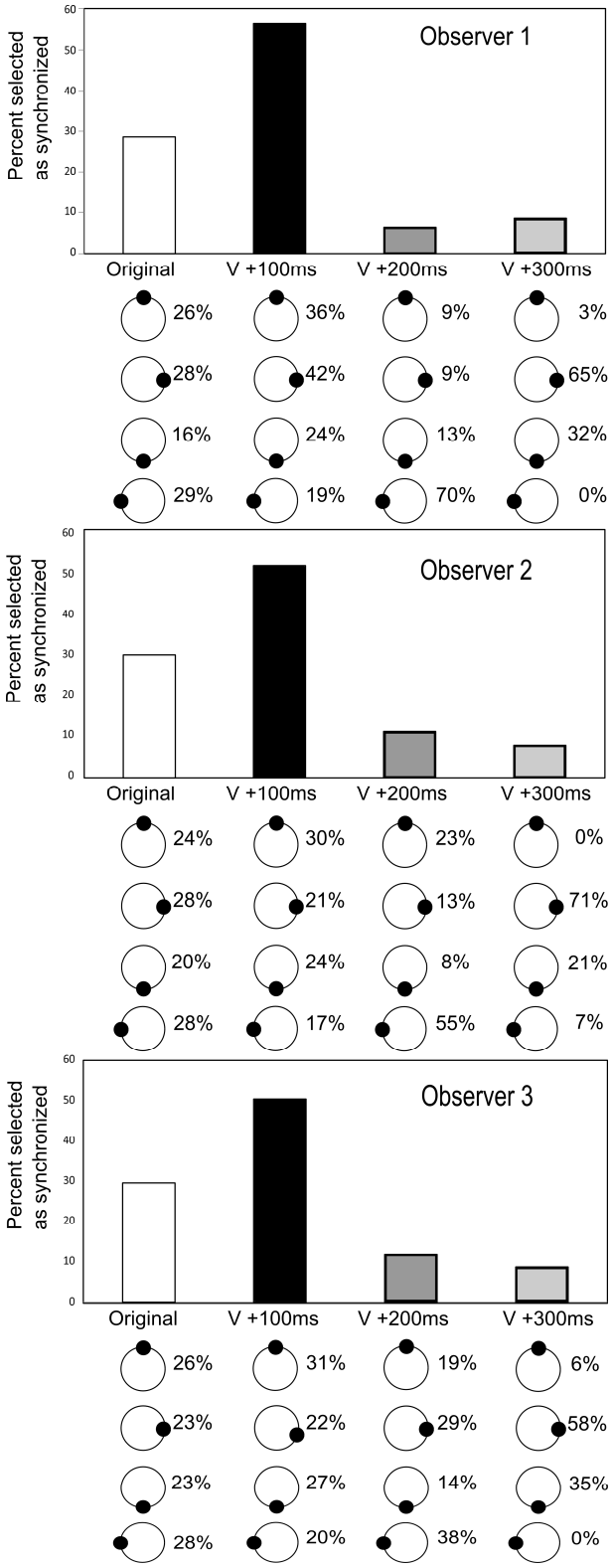

Figure 4. Mean percentage of the videos selected as being in sync for each of the 3 participants.

A chi-square test on the overall data showed that the distribution of the responses differed from that expected by chance, $\chi 2 = 559.8$, $p < 0.05$. Comparing the number of

responses to the original and +V100 responses (summed) against those for the other two videos (summed) again revealed that the response distributions differed, $\chi 2 = 440.8$, $p < 0.05$.

The positions at which the displayed face was selected as being in synchrony with the auditory signal (see the percentages associated with the circles indicating the displayed position in Figure 3) had little effect on the synchrony detection, at least for the original and +V100 videos (which accounted for 82% of responses). This was confirmed by chi-square analysis, for the original videos, $\chi 2 = 2.16$, $p > 0.05$ and for the +V100 ones, $\chi 2 = 4.4$, $p > 0.05$.

An ANOVA (repeated measures for the delay and position variable) was conducted on the overall data to determine if there was a difference in the number of times the V+100 video was selected compared to the original video and whether this effect interacted with position. There was a difference between the two video delays (with the +V100 video attracting more responses), $F(1,6) = 409.90$, $p < 0.05$. There was no difference in number of response selections as a function of where the videos were displayed, $p < 0.05$, and no interaction between the delay and position variables, $p < 0.05$.

Figure 4 shows a breakdown of the data for each participant. As can be seen, these data show very much the same pattern as the averaged data.

A series of chi-square tests showed that the distribution of the responses differed from chance for each participant, $\chi 2 = 232.0$, $p < 0.05$; $\chi 2 = 174.0$, $p < 0.05$; $\chi 2 = 159.1$, $p < 0.05$ (respectively).

We also compared the response totals from the original and V+100 ms conditions against those of the other two conditions (+V200 and +V300 ms) for each participant. The analyses showed that responses to the original and V+100 videos differed from those of the other two, $\chi 2 = 176.4$, $p < 0.05$; $\chi 2 = 140.6$, $p < 0.05$; $\chi 2 = 127.2$, $p < 0.05$.

3. Discussion

Previous research makes it clear that the temporal window over which a person perceives AV synchrony is not fixed. For example, the dimensions of this window are affected by recent experience of AV timing [15] and different ways of measuring it can produce different estimates [11]. The dynamic nature of this AV temporal window makes it important to estimate boundary conditions for AV synchrony perception, e.g., is it possible to obtain estimates that show a relatively narrow temporal window? The current study used a synchrony-driven visual search paradigm to estimate the synchrony window for AV speech (as this paradigm produced a narrow integration window for non-speech AV signals, [13]).

We found that the bulk of the videos selected as synchronous (82% of responses) were for stimuli that had the original AV synchrony (29%) or where the video component was shifted ahead to the audio by 100 ms, V+100 (53%). This estimate of the temporal synchrony window appears to be much narrower than that found with the SJ paradigm, where high rates (~75-80%) of simultaneity responses occur in cases where the visual signal has been shifted by several hundred milliseconds, e.g., [1; 2; 3].

Unlike, the study by Alais and colleagues [13] we found no evidence for a right quadrant bias in the response sensitivity. This may have been due to the difference between the types of stimuli used in the two studies ([13] used modulating discs and tones) or perhaps due to the number of elements presented in an array ([13] had 19 compared to four in the current study).

Another difference between the current results and that of [13] was that in the current study, the stimulus that was most often selected as synchronous was the one in which the visual speech was shifted ahead of the auditory component by 100 ms (3 video frames), in [13] the synchrony window was centred on zero phase. Once again, this disparity may have been due to difference in the setups. In the current study we only examined shifts in AV synchrony in which the visual component preceded the auditory one (in [13] an equal number of +Visual and –Visual shifts were tested).

It should be pointed out, however, that finding that the point of subjective AV simultaneity is shifted toward visual first is actually not new [1; 6] and a number of ideas have been suggested for why this might be the case. These range from explanations based on differences in AV transduction to those that make reference to the customary association of signals (e.g., in speech typically face and jaw motion occurs before acoustic speech).

So, why might the synchrony search task give a comparatively narrow estimate of the AV temporal integration window compared to other measures? This paradigm provides stimuli that have different AV synchronies and a comparison with these will provide a basis for synchrony selection. In addition, as the results of the auditory assisted visual search task show [12], AV search within multi-element displays can be very efficient (at least with AV signals that have clear transient onsets). Thus, the synchronous AV video may stand out against the others and then subsequent comparisons with these videos can provide confirmation of synchrony.

4. Acknowledgements

The authors acknowledge support from the Australian Research Council (DP130104447).

5. REFERENCES

[1] Dixon, N. F., & Spitz, L. (1980). The detection of auditory visual desynchrony. Perception, 9, 719–721.

[2] Maier, J. X., Di Luca, M., & Noppeney, U. (2011). Audiovisual asynchrony detection in human speech. Journal of experimental psychology. Human perception and performance, 37, 245.

[3] van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. Neuropsychologia, 45, 598-607.

[4] Love, S. A., Petrini, K., Cheng, A., & Pollick, F. E. (2013). A Psychophysical Investigation of Differences between Synchrony and Temporal Order Judgments. PloS one, 8, e54798.

[5] Jones, J. A., & Jarick, M. (2006). Multisensory integration of speech signals: the relationship between space and time. Experimental Brain Research, 174, 588-594.

[6] Vatakis, A., & Spence, C. (2006). Audiovisual synchrony perception for music, speech, and object actions. Brain research, 1111, 134.

[7] Conrey, B., & Pisoni, D. B. (2006). Auditory-visual speech perception and synchrony detection for speech and nonspeech signals. The Journal of the Acoustical Society of America, 119, 4065.

[8] Vatakis, A., & Spence, C. (2006). Audiovisual synchrony perception for speech and music assessed using a temporal order judgment task. Neuroscience letters, 393, 40-44.

[9] Vroomen, J., & Stekelenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial

audiovisual stimuli. Journal of Cognitive Neuroscience, 22, 1583-1596.
[10] Vroomen, J., & Keetels, M. (2010). Perception of intersensory synchrony: a tutorial review. Attention, Perception, & Psychophysics, 72, 871-884.
[11] van Eijk, R. L., Kohlrausch, A., Juola, J. F., & van de Par, S. (2008). Audiovisual synchrony and temporal order judgments: Effects of experimental method and stimulus type. Perception & psychophysics, 70, 955-968.
[12] Van der Burg, E., Olivers, C. N., Bronkhorst, A. W., & Theeuwes, J. (2008). Pip and pop: nonspatial auditory signals improve spatial visual search. Journal of Experimental Psychology: Human Perception and Performance, 34, 1053-1065.
[13] Alais, D. Cass, J & van der burg, E. (2013). Spacial and temporal precision of visual search driven by audiovisual synchrony. 40th Australasian Experiment Psychology Conference, 3-6 April, 2013. Adelaide.
[14] Alsius, A., & Soto-Faraco, S. (2011). Searching for audiovisual correspondence in multiple speaker scenarios. Experimental brain research, 213, 175-183.
[15] Roseboom, W., Nishida, S., & Arnold, D. H. (2009). The sliding window of audio–visual simultaneity. Journal of Vision, 9(12):4, 1–8, http://journalofvision.org/9/12/4/, doi:10.1167/9.12.4.
[16] Lee, A (2010). VirtualDub 1.9.11. Build 32842. [software] home page. URL: www.virtualdub.org.
[17] Kim, J., Cvjeic, E., & Davis, C. (in press). Tracking eyebrows and head gestures associated with spoken prosody. Speech Communication.
[18] Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. Behavior Research Methods, Instruments, & Computers,35, 116-124.