Crossmodal Enhancement: Investigating the role of timing in audio-visual enhancement.

Sheena Luu and Willy Wong
Institute of Biomaterials and Biomedical Engineering
and Edward S. Rogers Sr. Department of Electrical and Computer Engineering,
University of Toronto, Toronto, Canada
sheena.luu@utoronto.ca   willy@eecg.utoronto.ca

Abstract

The results of a set of three psychophysical experiments investigating crossmodal enhancement using sinusoidally amplitude modulated (SAM) audio and visual stimuli are presented. The experiments are designed to probe the importance of timing in crossmodal enhancement. Our findings suggest that synchrony is a requisite for enhancement to occur. The results also suggest that although some enhancement may be due to the visual stimuli providing timing information for the detection of the audio stimuli, timing information alone cannot explain the full magnitude of the observed enhancement. Given the importance of synchrony perception in crossmodal integration, a model of audio-visual synchrony detection based on windowed cross-correlation is also introduced and discussed.

Current literature suggests that auditory perception can be enhanced by signals received in the visual domain. However, most studies have used either really simple signals (noise bursts and light flashes), which can be described as being ‘irrelevant’ or ‘uninformative’ to the auditory signal, or highly complex signals like audio-visual speech. In this paper we describe experiments that attempt to fill in the gap between these two extremes through the use of sinusoidally amplitude modulated (SAM) audio and visual signals. We investigate audiovisual integration through psychophysical experiments by comparing the ability of participants to detect SAM auditory stimuli in the presence and absence of a co-modulated visual stimuli. We also vary the temporal relationship between the visual and auditory stimuli in order to probe the underlying process in the brain that integrates information from the different senses.

Experiment 1

The purpose of experiment 1 is to answer the question: *Can a co-modulated visual signal enhance auditory detection of a sinusoidally amplitude modulated (SAM) tone?* The results of Experiment 1 provide a baseline to which the other experiments can be compared.

Subjects

There were 6 participants in this experiment (2 female, 4 male, mean age 26.5).

Apparatus

Video stimuli were presented via a CRT monitor (refresh rate: 160Hz, resolution: 640x480). Audio stimuli were presented via headphones. Stimuli were presented via test software developed in house using C++ and DirectX 9.0c libraries. Relative timing of audio and visual signals was verified using a photoresistor circuit and oscilloscope and was determined to have an error within ±3ms.

Stimuli

Experiments were run in a sound-attenuated chamber with reduced ambient light. The audio-visual stimulus was 2 seconds in duration.

The auditory stimulus was a suprathreshold 1kHz pure tone, sinusoidally amplitude-modulated at a frequency of 3Hz, at a sampling frequency of 44.1kHz. There were two types of visual stimuli. In the Audio-only condition, the visual stimulus was simply a static eye-fixation point which appeared for the duration of the tone was played (see Figure 1a). In the co-modulated Audio-Visual condition, the visual stimulus was a Gaussian blob on a black background, with a maximum luminance of approximately 300cd/m^2. The visual stimulus was sinusoidally amplitude modulated with the same function as the auditory signal (see Figure 1b). In both conditions, the audio signal was the same.

Procedure

The auditory detection threshold for each subject in each condition (audio-only and audiovisual) was estimated separately using a two-interval, two-alternative forced choice (2-AFC) task. The same video is presented in both intervals but only one of the intervals (randomly determined) contained the tone. The subject's task was to answer the question: *In which interval was the audio presented?* Data was collected over a range of sound levels and then curve-fitted to estimate a threshold. Each session lasted approximately 20min and consisted of randomized presentation of 5 sound levels, 20 trials at each sound level.

Results and Discussion

Curve fit using a hyperbolic tangent allowed us to estimate the auditory detection threshold at the 75% correct point on each psychometric curve. The threshold results are shown in Figure 1c. Each data point is the average accuracy over 80 trials. On the plot, AV stands for audiovisual stimuli with a 3 Hz modulating frequency. Similarly A stands for audio-only stimuli with a 3Hz modulating frequency.

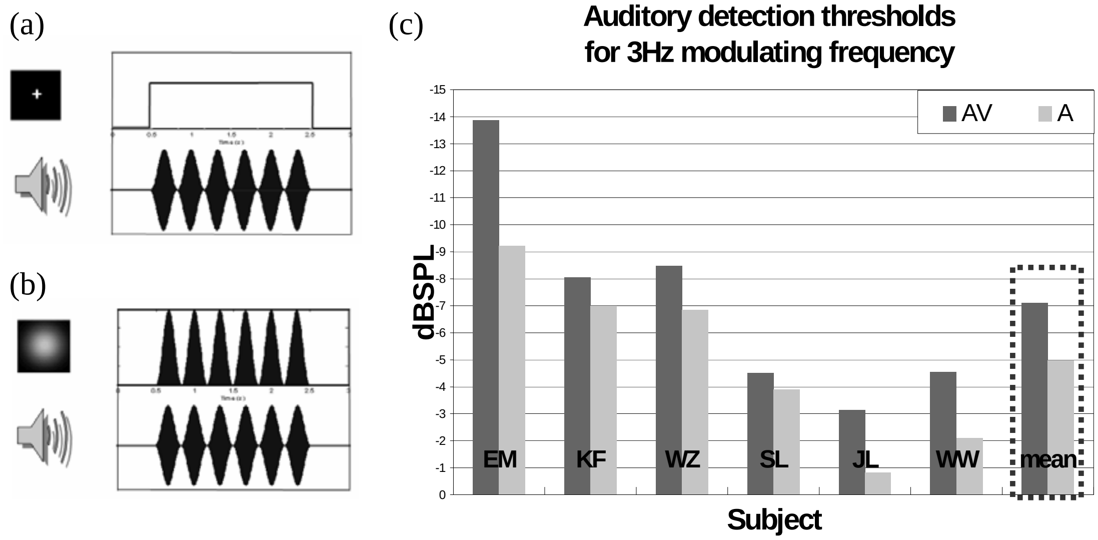

Fig. 1. Experiment 1 - (a) Audio only stimulus, (b) Audio-visual stimulus (c) Resulting auditory detection thresholds in the Audio-only and Audio-visual conditions.

Each subject had a lower threshold in the Audio-visual condition compared to the Audio-only condition. The mean Audio-only threshold over all subjects is -5.0 dBSPL. This value is compared to the mean Audio-visual threshold over all subjects, which is -7.1 dBSPL. Thus, on average, an improvement of approximately 2.1 dB is observed due to crossmodal stimulation. A two-tailed paired t-test shows that this difference is significant (p=0.015).

The results found here are in agreement with past experiments from other studies (see Grant and Seitz 2000; Odgaard et al 2004) where similar enhancement magnitudes have been reported.

Experiment 2

Experiment 2 is designed to answer the question: *Is AV synchrony perception related to crossmodal enhancement?*

Synchrony perception is the perception that two sensory stimuli occur simultaneously. Fujisaki and Nishida 2005 found that with sinusoidally amplitude modulated audio-visual stimuli, the threshold for audiovisual synchrony detection is approximately 3Hz. At higher modulating frequencies, discriminating between in-phase and 180^o out-of-phase audio-visual signals becomes increasing difficult. We have made use of these findings to design Experiment 2.

Three of the participants in Experiment 1 returned to take part in Experiment 2. We repeated Experiment 1 at two other frequencies of 1.5Hz and 6Hz.

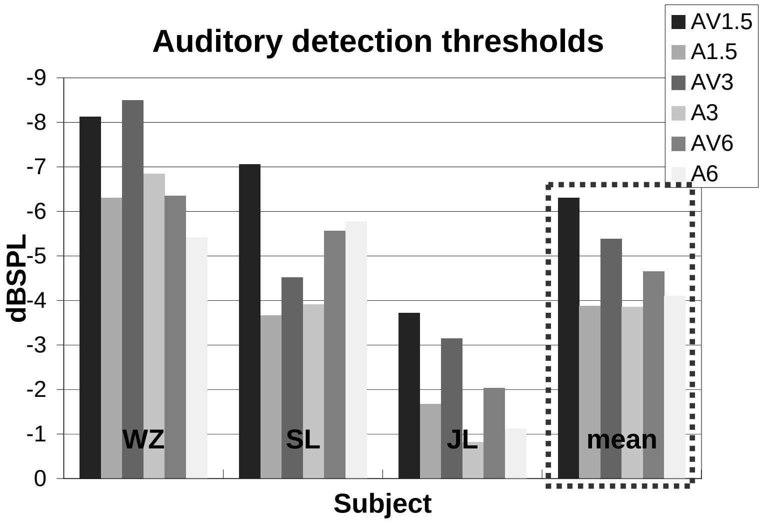

Fig. 2. Experiment 2 - Resulting auditory detection thresholds for the Audio-visual and Audio-only conditions for 3 modulating frequencies (1.5Hz, 3Hz and 6Hz).

Results and Discussion

The results are plotted in Figure 2. The graph shows that for low frequencies (1.5Hz and 3Hz) the co-modulated visual stimulus improved auditory detection compared to the audio-only stimulus. However, the advantage conferred by a correlated visual signal is noticeably diminished at higher frequencies (6Hz).

A comparison of our data with the synchrony detection data from Fujisaki and Nishida (2005) suggests that audiovisual synchrony perception and audiovisual enhancement are interrelated. Audiovisual synchrony perception may play a role in the mechanism of audiovisual enhancement.

Experiment 3

Timing information, in knowing when to listen for an auditory signal, is known to reduce thresholds of detection (Egan et al 1961). “Peak listening” (Grant and Seitz 2000) is one possible explanation for the visual enhancement of speech detection. Experiment 3 is designed to answer the question: *Is the observed crossmodal enhancement due to the visual stimulus providing a timing cue for when to listen for the audio signal?*

Three of the participants in Experiment 1 returned to take part in Experiment 3. The procedure is the same as in Experiment 1, but uses the following modified visual signals. The Audio-visual 180^o out-of-phase condition is illustrated in Figure 3a. In this condition, the visual signal provides full timing information for when to listen for the peaks in the auditory signal. However, the stimuli are 180^o out-of-phase, so the intensities of the audio and visual stimuli change in opposite directions. The Audio-visual cue condition is illustrated in Figure 3b. In this condition the visual signal is a dynamically updated time-line with the auditory sinusoidal envelope overlaid on the time bar. Once again, full timing information is provided by the visual cue. Experiment 3 was conducted at a modulating frequency of 3Hz.

In designing these video signals, we have avoided a comodulation relationship between the audio and visual stimulus intensities while allowing the video signal to convey full timing information to the participant. If enhancement is due solely to the visual signal conveying timing information to the subject, then under these new visual conditions, we should observe an enhancement that is similar in magnitude to the enhancement in the original comodulation condition.

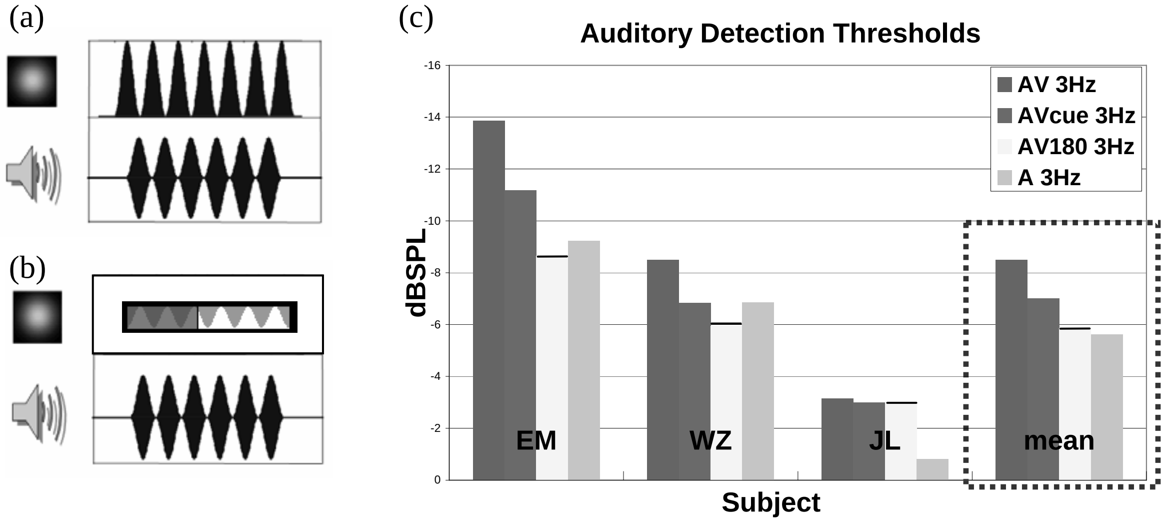

Fig. 3. Experiment 3 - (a) Audio-visual 180^o out-of-phase stimulus, (b) Audio-visual timing cue stimulus (c) Resulting auditory detection thresholds

Results and Discussion

The results for Experiment 3 are shown in Figure 3c. For ease of comparison, the graph also shows the results for these participants from Experiment 1. All three subjects showed an improvement in detection threshold for the Audio-visual cue condition; however, on average the improvement from the timing cue has only half the magnitude of the enhancement observed in the co-modulated Audio-visual condition.

The average thresholds for the Audio-only and Audio-visual 180^{o} out-of-phase are essentially the same. These results suggest that the improvement in auditory detection threshold conferred by a co-modulated visual signal is not due solely to providing a timing cue for when to listen for the peaks of the audio signal.

Discussion of experimental results

Experiment 1 has shown that a corresponding synchronous visual signal can improve auditory detection of a sinusoidally amplitude modulated tone significantly, by 2 dB on average. Experiment 2 has shown that the amount of enhancement is affected by frequency of the modulation of the audiovisual signals. Our results show that the relationship between enhancement and modulation frequency is very similar to the temporal frequency characteristics of synchrony detection in humans. Experiment 3 has shown that he amount of enhancement cannot be explained simply by the visual signal providing a timing cue for when to listen for the high energy peaks of the auditory signal. Thus, although the ability to perceive synchrony plays a role, the role is not simply the providing of characteristic timing information of the signal.

Given the importance of synchrony detection in our results, we take this opportunity to discuss briefly a windowed cross-correlation based model of synchrony detection.

A windowed cross-correlation based model of synchrony detection

We propose that synchrony and asynchrony discrimination can be modeled as a cross-correlation based process. The model is based on the following four assumptions.

Assumption 1: Limited temporal resolution gives rise to temporal uncertainty in both the auditory and visual inputs. It is widely accepted that individually the sensory modalities have limited temporal resolution due to temporal integration in the early processing stages of each modality. We model temporal uncertainty using Gaussian-shaped temporal integration windows in the auditory and visual preprocessing stage with standard deviations of 5ms and 50ms respectively which are appropriate for the stimuli used in our experiments (Barlow 1958; Plack and Moore 1989). A larger window width indicates a coarser temporal resolution and thus larger temporal uncertainty.

Assumption 2: There is a relative latency in the neural processing and transmission of auditory and visual inputs. It is known that there is a difference in transmission and processing time between auditory and visual information in the brain. This relative delay between auditory and visual inputs is the first parameter in the model.

Assumption 3: Synchrony detection compares relative changes in the audio and visual information streams. One way of detecting audiovisual synchrony is by monitoring whether changes in the two streams of information are temporally coincident. Mathematically this suggests differentiation, which may be accomplished biologically through neurons with phasic responses.

Assumption 4: Synchrony processing involves a calculation of cross-correlation. Mathematically, cross-correlation is used to determine how well two streams of information match at various delays. Cross-correlation is well-established in modeling physiological systems that require timing comparisons, see for example interaural hearing (Jeffress 1948). It is important to realize that the output of the cross-correlation mechanism is not a single value; instead, it is an array of values corresponding to how well the streams match at delays ranging from negative (video lag) to positive (audio lag). The output of the cross-correlator will have a maximum value at the estimated delay between the auditory and visual inputs. The width of the temporal window over which correlation is calculated is the second parameter of the model. Evidence for a cross-correlator in audiovisual synchrony detection has been presented by Banks et al (2006).

Decision Device

The decision device computes the best lag estimate given the cross-correlator output. If the input were single pulse stimuli, finding the best lag estimate is as simple as finding the single peak in the cross-correlator output. However, given pulse train stimuli, more sophisticated population coding techniques will be required.

The internal sensory environment is inherently noisy. The best lag estimate from the output of the cross-correlator will therefore be affected by noise. By the central limit theorem, we assume the estimate follows a normal distribution centered on the mean delay between the audio and visual inputs. A second, 'noise-only' distribution represents the internal standard of synchrony. To first order, we assume both distributions have identical variances. The variance is the third parameter of the model. The window of synchrony detection is taken to cover two standard deviations from 0-delay. The window indicates the tolerance over which the signals are considered "synchronous" to the perceptual system.

The current model only processes information about the envelope of the signal. Other signal characteristics such as the frequency of the tone, or the size and shape of the visual input may affect synchrony detection, but are not included in the preliminary model.

Acknowledgements

This research was sponsored by grants from the Defence and Research Development of Canada, NSERC, and the University of Toronto.

References

Barlow HB (1958) Temporal and spatial summation in human vision at different background intensities. *Journal of Physiology* 141; pp. 337–350

Banks MS, Burr D, Morrone C (2006) Auditory-visual temporal discrimination: Evidence for usage of a temporal cross-correlator? *International Multisensory Research Forum (Dublin)*

Egan, J. P., Greenberg, G. Z., & Shulman, A. I. (1961) Interval of time uncertainty in auditory detection. *Journal of the Acoustical Society of America, 33;* pp. 771–778.

Fujisaki, W. & Nishida, S. (2005) Temporal frequency characteristics of synchrony-asynchrony discrimination of audio-visual signals. *Experimental Brain Research*, *166,* 3-4; pp. 455-464.

Grant, K.W., & Seitz, P. (2000) The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America, 108*; pp. 1197–1208.

Jeffress LA (1948) A place theory of sound localization. *Journal of Comparative and Physiological Psychology 41*; pp. 35–39

Odgaard EC, Arieh Y, Marks LE (2004) Brighter noise: Sensory enhancement of perceived loudness by concurrent visual stimulation. *Cognitive, Affective, & Behavioral Neuroscience 4,* **2**; pp. 127–132

Plack CJ, Moore BC (1989) Temporal window shape as a function of frequency and level. *Journal of the Acoustical Society of America* 87, **5**; pp. 2178–2187