AUDIOVISUAL SYNCHRONY THRESHOLDS

Sheena Luu, Ewen MacDonald, Hafiz Noordin and Willy Wong
Institute of Biomaterials and Biomedical Engineering
and Department of Electrical and Computer Engineering, University of Toronto

INTRODUCTION

An important factor in multisensory integration is synchrony. From a sensory perspective, synchrony refers to the perception that two events occurred at the same time. Past studies in audio-visual synchrony have revealed a number of interesting findings. Namely, there is a difference in the minimal detectable lag depending on whether the audio signal precedes the video signal or vice versa [1] [2] and the perception of synchrony can be recalibrated after adaptation to asynchronous stimuli [3].

Our current study builds upon earlier work both experimentally and theoretically. We have carried out experiments to explore the effect of prior knowledge and expectation on the detection of synchrony. By expanding on earlier methodologies, we have shown that the perception of synchrony is not significantly affected by prior knowledge of lag type (i.e. visual signal precedes audio signal or vice versa). This result suggests that higher level cognitive processes like expectation do not play a significant role in synchrony perception. We also introduce here a systems-level model based on cross-correlation which is compatible with both the observations of our experiments and with the observations of other studies.

EXPERIMENTS

Subjects

Participants were the three authors and five volunteers (6 men, 2 women). Subjects had no known hearing loss and normal or corrected-to-normal vision. The five volunteers were naive as to the purpose of the experiment. Experiments were approved by the ethics board of the University of Toronto (#18435) and informed consent was obtained from each subject prior to participation in the experiments.

Apparatus

A Dell Optiplex PC with a 3.00GHz Pentium-4 processor and 512MB RAM running the Windows XP operating system was used as the platform for the experiments. Video stimuli were presented via an ATI Radeon 9250 PCI video card and a Samsung SM950P CRT monitor (refresh rate: 160Hz, resolution: 640x480). Audio stimuli were presented via headphones (Madsen Electronics TDH 3910) connected to a Digital Audio Labs CardDeluxe sound card using ¼" TRS connectors. Stimuli were presented via test software developed in house using C++ and DirectX 9.0c libraries. Timing of audio and visual signals was verified using a photoresistor circuit and oscilloscope to measure the onset of audio and visual signals. The audio and visual synchrony was determined to have an error within ±3ms.

Stimuli

The auditory stimulus was a suprathreshold 1kHz pure tone at a sampling frequency of 44.1kHz. The tone had a 1ms rise-time and was steady at 70dBSPL for 160ms before decaying to silence over 330ms.

The visual stimulus was a white disk with luminance approximately 300cd/m^2 on a black background presented for 1s. The disk was measured to be approximately 3.2cm, which approximately subtended 5^o of the foveal field. Experiments were run in a sound-attenuated chamber with reduced ambient light.

The auditory and visual stimuli were designed so that synchrony judgments would be based solely on onset times and would not be influenced by the relative offset times of the stimuli.

Asynchronous audiovisual stimuli were created by introducing a relative delay in the onsets of the audio and visual signals. Audio lag stimuli are asynchronous stimuli in which the visual signal precedes the audio signal. Similarly, video lag stimuli are asynchronous stimuli in which the audio signal precedes the video signal. In our experiments asynchronous delays (both audio lag and video lag) ranged from 10ms to 150 ms in 20 ms increments.

Procedure

Each subject participated in three, two-alternative forced choice experiments. The three experiments were very similar and differed in the type of asynchronous stimuli presented. The *ALAG* experiment consisted of only synchronous and audio lag stimuli. The *VLAG* experiment consisted of only synchronous and video lag stimuli. And the *Mixed* experiment was made up of synchronous stimuli randomly paired with audio lag and video lag stimuli. The subject was informed of the type(s) of asynchrony to expect prior to each session. The *ALAG* and *VLAG* experiments are collectively referred to as the 'Separate' testing paradigm and the *Mixed* experiments are referred to as the 'Mixed' paradigm.

Each subject participated in the three experimental types over four sessions per experiment. One experimental session lasted approximately an hour (including a training run at the beginning of each session) and consisted of a series of two alternative forced choice trials. Each trial consisted of two intervals in which a synchronous audiovisual stimulus was presented in one interval and an asynchronous stimulus (with a randomly chosen asynchronous delay) was presented in the other. After attending to both intervals, the subject responded to the question "Which stimulus was most synchronous" by selecting one of two buttons labeled "First" or "Second" indicating the first or second audiovisual stimulus respectively. Responses were collected via a computer keyboard. A uniformly distributed random delay of 0ms to 500ms was inserted between stimulus presentations to ensure that subject responses would not be influenced by inter-stimulus time intervals. Between audiovisual stimuli presentations the screen was dark and blank and no sound was played.

For each subject, 60 data points were collected for each asynchrony delay in each experimental paradigm.

Results

By plotting the percentage of correct responses for each asynchrony delay, the data for each subject was used to create four psychometric curves, one for each of the two delay types in each of the two paradigms (separate audio lag, separate video lag, mixed audio lag and mixed video lag).

In general, the psychometric curves extended from approximately 50% for an asynchronous delay of 10ms to almost 100% for a delay of 150ms. The results can be interpreted as follows: Audio and visual signals presented at a relative delay of 10ms will appear synchronous. That is, the subject cannot reliably detect which of the two intervals contain the delay and hence will be guessing in their response. At the other end, for a delay of 150ms or more, subjects responded correctly almost 100% of the time because the stimulus was clearly asynchronous.

A curve fit using a hyperbolic tangent allowed us to interpolate the data to infer an audiovisual asynchrony threshold at the 75% correct point on each of the psychometric curves. The thresholds for each subject are a measure of the maximum tolerable lag between audio and visual stimuli before the perception of synchrony is broken. In other words, the threshold values indicate the relative delay between auditory and visual onsets before they appear asynchronous.

The average threshold in each experiment over all subjects is shown in Figure 1.

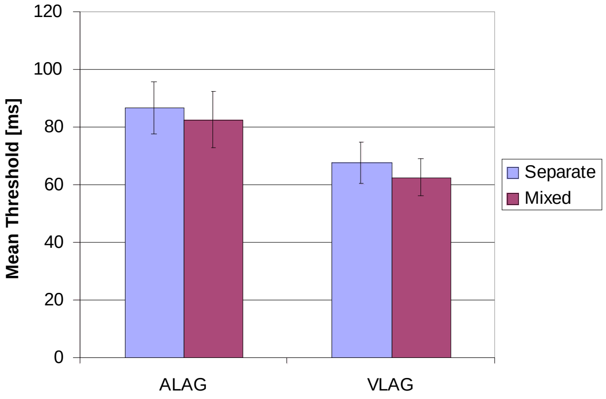

Figure 1: Average audiovisual synchrony thresholds for each type of lag and each experimental paradigm.

A repeated measures ANOVA confirmed that there was a significant main effect (F(1,6)=7.425, p=0.036) of lag type (audio or video lag). There was no significant main effect of paradigm (Separate versus Mixed) and no significant interaction between paradigm and lag type.

Discussion

The results we have found here are in agreement with past experiments from other groups [1] [2] who have indicated that most people have a larger tolerance for perceiving synchrony when audio lags video than when audio precedes video. This asymmetry in synchrony judgment may be a result of adaptation to the physical world where light travels faster than sound [1] [2] [3]. Another possible factor is the unequal neural latencies in the transmission of audio and visual signals in the brain. The auditory and visual pathways to the superior colliculus have delays of approximately 13ms and 80ms respectively [6]; thus, a delay in the presentation of the audio signal would balance this transmission inequality.

A comparison of the results from the single-lag-type and the mixed-lag-types experiments lead us to conclude that prior knowledge and expectation do not have a significant role in audiovisual synchrony detection. The additional information of the lag type to expect did not make the task any easier, as evidenced by the insignificant difference in performance between the Separate and Mixed testing paradigms. This observation suggests that the multisensory temporal processing required for synchrony perception is a lower-level process completed in relative isolation from higher-level cognitive information.

A CROSS-CORRELATION BASED MODEL

Based on our experiments described above and on previous studies by other groups, 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. Because it is not clear at what level bimodal processing takes place, it may be that the temporal uncertainty of sensory information in this context is not fixed and can be modified by factors such as intensity of presentation [5] and possibly attention. In general, vision is known to have greater uncertainty than audition, which is the basis for the so-called auditory dominance in temporal judgments [4] [5].

As a first approximation, we model temporal uncertainty using Gaussian-shaped temporal integration windows are used in the auditory and visual preprocessing step with standard deviations of 5ms and 50ms respectively [7] [8]. A larger window width indicates a coarser temporal resolution and thus a larger temporal uncertainty. (More detailed models of temporal integration in vision and hearing can be found in [7] and [8].)

Assumption 2: There is a relative latency in the neural processing or transmission of auditory and visual inputs.

Auditory and visual pathways to the superior colliculus have delays of approximately 13ms and 80ms respectively [6]. Because multisensory processing is not localized specifically in the superior colliculus, these particular delays may not be the only factors in the temporal incongruence between auditory and visual information.

The relative delay in processing or transmission between auditory and visual inputs is the first parameter in the model.

Assumption 3: Synchrony processing involves a calculation of cross-correlation. The output of the cross-correlation yields an estimate of the delay between the auditory and visual inputs.

Our model simply assumes that processing is distributed over multimodal areas and the processing ultimately can be modeled based on cross-correlation.

Cross-correlation is well-established in modeling physiological systems, including interaural hearing as well as low-level motion detection in vision. Modeling crossmodal synchrony detection as a cross-correlator was hypothesized in [4] but a computational model was not put forth.

In this model, cross-correlation is calculated over a sliding window of fixed width. The size of this window, T, adds a second parameter to the model. The output of the cross-correlator will have a maximum value at the estimated delay between the auditory and visual inputs.

A high-level diagram of the model illustrating the first three assumptions is shown in Figure 2.

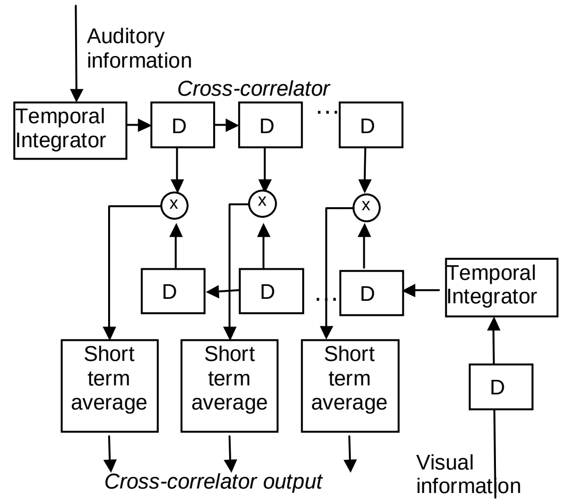

Figure 2: High-level diagram of the cross-correlation based model.

Assumption 4: Decision making can be simulated using a signal detection-type model.

We assume that the output of the cross-correlation is noisy, and synchrony detection is based on Gaussian-shaped noise curves centered on the estimated delay between auditory and visual inputs. In addition, we assume a Gaussian-shaped curve centered at 0-delay to represent the internal standard of synchrony.

To model our experimental results, it is important to notice that the classic signal detection model for a two-alternative forced choice experimental paradigm does not apply here because participants are not asked to simply distinguish between the two presented stimuli, but actually asked to identify which appears *more synchronous*. Thus, the four cases to be considered are illustrated in Table 1.

Table 1: Two Alternative Forced Choice Outcomes

	Synchronous Stimulus	**Asynchronous Stimulus**
1	Perceived as synchronous	Perceived as asynchronous
2	Perceived as synchronous	Perceived as synchronous
3	Perceived as asynchronous	Perceived as asynchronous
4	Perceived as asynchronous	Perceived as synchronous

Combining this logic with the noisy time delay estimates, the percentage of correct responses can be modeled probabilistically as in (1).

$$\begin{aligned} P_{correct} &= P(s\mid S)P(a\mid A) + w_a P(a\mid S)P(a\mid A) \\ &+ w_s P(s\mid S)P(s\mid A) \\ w_a &= \left(\frac{P(a\mid A)}{P(a\mid S)+P(a\mid A)}\right) \\ w_s &= \left(\frac{P(s\mid S)}{P(s\mid S)+P(s\mid A)}\right) \end{aligned} \tag{1}$$

Where *s* is the perception of synchrony, and *S* is the synchronous stimulus. Similarly, *a* is the perception of asynchrony, and *A* is the asynchronous stimulus.

Model Simulation

A simple optimization search was done to find a set of values for the model parameters to fit the experimental results. The average experimental results from all 8 subjects were used to fit the 3 parameters of the model. The model parameters were the relative neural latency of auditory and visual transmission (δ), the size of the covariance window (T), and the standard deviation of the Gaussian noise (σ_n)affecting the decision.

The experimental and simulated results are shown in Figure 2. The experimental curve was created from the average over all subject data from the Separate paradigm experiments. Negative audio delays indicate that the auditory stimulus preceded the video stimulus. Notice that the curve is not symmetrical around zero delay and that subjects showed greater tolerance for perceiving synchrony when audio lagged video than when video lagged audio.

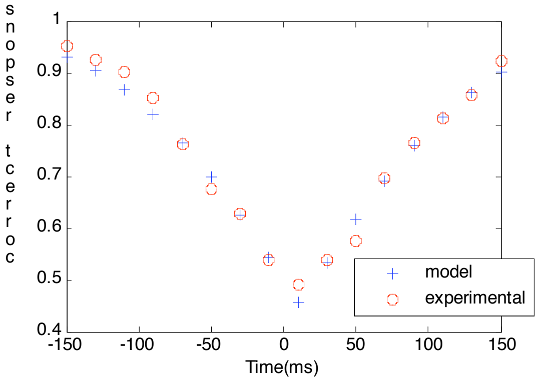

Figure 3: Average experimental and simulated results. Percentage of correct responses is plotted for each audio delay.

The Pearson correlation coefficient, r, between the experimental and simulated data was greater than 0.993 for the following parameter values: covariance calculated over a window of T=233ms; a visual delay of δ=12ms relative to auditory transmission and processing; and the Gaussian noise affecting decision making had a standard deviation of σ_n=68ms.

Fitting the model to the experimental data yields parameter values that lie within a biologically plausible range.

CONCLUSION

The results of our experiment are compatible with past observations that subjects have a larger tolerance for perceiving synchrony when the auditory signal lags the visual signal than vice versa. Moreover, the results of our experiment suggest that higher level cognitive information, specifically, knowledge and expectation of the lag type, have a limited role in the perception of synchrony. We have also presented a cross-correlation based model that can simulate the results of our synchrony threshold experiment.

ACKNOWLEDGEMENTS

This research was supported by the NSERC and Defence Research and Development Canada (DRDC).

REFERENCES

[1] S. van de Par and A. Kohlrausch, “Integration of auditory-visual patterns,” IPO Annual Progress Report, vol. 34, pp. 94–102, 1999.

[2] N. Dixon and L. Spitz, “The detection of audiovisual desynchrony,” Perception, vol. 9, pp. 719–721, 1980.

[3] W. Fujisaki, S. Shimojo, M. Kashino, and S. Nishida, “Recalibration of audiovisual simultaneity,” Nature Neuroscience, vol. 7, no. 7, pp. 773–778, 2004.

[4] W. Fujisaki and S. Nishida, “Temporal frequency characteristics of synchronyasynchrony discrimination of audio-visual signals,” Exp Brain Res, vol. 166, pp. 455– 464, 2005.

[5] T. S. Andersen, K. Tiippana, and M. Sams, “Factors influencing audiovisual fission and fusion illusions,” Cognitive Brain Res, vol. 21, pp. 301–308, 2004.

[6] Spence, C. and Driver, J. “A new approach to the design of multimodal warning signals” In Harris, D., editor, *Engineering psychology and cognitive ergonomics*, volume 4, pages 455-461. Ashgate Publishing, Hampshire, 1999.

[7] H. B. Barlowe, “Temporal and Spatial Summation in Human Vision at Different Background Intensities”, Journal of Physiology, pp. 337-350, vol. 141, no. 2, 1958.

[8] B. J. Moore. “An introduction to the Psychology of Hearing” 5th edition, Academic Press, California, 2003.