# The Perception of Cross-Modal Simultaneity

Daniel J. Levitin[†º], Karon MacLean[‡º], Max Mathews[?º] and Lonny Chu[?º]

†McGill University, Montreal, Quebec, Canada H3A 1B1
‡University of British Columbia, Vancouver, Canada V6T 1Z4
?CCRMA, Stanford University, Stanford, California USA 94306
and ºInterval Research, Palo Alto, California, USA 94306

**ABSTRACT**

One of the oldest questions in experimental psychology concerns perception of simultaneous events, particularly when input arrives through different sensory channels (sight / sound or touch / sound). How far apart in time must two events be to be perceived as sequential? This paper reports preliminary data from a cross-modal simultaneity task designed for ecological validity. Results indicate a smaller threshold for successiveness than that found in previous experiments, which used more artificial tasks. The present findings are relevant to theories of time, order and perception.

**Keywords**

Simultaneity, synchrony, cross-modal perception.

**INTRODUCTION**

An unsolved problem in cognitive science concerns the perception of simultaneous events, particularly when the information impinging on the sensory receptors comes from two different sensory modalities. For example, an event in the external world may give rise to both visual and auditory signals that may or may not be received at the sensory receptors of a human or a machine at the same time. How does the information processor decide if the sensory inputs were simultaneous?

The problem is fundamentally important for object identification because simultaneity is one of the most powerful cues available for determining whether two events define a single or multiple objects [3]. It is unsolved at both the macro and micro level. Neuroscientists know that the senses require varying amounts of time to process input [15], yet they do not know precisely how or where the brain integrates information to yield simultaneity judgments. Cognitive psychologists do not know how accurate human simultaneity judgments are; we do not have good models of what processes are involved, nor what the sources of variance in such judgments might be ([8] and others address the modeling problem).

The problem has practical relevance for those who design multimodal user interfaces (e.g. virtual reality systems, computer games) and seek experiential realism. Today's serial computational architectures preclude perfectly synchronous event presentation, and both Wintel and non-realtime Unix operating systems introduce additional latencies in their background operations. The simultaneity problem is also relevant to the entire range of experimental psychology and cognitive science. This includes sensation, perception, attention, psycholinguistics, music perception, neuroscience and the emerging field of anticipation and ordered systems. To accurately resolve the simultaneity problem, information processors (both human and machine) need to invoke complex mechanisms for anticipation, comparison, feedback and recursion.

In this paper, we will review the history of the problem, present an experiment performed in our psychophysical laboratory, and suggest anticipatory mechanisms to account for the phenomena observed.

**HISTORY**

In 1796, Maskelyne was the astronomer royal at the Greenwich Observatory. Stellar transits were then observed using the "eye and ear method" developed by Bradley (Figure 1). At the right level of magnification, a particular star took roughly one second to travel across the telescope's eyepiece; the astronomer would follow the star's position and noted its position relative to the sound of successive tic-tocs on a clock in the observatory. This involved a judgment about the simultaneity of a visual event (the star's location) in conjunction with an auditory event (the sound of the clock), and was believed accurate within a few hundred milliseconds.

The recording of stellar transits became known as the Greenwich Observatory problem after Maskelyne discovered that the observations of his assistant differed from his own by up to 800 milliseconds. He dismissed the assistant, but similar discrepancies reported elsewhere led to theories that time perception and simultaneity

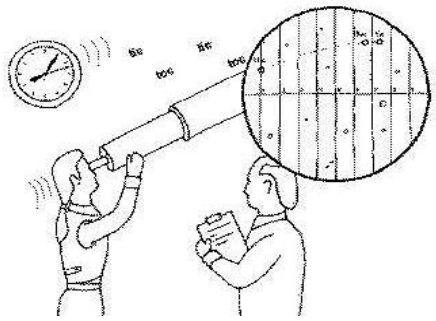

Figure 1: Bradley's "eye and ear" method required comparison of sensory inputs from the eye and the ear.

judgments varied by individual. In fact, Fechner and Wundt launched the fields of psychophysics and experimental psychology with studies of simultaneity perception [2].

One clue came from Arago's experiments in 1843. The variability was greatly reduced by using two observers: one watched and called out when the star crossed a gridline, and the other compared the call with the sound of the clock. Arago's method called for an *intra*modal judgment of sound vs. sound. Wundt ([2]) calculated thresholds for intramodal simultaneity perception, and found them to be 2 ms for sound; 27 ms for touch; and 43 ms for sight; these measure still stand today.

## PRIOR RESEARCH FINDINGS

Past investigation of intermodal asynchrony has focused audio/video. There are two general findings. (1) There appear to be large individual differences in perception thresholds. (2) Thresholds are asymmetric: people are more likely to perceive events as synchronous when the audio *precedes* the video than vice versa. This might be explained by human evolution in a world where sound travels more slowly than light. We can begin to quantify these thresholds based on six recent studies, the findings of which are arranged in a time line in Figure 2.

Dixon & Spitz [5] found a difference in thresholds for speech and non-speech stimuli. Subjects watched a video of hammering or a person speaking English words, while an audio track led and lagged the video. Subjects noticed asynchrony in the non-speech stimuli when the sound was 75 ms early or 175 ms late; but for speech, only at {-130, 250 ms}. The authors attribute this difference to our evolutionary experience: certain consonants (such as the English /p/) require the speaker to position the lips before sound comes out. Reinforcing this, McGrath and Summerfield [12] had professional lip-readers watch a video with varied audio asynchrony. Asynchrony threshold was taken as the point where the subjects' performance dropped significantly {-65 ms, 140 ms}.

Other studies averaged response to sound leading and lagging visual stimuli (absolute values shown Figure 2): 90 ms (Allan & Kristofferson [1]), 100 ms (Ganz [6]).

Jaskowski [9] asked subjects to make a three alternative forced-choice judgment about whether a target event was

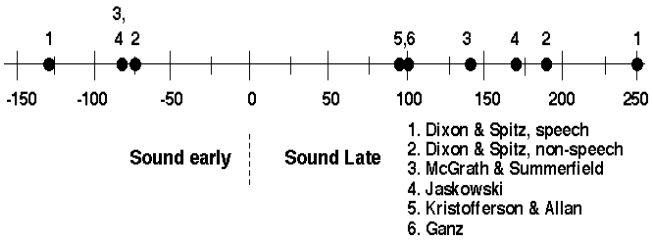


Figure 2: Recent estimates of the auditory-visual simultaneity threshold.

before, after, or simultaneous with a reference event, and reported a threshold of {-65 ms, 165 ms}. A problem with this study is that the before/after judgment amounted to a study of temporal order judgment (TOJ), and not a direct judgment of simultaneity. Although for most of this century researchers have used before/after experiments to infer simultaneity, it is now believed that separate neural processes are involved in judging simultaneity and successiveness [13].

Thus, accordingly to the most liberal estimates, asynchronous events are detected when a sound is approximately 75 ms early or 90 ms late. This is much slower than for *intra*modal judgments (Wundt). If the brain is a parallel processor, *inter*modal simultaneity judgments ought to take no longer than the longest intramodal judgment plus a modest interval for intermodal comparison. It is difficult to account 30 ms in such a comparison.

Why the low sensitivity? Perhaps subjects found the tasks strange or without ecological validity; human subjects can perform better in cognition-perception experiments if given the opportunity to perform in a natural setting with minimal interference [11]. Moreover, by relying on computer monitors, film and video, the cited studies faced significant latencies (as large as 40 ms for 16 mm film and no better than 25 ms for video). Our new experiment addressed both of these issues, and introduced the haptic modality.

## TESTING ASYNCHRONY

### Paradigm and Apparatus

We needed to systematically vary asynchrony between sensory cues in different channels; i.e. tightly control when the subject experienced the sound of a real-world event relative to its sight and feel. This meant erasing the event's real sound and reproducing it at a predefined interval before, with or after the visual/haptic event.

Striking a pad with a mallet proved amenable to such manipulation. We began with Matthews' Radio Baton ([14], Figure 3), a platform (the drum surface) with capacitive proximity sensing circuitry on the vertical axis, and a wand that acts as an antenna. A DOS PC acquired analog wand position at approximately 10 kHz.

The subject wore Sony MDR-V6 headphones for external blocking and internal playback, and heard a mono digitized sample of a stick hitting a drum played by a Linn machine at 75 dB(A). The drum sample was triggered by the PC at various temporal offsets from haptic impact, chosen randomly between +-200 msec. We recorded actual time of the acoustic stimulus with a microphone (to 0.1 msec) to bypass trigger variability, and post-computed actual time of haptic impact (verified acoustically to within 0.1 msec) with a smoothed direction reversal algorithm, to overcome drift, noise and imprecision in the Baton's raw signal.

Figure 3: Experimental set-up. A blindfolded Actor strikes while hearing the impact through headphones, offset from actual impact time. An Observer watches from an isolated room, hearing the same sound.

We estimated impending impact time by tracking the wand's velocity and acceleration, and delivered the acoustic stimulus within a few milliseconds of its target. The accuracy of this scheme increased as the desired acoustic precedence decreased and then became negative; but, having precise records of *actual* relative acoustic and haptic events, and given sufficient repetition, an adequate distribution was achieved.

### Subjects

Eight subjects (4M/4F, 20-40 years) were recruited from Stanford University via posted notices. Each completed three 2-hour sessions for $20/hour.

### Experiment Procedure

Subjects were tested in yoked pairs. One was randomly designated the "actor" and one the "observer;" roles were switched after each 90-trial block. Prior to each block, subjects practiced their roles in 10 practice trials. Each completed 720 trials in three days. The actor wore headphones and a blindfold and hit the drum surface with the wand during each trial; thus, the actor received information about the event through touch and sound. The observer stood in a sound isolated room behind double glass 2 meters away, and received the same asynchronous auditory signal as the actor through headphones (Figure 3). Thus the observer experienced the event through sight and sound.

After each trial, the subject said if the sound occurred at the "same" or a "different" time (a 2-alternative, forced choice in touch for actors, sight for observers), and then made a 3-point confidence rating ("not at all sure," "somewhat sure," and "very sure"). An experimenter in each room recorded the subjects' responses.

## RESULTS

A simple and intuitive way to consider the data is to plot the percentage of the time subjects judged the stimuli to be synchronous as a function of the sound-contact asynchrony (indicated by responses of "same"). If we consider the 75% line to indicate subjects' detection threshold (i.e., the asynchrony beyond which fewer than 75% of responses incorrectly judge simultaneous), we find that actors detected asynchronies at -25 and +42 msec, and observers detected asynchrony at -41 and 45 msecs (Figure 4). These figures are not adjusted for response bias and do employ subjects response ratings (a signal detection analysis will be reported in the future). However, these findings suggest a substantially smaller simultaneity window than the most liberal earlier estimates of {-75 ms, 90 ms}. The new figures are not far off in fact from Wundt's intramodal simultaneity threshold estimates. If we take the best case - in which subjects are most sensitive and judgi fsng sounds that are early - the current subjects are performing an intermodal task at rates approaching Wundt's intramodal findings. This suggests that the work of the comparator module is being accomplished rather quickly as one might expect - perhaps as quickly as 1 or 2 ms.

## DISCUSSION

Understanding human simultaneity detection logically involves anticipatory mechanisms. These are posited to exist in most (if not all) living organisms as a response to environmental contingencies [10], and as a solution to particular perceptual challenges [7]: try to predict future situations, and begin to adapt in advance. Since sensory event processing time varies greatly among sensory channels, the brain may have evolved a strategy invoking anticipation. By this we mean that the arrival of sensory input through one sensory channel causes a neural module to prepare for (that is, to *predict*) the arrival of related information through a different sensory channel.

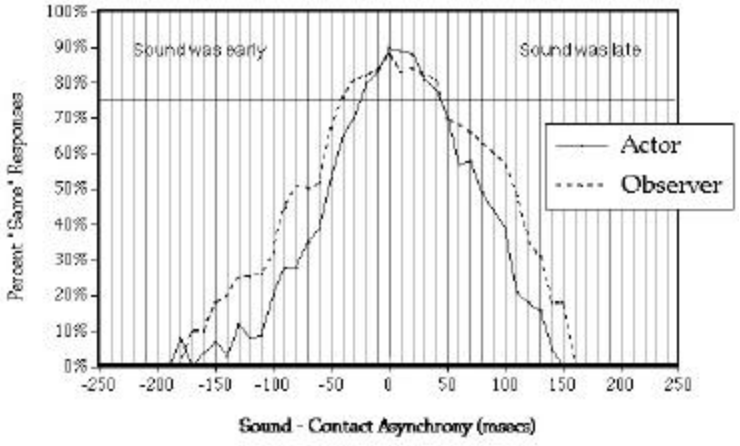


Figure 4: New data: judgments (uncorrected for response bias) of simultaneity when audio was early or late relative to tactile (solid line) or visual (broken) input. Curves indicate fraction of "same" responses out of all trials at that asynchrony, binned at 10 msec. More data was taken at small asynchronies. (8 subjects, 11,172 trials)

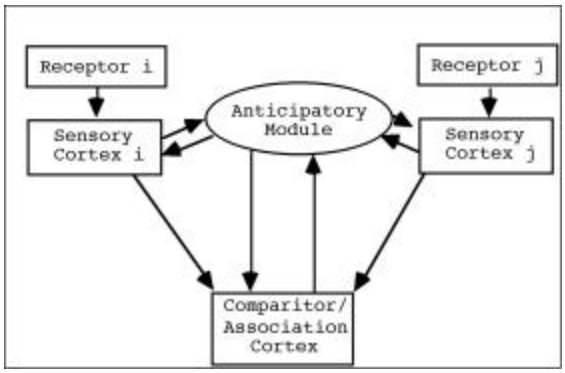


Figure 5: How an anticipatory module in the brain might coordinate simultaneity judgments with a comparator in the association cortex.

For example, after seeing a large boulder fall off a cliff in the distance, the brain (through millions of years of evolutionary experience with similar events) anticipates that the corresponding sound will follow shortly. Figure 5 illustrates one way this might work.

After sensory input is received at a given sensory receptor *i*, it is passed on to sensory cortex, exciting an anticipatory module. This module lowers the activation threshold in cortical areas associated with other senses, creating expectation that new information will soon arrive. The module has both feedback and feedforward connections to monitor new activity, and to interface with a "comparator" (probably instantiated in the superior colliculus) to decide whether signals that arrive at the comparator describe different sensory aspects of the same real-world event. Inputs may receive a time stamp as they reach the receptors, so that regardless of cortical processing time, the brain can subsequently determine item arrival times with respect to each other and an external world clock; but it is not yet understood how [4].

## CONCLUSIONS

A paradigm with greater ecological validity produced a smaller simultaneity threshold for input from two different sensory modalities than was previously found, by as much as 30 msec in either direction. However, the window within which people judge two asynchronous events to be the same not zero. This has implications for designers and computer scientists creating virtual reality environments and multi-media systems. Popular computer operating systems do not provide precise control over event scheduling and timing; awareness of end-user latency will allow designers to spend their effort more efficiently, providing precise synchronization only where needed.

A successful solution to the question of simultaneity perception may involve anticipatory and comparative mechanisms to accurately resolve information processed in a parallel and asynchronous fashion in the brain.

## ACKNOWLEDGMENTS

This research was supported by Interval Research Corporation. The authors thank Oliver Bayley, Jennifer Orton and Helen Shwe for their help.

## REFERENCES

[1] L. G. Allan and A. B. Kristofferson, "Successiveness discrimination: Two models," *Perception & Psychophysics*, vol. 15:1, pp. 37-46, 1974.

[2] E. Boring, *A History of Experimental Psychology*. New York: Pendragon, 1923.

[3] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.

[4] D. C. Dennett, *Consciousness explained*. Boston: Little, Brown & Company, 1991.

[5] N. F. Dixon and L. Spitz, "The detection of auditory visual desynchrony," *Perception*, vol. 9:6, 1980.

[6] L. Ganz, *Journal of Experimental Psychology: Animal Behavior Processes*, vol. 33, 1975.

[7] H. M. Gross, A. Heinze, et al, "Generative character of perception: A neural architecture for sensorimotor anticipation.," *Neural Networks*, vol. 12:7-8, 1999.

[8] R. B. Ivry and R. E. Hazeltine, "Perception and production of temporal intervals across a range of durations." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 21:1, 1995.

[9] P. Jaskowski, "Simple reaction time and perception of temporal order: Dissociations and hypotheses," *Perceptual and Motor Skills*, vol. 82:3, Pt. 1, 1996.

[10] J. A. Kelso, "Anticipatory dynamical systems, intrinsic pattern dynamics and skill learning," *Human Movement Science*, vol. 10:1, pp. 93-111, 1991.

[11] D. J. Levitin, "Absolute memory for musical pitch: Evidence from the production of learned melodies," *Perception and Psychophysics*, vol. 56:4, 1994.

[12] M. McGrath and Q. Summerfield, "Intermodal timing relations and audio-visual speech recognition by normal-hearing adults," *Journal of the Acoustical Society of America*, vol. 77, 1985.

[13] L. Mitrani, S. Shekerdijiiski, and N. Yakimoff, "Mechanisms and Asymmetries in Visual Perception of Simultaneity and Temporal Order," *Biological Cybernetics*, vol. 54, 1986.

[14] W. Putnam and R. B. Knapp, "The Radio Baton, in *Input/Data Acquisition System Design for Human Computer Interfacing*," 1996.

[15] E. Zeki, *A Vision of the Brain*. New York: Oxford University Press, 1993.