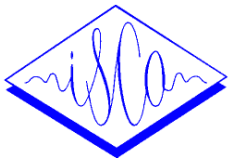

AVSP 2003 – International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4–7, 2003

Audiovisual Asynchrony Detection for Speech and Nonspeech Signals

Brianna L. Conrey
David B. Pisoni

Department of Psychology, Indiana University
Bloomington, Indiana, USA 47405-1301

bconrey|pisoni@indiana.edu

Abstract

This study investigated the "intersensory temporal synchrony window" [1] for audiovisual (AV) signals. A speeded asynchrony detection task was used to measure each participant's temporal synchrony window for speech and nonspeech signals over an 800-ms range of AV asynchronies. Across three sets of stimuli, the video-leading threshold for asynchrony detection was larger than the audio-leading threshold, replicating previous findings reported in the literature. Although the audio-leading threshold did not differ for any of the stimulus sets, the video-leading threshold was significantly larger for the point-light display (PLD) condition than for either the full-face (FF) or nonspeech (NS) conditions. In addition, a small but reliable phonotactic effect of visual intelligibility was found for the FF condition. High visual intelligibility words produced larger video-leading thresholds than low visual intelligibility words. Relationships with recent neurophysiological data on multisensory enhancement and convergence are discussed.

1. Introduction

Temporal correlation of auditory and visual stimuli is known to be critical in producing audiovisual (AV) enhancement [2]. Investigations of the temporal limitations of AV enhancement are therefore important to our theoretical understanding of AV processing. Desynchronizing auditory and visual inputs is one tool that can be used to conduct such investigations.

Previous studies have examined thresholds of AV asynchronies at which either the asynchrony can be detected or no AV gain is experienced in integration. These studies have typically attempted to provide estimates of an "intersensory temporal synchrony window" [1]. For example, Bushara and colleagues [3] have reported that AV asynchronies are not detected above chance levels between A50V ms (audio leading video by 50 ms) and V100A ms (video leading audio by 100 ms). Several other studies have reported poor asynchrony detection for nonspeech AV signals between A80V ms and V160A ms [1, 4, 5]. For speech stimuli, Grant, van Wassenhove, and Poeppel [6] found that AV asynchronies were not reliably detected between A35V ms and V225A ms.

Dixon and Spitz [4] describe the only previous comparison of asynchrony detection between speech and nonspeech signals that we have found in the literature. They reported that asynchronies in a film of a hammer hitting a nail could not be detected at auditory delays between -74.8 and 187.5 ms, whereas asynchronies in a film of a man reading prose could not be detected at auditory delays between -131 and 257.9 ms. In another study, Munhall and colleagues [7] reported similar thresholds for AV speech in an investigation of the effect of temporal asynchrony on the McGurk effect. They found that participants experienced the McGurk effect with asynchronies between A60V ms and V240A ms, indicating integration of auditory and visual information within that time window. Also, more recently Grant and Greenberg [8] reported increases in intelligibility for AV sentences compared with their auditory- and visual-alone counterparts between A40V ms and V160A to V200A ms. Several other studies have found that auditory delays of 200 to 250 ms are the upper limit for integration of AV speech [4, 9-11].

The present study had two objectives. First, we wanted to obtain fine-grained measures of the temporal synchrony windows for three types of AV stimuli using the same set of participants. Second, we wanted to assess the effects of phonotactic context on the characteristics of these windows. To accomplish these two goals, we used a speeded detection task methodology.

2. Methods

2.1. Participants

Participants were 15 undergraduate students at Indiana University (5 male and 10 female, mean age of 19.33 years). Eight received partial credit in an introductory psychology course for their participation; the other seven were paid $10 for their services. All participants were right-handed, monolingual native speakers of American English with no history of hearing or speech disorders and normal or corrected-to-normal vision. The experiment took approximately one hour to complete.

2.2. Procedures

The experimental design consisted of three AV conditions: full-face video (FF), point-light display video (PLD), and a nonspeech control condition (NS). The visual stimuli were presented on an Apple Macintosh G4 computer. Auditory stimuli were presented over Beyer Dynamic DT headphones at 77 dB SPL. PsyScope version 1.5.2 was used for stimulus presentation. All participants were tested on each of the three conditions, which were presented in the same order for all participants.

The NS stimuli were based on those used in an experiment by Bushara and colleagues [3]. The visual display consisted of a red circle approximately 4 cm in diameter, and the auditory signal was a 2000-Hz tone. Both the visual and auditory stimuli were 100 ms in duration.

For the FF condition, 10 familiar English words were chosen from the Hoosier Audiovisual Multitalker Database [12, 13], a previously recorded database of digitized AV movies consisting of single talkers speaking isolated monosyllabic words. All 10 words chosen for the present study were spoken by the same female talker who had been determined to be the most intelligible of the eight talkers in the database [14]. Because phonotactic context effects were of interest, five of the words chosen had high visual-only intelligibility scores and five had low visual-only intelligibility scores based on an earlier experiment [14].

The PLD condition also had 10 words, similarly chosen from the most intelligible female talker from a previously recorded audiovisual database of isolated single-syllable English words [15]. For the PLD movies in this database, the talker had 30 glow-in-the-dark dots glued to the lower half of her face, including her cheeks, jaw, chin, lips, upper and lower teeth, and tongue tip [16]. The video recordings were made so that only the movement of the dots was visible. Visual-only intelligibility data did not exist for whole words in the PLD database, but viseme-confusability matrices for the PLD movies obtained from one talker [17] were used to choose five words that were predicted to have high visual intelligibility and five others that were predicted to have low visual intelligibility.

We wanted to obtain a full subjective temporal synchrony window for each participant, so we measured AV asynchrony detection using the method of constant stimuli over a wide range from A300V ms to V500A ms. These temporal asynchronies were chosen based on piloting in our lab. The videos used were recorded at a rate of 30 frames per second, so each stimulus could differ by 33 ms. This resulted in 25 asynchrony levels, nine with audio leading, one synchronous, and 15 with video leading.

The test stimuli were created using Final Cut Pro 3. For the asynchronous speech stimuli, the portions of the audio and video tracks that did not overlap with each other were edited from the stimulus video. Thus, the participants could not rely on global temporal cues such as the audio track coming on while the screen was blank to determine if the video was synchronous, but instead they had to make their judgments about synchrony based on whether information was temporally "matched" across the auditory and visual modalities.

The stimuli were presented in a speeded single-interval detection task. On each trial, the participants were asked to respond as quickly as possible using a button box if the AV stimulus was "in sync" or "not in sync." They received instructions at the beginning of each condition and were presented with several examples of synchronous (0 auditory delay) and asynchronous (A300V ms and V300A ms) movies before the FF and PLD conditions. Each condition consisted of 250 randomized trials, 10 for each of the 25 asynchrony levels. In the NS condition, all trials used the same visual and auditory stimuli, described earlier. In the FF and PLD conditions, each of the 10 words was presented once at each asynchrony level. At the onset of each trial, a fixation mark ("+") flashed on the screen for 200 ms and was followed by 300 ms of blank screen before the test stimulus appeared.

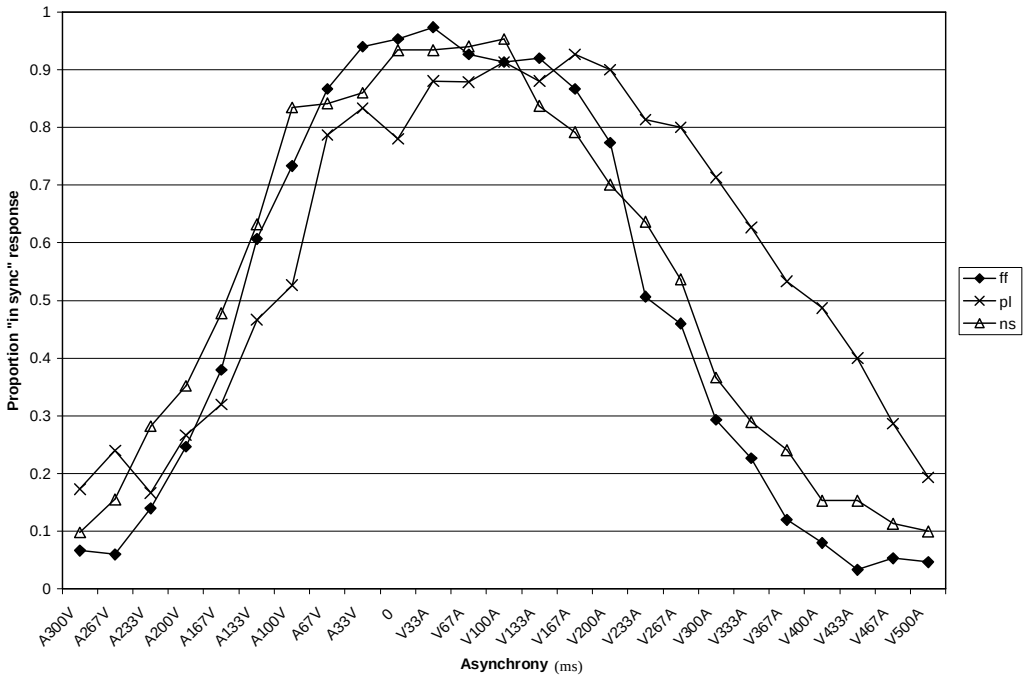

Figure 1: Average proportion "in sync" responses

Table 1: MPS and threshold averages (ms)

Condition	MPS		Threshold			
			Audio-leading		Video-leading	
	M	*SD*	*M*	*SD*	*M*	*SD*
NS	45.82	40.40	157.68	47.05	249.32	64.81
FF, average	46.73	29.00	133.54	54.86	226.99	59.99
FF, high VI	57.09	29.07	131.93	53.44	246.11	72.26
FF, low VI	37.79	34.74	127.87	78.18	203.46	71.61
PLD, average	113.27	35.36	157.92	93.05	384.46	122.50
PLD, high VI	108.74	37.44	152.08	95.36	369.56	128.95
PLD, low VI	116.90	39.43	176.12	130.72	409.92	156.57

3. Results

The proportion of "in sync" responses was determined at each asynchrony level for each participant. Figure 1 shows the average "in sync" responses for the FF, PLD, and NS conditions at each asynchrony level. Figures 2 and 3 show the average "in sync" responses for high and low visual intelligibility words in the FF and PLD conditions, respectively. It should be noted that although the average PLD curves do not reach 100% "in sync" response, all but one of the individual participants did respond "in sync" for 100% of at least one asynchrony level. That same participant was also the only one who failed to respond "in sync" for 100% of at least one asynchrony level in the NS condition.

The data were transformed so that the proportion correct "in sync" response for audio-leading-video levels (including the 0 asynchrony level) was in the range [0, 0.5] and the proportion correct for video-leading-audio was in the range [0.5, 1.0]. Following methods described by Spence and colleagues [18], each participant's data was then fitted with a logistic sigmoid curve of the form:

$$P(\text{"}in.sync\text{"}) = \frac{1}{1+e^{-slope*(asynchrony.level-MPS)}} \qquad (1)$$

where MPS is the "mean point of synchrony." The audio-leading and video-leading thresholds for synchrony were calculated as the 25% and 75% points on the curve, given by *MPS* ± (ln 3 / *slope*). Means and standard deviations for the MPS and thresholds in the three conditions are summarized in Table 1.

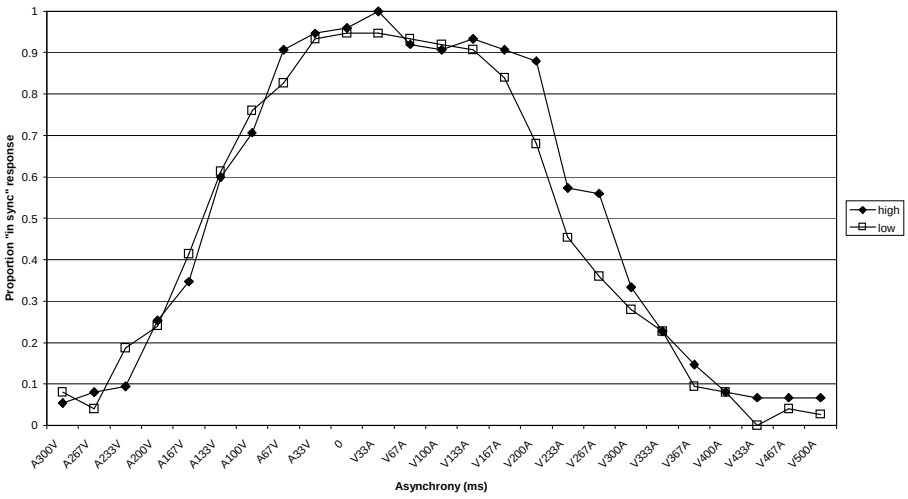

Figure 2: High and low visual intelligibility "in sync" responses, FF condition

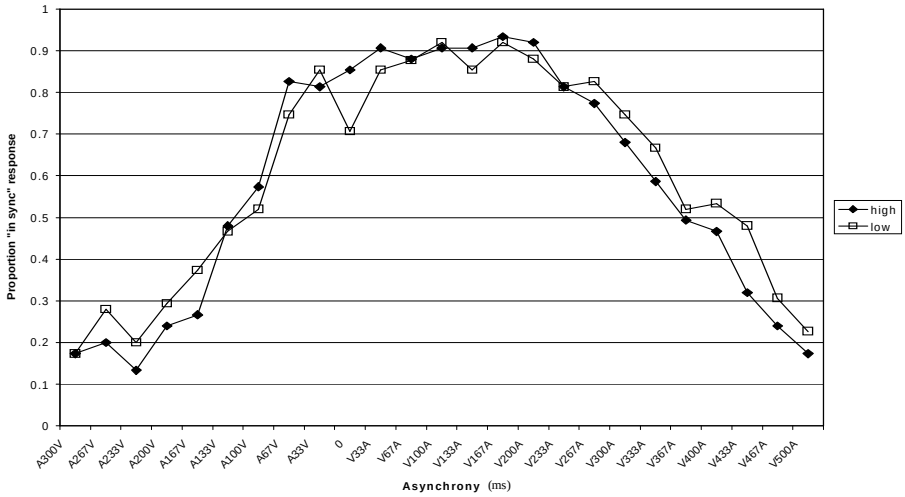

Figure 3: High and low visual intelligibility "in sync" responses, PLD condition

3.1. MPS Results

The MPS was significantly larger than 0 in all three conditions of the experiment (FF: $t(14) = 6.240$; PLD: $t(14) = 12.445$; NS: $t(14) = 4.393$; all p's $\leq .001$). A one-way repeated measures ANOVA revealed a significant effect of condition (NS, FF, PLD) on MPS ($F(2, 28) = 27.144$, $p < .001$). Paired t-tests using a Bonferonni correction for multiple comparisons ($\alpha = .05/7 = .007$) indicated that the NS and FF conditions did not differ significantly from each other on MPS ($t(14) = -.078$, $p > .05$). However, the PLD condition had a MPS that was significantly larger than either the NS ($t(14) = 5.730$, $p < .001$) or the FF ($t(14) = 8.732$, $p < .001$) conditions.

A two-way repeated measures ANOVA on visual intelligibility in the FF and PLD AV speech conditions revealed an additional interaction of Condition x Visual Intelligibility ($F(1, 14) = 6.008$, $p = .028$) for the MPS. For the FF condition, the high visual intelligibility words had a MPS that was on average 19.3 ms larger than the MPS for low visual intelligibility words. This difference, although small, was significant using a paired t-test ($t(1, 14) = 3.243$, $p = .006$). Twelve of the 15 participants showed the visual intelligibility effect. For the remaining three participants, the FF high visual intelligibility MPS's were 4.44, 7.89, and 17.23 ms lower than the low visual intelligibility MPS's. The PLD words showed the opposite result for MPS. The high visual intelligibility words had a smaller MPS than the low visual intelligibility words; however, this difference was not significant ($t(14) = -1.147$, $p > .007$).

Neither high nor low visual intelligibility MPS's in the FF condition differed significantly from the MPS in the NS condition ($t(14) = .998$, $t(14) = -.634$, respectively; p's $< .001$).

3.2. AV Synchrony Threshold Results

3.2.1. Audio-leading Thresholds

For the audio-leading thresholds, a one-way repeated measures ANOVA showed no effect of condition ($F(2, 28) = 1.682$, $p = .204$). A two-way ANOVA with FF and PLD visual intelligibility showed no significant Condition x Visual Intelligibility interaction ($F(1, 14) < 1$). None of the paired t-test comparisons was significant.

3.2.2. Video-leading Thresholds

A one-way repeated measures ANOVA on the video-leading thresholds revealed a significant effect of condition ($F(2, 28) = 24.562$, $p < .001$). As for the MPS analysis, paired t-tests indicated that while the NS and FF video-leading thresholds were not significantly different ($t(14) = 1.187$, $p > .05$), both the NS and FF conditions had significantly smaller video-leading thresholds than the PLD conditions ($t(14) = -7.237$ and $t(14) = -4.393$, respectively; p's $\leq .001$).

Again, a two-way repeated measures ANOVA was performed to assess the effects of on visual intelligibility in the FF and PLD conditions. This analysis revealed a significant Condition x Visual Intelligibility interaction ($F(1, 14) = 4.623$, $p = .05$). Paired t-tests using a Bonferonni correction of $\alpha = .007$ established that high visual intelligibility words in the FF condition had a significantly larger video-leading threshold than low visual intelligibility words ($t(14) = 3.281$, $p = .005$). Neither high nor low visual intelligibility FF video-leading thresholds were significantly different from the NS condition, however ($t(14) = -.164$ and $t(14) = -2.345$, respectively; p's $> .007$).

4. Discussion

Our findings indicate that participants judged larger asynchrony levels as subjectively synchronous when the video led the audio than when the audio led the video.

Similar AV processing asymmetries have been observed in electrophysiological data [19-21]; behavioral studies using simple stimuli [1, 3, 22]; AV speech asynchrony detection tasks [4-6]; and AV speech integration tasks [7, 8]. Furthermore, the average intersensory synchrony window, which ranged from approximately A150V ms to V240A ms in the NS and FF conditions, was comparable to those observed in earlier studies of AV speech asynchrony detection and integration [4, 6, 7]. The video-leading threshold for the PLD condition is comparable to that reported by Grant and Seitz [23], although they did not include audio-leading stimuli in their experiment and used only hearing-impaired adults as participants. It is unclear why the nonspeech asynchrony detection thresholds differ from those reported previously [1, 3-5; but see 22], but further investigations are planned into how stimulus characteristics might influence thresholds for nonspeech signals.

The average MPS for the NS and FF conditions was around 45 ms, indicating that the likelihood of an "in sync" judgment was maximal when visual input led auditory input by approximately that time interval. Interestingly, Schroeder and Foxe [24] have reported that visual feedback reaches posterior auditory cortex at 50 ms poststimulus, while auditory feedforward input arrives at around 11 ms poststimulus. They suggest that input from visual feedback circuits into auditory areas may provide a substrate for early AV interactions in auditory processing. Several fMRI studies [25-28] and an EEG study using independent component analysis [29] have reported enhanced auditory association cortex activation during AV speech perception. Based on the estimates of Schroeder and Foxe [24], visual and auditory inputs could be expected to arrive simultaneously at posterior auditory cortex if the visual stimulus occurs approximately 40 ms before the auditory stimulus. Recent data from event-related potential studies in humans in fact suggest that the earliest audiovisual interaction in cortex can be detected over posterior cortex 40 to 46 ms after the presentation of an audiovisual stimulus [30-32], and a recent behavioral study of audiovisual processing indicated that optimal performance on several perceptual tasks occurred with auditory delays of between 50 and 100 ms [22].

In addition, it has been reported that multisensory enhancement is most likely to occur at the neuronal level when signals from multiple sensory modalities occur within 100 ms of each other [19]. Similar multisensory interactions have been observed for asynchronies of 100 ms or less at a behavioral level [33]. If an optimal level of AV enhancement indeed occurs at around V40A ms, then, assuming the 100-ms window for enhancement is centered around that point, we might expect to see evidence of multisensory enhancement between A170V ms and V230A ms for 100-ms-long signals. This intersensory synchrony window is in fact similar to the window estimates obtained with the 100-ms-long NS signals as well as with the FF signals.

The PLD temporal synchrony window was somewhat larger. It has been proposed that information relevant for AV speech processing is provided not only in the perioral region, but all over the face [34]. Because the PLD stimuli contained information from the perioral region only, it is possible that these highly impoverished stimuli may have required more cognitive processing. Another possibility is that since the PLD stimuli captured less detailed information than the FF stimuli, the potential window of multisensory enhancement was widened, an effect observed previously at the neuronal level for slow-moving visual stimuli when compared with their fast-moving counterparts [20].

One issue in investigations of multisensory processing has been whether behavioral judgments about multisensory stimuli reflect linguistic or nonlinguistic processes. The similarity of the NS and FF results suggests that at least for these two conditions our speeded asynchrony detection task captured processes that are common to both linguistic and nonlinguistic signals. However, the small but significant effect of visual intelligibility of the words in the FF condition may be indicative of some additional mandatory phonotactic processing of these patterns. It is interesting that this phonotactic effect appeared only when the video led the audio but not vice versa. Perhaps since auditory information is processed more quickly by the nervous system than visual information, visual intelligibility information does not have time to converge with audio-leading stimuli before the synchrony judgment is made. On the other hand, video-leading stimuli could provide additional priming cues for the upcoming auditory stimuli that could take effect before the auditory signals begin. A similar explanation can be proposed for the PLD condition, which also showed significant differences from the synchrony windows of the other conditions but only when the video signal led the audio signal. Further investigations of the interaction between phonotactic structure and the asymmetry of auditory and visual information are planned.

5. Acknowledgements

This research was supported by the NIH-NIDCD research grant R01 DC00111 to Indiana University. The first author was also supported by an Indiana University Chancellor's Fellowship and a National Science Foundation Graduate Research Fellowship.

The authors would like to thank Luis Hernandez for his technical expertise and Dr. Olaf Sporns for his advice and suggestions.

6. References

[1] Lewkowicz, D.J., *Perception of auditory-visual temporal synchrony in human infants.* Journal of Experimental Psychology: Human Perception and Performance, 1996. **22**: p. 1094-1106.

[2] Stein, B. and M.A. Meredith, *The merging of the senses*. 1993, Cambridge, MA: MIT Press. 211.

[3] Bushara, K.O., J. Grafman, and M. Hallett, *Neural correlates of auditory-visual stimulus onset asynchrony detection.* Journal of Neuroscience, 2001. **21**(1): p. 300-304.

[4] Dixon, N. and L. Spitz, *The detection of audiovisual desynchrony*. Perception, 1980. **9**: p. 719-721.

[5] McGrath, M. and Q. Summerfield, *Intermodal timing relations and audio-visual speech recognition by*

normal-hearing adults. Journal of the Acoustical Society of America, 1985. **77**(2): p. 678-684.

[6] Grant, K.W., V. van Wassenhove, and D. Poeppel. *Discrimination of auditory-visual synchrony*. in *AVSP 2003 International Conference on Auditory-Visual Speech Processing*. 2003.

[7] Munhall, K.G., et al., *Temporal constraints on the McGurk effect*. Perception & Psychophysics, 1996. **58**(3): p. 351-362.

[8] Grant, K.W. and S. Greenberg. *Speech intelligibility derived from asynchronous processing of auditory-visual information*. in *AVSP International Conference on Auditory-Visual Speech Processing*. 2001.

[9] Pandey, C.P., H. Kunov, and M.S. Abel, *Disruptive effects of auditory signal delay on speech perception with lip-reading.* The Journal of Auditory Research, 1986. **26**: p. 27-41.

[10] Massaro, D. and M. Cohen, *Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables.* Speech Communication, 1993. **13**: p. 127-134.

[11] Massaro, D., *Perceiving talking faces: From speech perception to a behavioral principle*. 1998, Cambridge, MA: MIT Press.

[12] Lachs, L. and L.R. Hérnandez, *Update: The Hoosier Audiovisual Multitalker Database*, in *Research on Spoken Language Processing Progress Report 22*. 1998, Speech Research Laboratory, Indiana University: Bloomington, IN. p. 377-388.

[13] Sheffert, S.M., L. Lachs, and L.R. Hérnandez, *The Hoosier audiovisual multitalker database*, in *Research on Spoken Language Processing no. 21*. 1996, Speech Research Laboratory, Indiana University: Bloomington, IN. p. 578-583.

[14] Lachs, L. and D.B. Pisoni, *Visual recognition of spoken words without audition.* under review.

[15] Lachs, L., *Vocal tract kinematics and crossmodal speech information*, in *Research on Spoken Language Processing Technical Report*. 2002, Indiana University: Bloomington. p. 1-90.

[16] Rosenblum, L.D. and H.M. Saldaña, *An audiovisual test of kinematic primitives for visual speech perception.* Journal of Experimental Psychology: Human Perception and Performance, 1996. **22**(2): p. 318-331.

[17] Bergeson, T.R., J.T. Reynolds, and D.B. Pisoni. *Perception of point light displays of speech by normal-hearing adults and deaf adults with cochlear implants*. in *AVSP 2003 International Conference on Auditory-Visual Speech Processing*. 2003.

[18] Spence, C., et al., *Multisensory temporal order judgments: When two locations are better than one.* Perception & Psychophysics, 2003. **65**(2): p. 318-328.

[19] King, A.J. and A.R. Palmer, *Integration of visual and auditory information in bimodal neurones in the guinea-pig superior colliculus.* Experimental Brain Research, 1985. **60**: p. 492-500.

[20] Meredith, M.A., J.W. Nemitz, and B.E. Stein, *Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors.* The Journal of Neuroscience, 1987. **7**(10): p. 3215-3229.

[21] Meredith, M.A., *On the neuronal basis for multisensory convergence: A brief overview.* Cognitive Brain Research, 2002. **14**: p. 31-40.

[22] Lewald, J. and R. Guski, *Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli*. Cognitive Brain Research, 2003. **16**: p. 468-478.

[23] Grant, K.W. and P.F. Seitz, *Measures of auditory-visual integration in nonsense syllables and sentences.* Journal of the Acoustical Society of America, 1998. **104**: p. 2438-2450.

[24] Schroeder, C.E. and J.J. Foxe, *The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex.* Cognitive Brain Research, 2002. **14**: p. 187-198.

[25] Calvert, G., et al., *Activation of auditory cortex during silent lipreading.* Science, 1997. **276**: p. 593-6.

[26] Calvert, G., et al., *Response amplification in sensory-specific cortices during crossmodal binding.* NeuroReport, 1999. **10**: p. 2629-2623.

[27] Calvert, G., R. Campbell, and M.J. Brammer, *Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex.* Current Biology, 2000. **10**: p. 649-657.

[28] Calvert, G. and R. Campbell, *Reading speech from still and moving faces: The neural substrates of visible speech.* Journal of Cognitive Neuroscience, 2003. **15**(1): p. 57-70.

[29] Callan, D.E., et al., *Multimodal contribution to speech perception revealed by independent component analysis: A single-sweep EEG case study.* Cognitive Brain Research, 2001. **10**: p. 349-353.

[30] Molholm, S., et al., *Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study*. Cognitive Brain Research, 2002. **14**: p. 115-128.

[31] Teder-Sälejärvi, W.A., et al., *An analysis of audio-visual crossmodal integration by means of event-related potential (ERP) recordings.* Cognitive Brain Research, 2002. **14**: p. 106-114.

[32] Giard, M.H. and F. Peronnet, *Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study.* Journal of Cognitive Neuroscience, 1999. **11**(5): p. 473-490.

[33] Shams, L., Y. Kamitani, and S. Shimojo, *Visual illusion induced by sound.* Cognitive Brain Research, 2002. **14**: p. 147-152.

[34] Vatikiotis-Bateson, E., et al., *Eye movement of perceivers during audiovisual speech perception.* Perception & Psychophysics, 1998. **60**(6): p. 926-940.