

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 26 (2003-2004)
Indiana University

**Detection of Auditory-Visual Asynchrony
in Speech and Nonspeech Signals¹**

Brianna L. Conrey and David B. Pisoni²

*Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405*

¹ This work was supported by NIH NIDCD R01 Research Grant DC00111 and NIH NIDCD T32 Training Grant to Indiana University. The first author was also supported by an Indiana University Chancellor's Fellowship and an NSF Graduate Research Fellowship. The authors would like to thank Luis Hernandez for technical support and Sara Phillips and Angelique Horace for help with data collection. We also thank Jim Craig, Jason Gold, Olaf Sporns, and Bryan Donaldson for advice on various aspects of the data analysis. We gratefully acknowledge the audience at the 2003 Audio-Visual Speech Processing workshop for their comments and valuable suggestions on a preliminary version of this work.

² Also DeVault Otologic Research Laboratory, Department of Otolaryngology—Head and Neck Surgery, Indiana University School of Medicine, Indianapolis, IN.

Detection of Auditory-Visual Asynchrony in Speech and Nonspeech Signals

Abstract. Two experiments were conducted to examine the temporal limitations on the detection of asynchrony in auditory-visual (AV) signals. Each participant made asynchrony judgments about speech and nonspeech signals presented over an 800-ms range of AV onset asynchronies. Consistent with previous findings, all conditions revealed a wide window of several hundred milliseconds over which AV signals were judged to be synchronous. In addition, signals in which the visual component led the auditory component were more likely to be judged as synchronous than signals in which the auditory component led the visual component. In contrast with earlier reports (Dixon & Spitz, 1980; McGrath & Summerfield, 1985), the present results also demonstrated a similar AV synchrony window for speech and nonspeech signals, even when these signals were matched for duration. Visual phonetic characteristics of the speech signals, however, did influence the size and shape of the AV synchrony window. Finally, the onset of the relevant aspects of the stimulus, rather than the duration or offset, was most important for asynchrony judgments for both speech and nonspeech signals. Relationships with recent data on neural mechanisms of multisensory enhancement and convergence are discussed.

Introduction

The temporal relationships among multimodal stimuli are critical in determining how and whether the stimuli will interact during perceptual processing (Meredith, 2002; Stein & Meredith, 1993). A fundamental question that arises in the study of multisensory perception is the window of time over which multisensory interactions can occur—that is, how close in time do stimuli in two or more sensory modalities need to be in order to interact in processing? One way of studying this issue is to measure the perception of multimodal events that are desynchronized to varying degrees. The present study investigated the detection of auditory-visual (AV) asynchrony for both speech and nonspeech signals.

Because earlier research has demonstrated interactions between audition and vision in human speech perception (Calvert et al., 1997; McGurk & MacDonald, 1976; Sumbly & Pollack, 1954), we studied AV asynchrony detection in meaningful speech signals. We also compared detection of AV asynchronies in speech with semantically meaningless nonspeech signals in order to gain insights into the cognitive and neural processes that are involved in this type of crossmodal task.

Literature Review

Several previous studies have examined AV synchrony detection for speech and nonspeech signals and explored the effects of AV asynchrony on McGurk illusions and speech intelligibility scores. In an early study, Dixon and Spitz (1980) presented adult subjects with films of two AV events, a man reading prose or a hammer hitting a peg. During the experiment, the auditory and visual tracks of the film became gradually out of sync. On half of the trials, the auditory track led the visual track, and on the other half the visual track led the auditory track. Subjects were asked to respond when they noticed that the auditory and visual tracks had become asynchronous. On average, subjects could not reliably detect AV asynchrony for the speech film over a range from when the auditory signal led by 131.1 ms (A131.1V ms) to when the visual signal led by 257.9 ms (V257.9A ms). The “intersensory synchrony window” (Lewkowicz, 1996) for the hammer hitting the nail was much smaller, between A74.8V and V187.5A ms.

McGrath and Summerfield (1985) measured the intelligibility of AV sentences presented in auditory white noise over several visual-leading conditions that ranged up to V160A ms. They reported that although there was a significant effect of delay overall, subjects did not show a significant decrement in speech intelligibility performance until delays of V160A ms were reached. Although the better lipreaders showed a linear trend of decreased performance with increased delay, this pattern was not observed in poor or average lipreaders.

In an additional experiment, McGrath and Summerfield (1985) assessed AV asynchrony detection thresholds using a Lissajou interference pattern as the visual stimulus and a 120-Hz rise/fall gated triangular wave as the auditory stimulus; these patterns were thought to simulate a pair of lips opening to articulate a CV syllable. In an adaptive testing procedure using an XAB forced-choice task, they estimated the 70.7% detection thresholds for auditory-leading and visual-leading asynchronies using these nonspeech patterns. On average, these thresholds were at A78.5V and V137.8A ms. McGrath and Summerfield found a moderate but nonsignificant association between higher nonspeech thresholds and lower lipreading scores. The average auditory-leading and visual-leading thresholds were at A78.5V and V137.8A ms. These thresholds are very close to those reported earlier for nonspeech stimuli by Dixon and Spitz (1980) (A74.8V and V187.5 ms), and Lewkowicz for an adult control group in an infant study (1996) (A65V and V112A ms). They are also similar to the results reported by Bushara, Grafman, and Hallett (2001) for a pilot behavioral task they used in a PET study (A56V and V114A ms).

In addition to Dixon and Spitz (1980), one other study, carried out by Grant, van Wassenhove, and Poeppel (2003), examined discrimination of AV asynchrony in sentences. In a two-interval forced-choice adaptive procedure, they found asynchrony discrimination thresholds of approximately A35V and V160 ms for unfiltered speech and A35V and V225A ms for bandpass-filtered speech. Dixon and Spitz (1980) reported thresholds between A131.1V and V257.9A ms for their speech materials. However, they used a different behavioral task and computed their thresholds with a different procedure.

In addition to the studies of AV synchrony detection and discrimination, several other studies have examined the perception of CV syllables as well as the McGurk illusion over a range of asynchrony levels. In general, AV temporal thresholds for the McGurk illusion are similar to those found in studies of AV synchrony detection and discrimination. For example, Massaro and Cohen (1993) measured AV perception of the syllables /ba/ and /ga/ at asynchronies of V200A, V100A, 0, A100V, and V200A ms. They reported that congruent AV presentations resulted in high accuracy of identification regardless of asynchrony level; however, with visual /ba/ and auditory /da/, /bda/ judgments increased and /da/ judgments decreased with increases in visual lead. In a second experiment, they reported crossmodal influences of the vowels /i/ and /u/ over the same range of asynchronies.

Massaro, Cohen, and Smeele (1996) assessed AV integration for CV syllables over 13 asynchrony levels, from A533V to V533A ms, and concluded that AV integration was not significantly disrupted for asynchronies of up to about A250V or V250A ms. Another study by Munhall, Gribble, Sacco, and Ward (1996), reported that subjects reliably displayed the McGurk effect for visual /aga/ paired with auditory /aba/ for asynchronies between A60V and V240A ms, but did not display the effect reliably when visual /igi/ was paired with auditory /aba/. Similarly, van Wassenhove, Grant, and Poeppel (2002) reported fusion /ta/ responses for visual /pa/ paired with auditory /ka/ between A50V and V200A ms.

Another group of studies explored the effects of AV asynchrony on speech intelligibility of sentences. Pandey, Kunov, and Abel (1986) measured sentence intelligibility when the auditory signal was delayed by 0, 60, 120, 180, 240, and 300 ms, and presented at SNRs of 0 and -10 dB. They reported

that AV speech intelligibility was not significantly affected compared with auditory-alone presentation at asynchronies up to V240A ms for the 0 dB SNR, but they did observe a significant decrease in intelligibility by V180A ms for the -10 dB SNR. They also tested a group of normal-hearing experienced lipreaders at auditory delays of 0, 80, 160, and 240 ms, using a SNR of -5 dB, and found that performance declined significantly by V160A ms.

Grant and Seitz (1998) reported that the intelligibility of AV sentences presented in auditory noise was unaffected for hearing-impaired adults until auditory delays of around 200 ms. In another study of integration in asynchronous AV sentences, Grant and Greenberg (2001) found a relatively constant benefit for bandpass-filtered auditory sentences presented audiovisually over asynchronies ranging from A40V to V160-200A ms.

Taken together, most of the behavioral studies—whether measuring detection, discrimination, syllable identification, or sentence intelligibility—have estimated that the synchrony window for AV signals covers a range of several hundred milliseconds. In addition, the studies have reported that this synchrony window is larger when the visual signal leads than when the auditory signal leads. Finally, the studies report a great deal of individual variability in AV asynchrony thresholds among subjects (cf. individual subject data in McGrath & Summerfield (1985), p. 683, Table I, for an example).

Previous studies also suggest that the size of the AV synchrony window may be a function of the specific stimuli used. For example, the overall congruity of the auditory and visual events may influence tasks involving asynchronous AV stimuli (Munhall et al., 1996). In addition, other studies suggest that AV asynchrony may be easier to detect in nonspeech than in speech stimuli. However, speech and nonspeech thresholds have only been estimated using the same task in one study (Dixon & Spitz, 1980), and the speech and nonspeech events chosen in that study were not comparable. The speech event—a man reading prose—contained continuous visual and auditory signals, whereas the nonspeech event—a hammer hitting a nail—was discrete in nature and differed in overall duration from the speech event.

In the present set of experiments, we were interested in how the properties of the auditory and visual signals might affect the AV synchrony window. One of our goals was to compare detection of AV synchrony for speech and nonspeech signals using the same subjects and statistical procedures as well as the same levels of asynchrony for both sets of signals. The speech stimuli were isolated spoken English words rather than samples of connected speech so as to make the speech stimuli consist of single events that were more comparable to the nonspeech stimulus, which was a static circle paired with a simple tone.

In addition, we wanted to examine whether context effects related to the visual properties of the speech might influence judgments of synchrony. To accomplish this, we manipulated the visual characteristics of the speech stimuli in two ways. First, we used one speech condition in which participants viewed point-light displays of a talker's face rather than a fully illuminated face, and second, we used words that had been judged to have either high or low visual-only intelligibility based on results from previous experiments (Bergeson, Reynolds, & Pisoni, 2003; Lachs, 1999; Lachs & Pisoni, in press-a, in press-b). By directly comparing asynchrony judgments of speech and nonspeech, and speech with varying levels of visual information available, we hoped to elucidate some of the variables that influence the size and shape of the AV synchrony window in order to ultimately link our behavioral findings with plausible neural mechanisms.

Experiment 1

In Experiment 1, we obtained synchrony judgments from each participant under three AV conditions: nonspeech signals (NS), full-face speech (FF), and point-light display speech (PLD). A wide range of AV asynchronies were studied, from A300V to V500A ms.

Methods

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 at the time of testing. The experiment took approximately one hour to complete.

Stimuli

The experimental design consisted of three AV conditions: full-face video (FF), point-light display video (PLD), and a nonspeech condition (NS). The NS stimuli were modeled after those used in a PET study by Bushara et al. (2001) that investigated the neural correlates of AV asynchrony processing for several asynchrony levels. The present study used a 4-cm diameter red circle paired with a 2000-Hz tone. As in the earlier Bushara et al. study, 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 (Lachs & Hernandez, 1998; Sheffert, Lachs, & Hernandez, 1996), which contains digitized AV movies consisting of single talkers speaking isolated monosyllabic words. The most intelligible of the eight talkers in the database was determined in a previous study, and auditory-only, visual-only, and audiovisual intelligibility data had been collected for all her utterances (Lachs, 1999; Lachs & Pisoni, in press-a, in press-b). In this study, all 10 FF words were spoken by this talker.

All of the FF words used for the present experiment had 100% speech intelligibility scores for both auditory-alone and AV presentation. In order to examine the possible effects of visual-only intelligibility on judgments of AV synchrony, five of the words chosen had high visual intelligibility (VI) and five had low VI. The low VI words all had 0% correct whole-word visual-only intelligibility; these words were back, give, pail, theme, and voice. The high VI words were doubt (10%), fall (80%), knot (40%), loan (30%), and reed (50%). The high VI words had the highest visual-only speech intelligibility scores among the words with 100% intelligibility scores for auditory-alone and AV presentation, and all of the high VI words had higher whole-word visual-only intelligibility scores than the talker's whole-word visual-only intelligibility average of 4.4% (standard error = 1.2%).

The PLD condition also used 10 words, spoken by the most intelligible female talker from a previously recorded audiovisual database of isolated single-syllable English words (Lachs, 2002; Lachs & Pisoni, in press-c). 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 (see (Rosenblum & Saldaña, 1996). The video recordings were made with a black background so that only the movement of the green glow-in-the-dark dots was visible. Whole-word visual-only speech intelligibility was close to 0% correct for all PLD words. However, viseme-confusability matrices for the PLD movies

obtained from the most intelligible talker (Bergeson et al., 2003) were used to choose five words that were predicted to have high VI and five others that were predicted to have low VI. The high VI words were boat, site, hope, mouse, and tile, and the low VI words were cod, gain, guide, reach, and thick.

The experimental stimuli used in this study were created using Final Cut Pro 3 (copyright 2003, Apple Computer, Inc.). In all cases, the visual and auditory stimuli were combined beforehand into precompiled movies rather than being assembled “on the fly” during the experiment by the computer. 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 movie. (The removed portions did not contain any speech sounds or active articulatory movements.) This was done so that the participants would be unable to rely on any global temporal cues such as the audio track coming on while the screen was blank to determine if the movie was synchronous. Instead, all participants had to make their judgments about synchrony based on whether the presented information was temporally matched across the auditory and visual modalities.

Previous research on AV asynchrony detection (Dixon & Spitz, 1980; Lewkowicz, 1996; Massaro & Cohen, 1993; Massaro et al., 1996; McGrath & Summerfield, 1985; Pandey et al., 1986) and pilot studies in our lab indicated that most normal-hearing young adult subjects should be able to judge AV stimuli as asynchronous with close to 100% accuracy when the auditory signal leads the visual signal by 300 ms (A300V ms) or less and when the visual signal leads the auditory signal by 500 ms (V500A ms) or less. The experimental stimuli used in this study covered this wide range of asynchronies. Because the videos used were recorded at a rate of 30 frames per second, each successive stimulus could differ by 33.33 ms. This resulted in 25 asynchrony levels covering a range of 800 ms, from A300V to V500A. Nine stimuli had auditory leads, one was synchronous, and 15 had visual leads.

Procedure

The visual stimuli were presented on an Apple Macintosh G4 computer. Auditory stimuli were presented over Beyer Dynamic DT headphones at 70 dB SPL. PsyScope version 1.5.2 (Cohen, MacWhinney, Flatt, & Provost, 1993) was used for stimulus presentation and response collection. All participants were tested in each of the three conditions, NS, FF, and PLD. The conditions were blocked and were always presented in the order NS, FF, and PLD.

The stimuli were presented in a single-interval asynchrony judgment task. On each trial, the participants were asked to judge whether the AV stimulus was synchronous or asynchronous (“in sync” or “not in sync”) and were encouraged to respond as quickly and as accurately as possible. Participants were instructed to press one button on a response box if the stimuli were synchronous and another if they were asynchronous. Response hand was counterbalanced across participants but kept constant for each participant on all three conditions of the experiment so as to minimize confusion about the instructions. Before beginning each condition, the participants received instructions and were presented with examples of synchronous and asynchronous movies.

Each of the three conditions 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, the red circle and the 3000-Hz tone described above. 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 computer screen for 200 ms and was followed by 300 ms of blank screen before the test stimulus was presented. The subject’s response cued the onset of the next trial.

Results

Throughout this report, we will refer to synchronous AV stimuli as the 0 condition, for 0-ms delay/lead. Because our figures represent auditory leads to the left side of 0 on the abscissa and visual leads to the right, “lower” will indicate further toward the auditory-leading side of the figure, and “higher” will indicate further toward the visual-leading side of the figure. Similarly, negative numbers will refer to the auditory signal leading the visual signal in time, and positive numbers to the visual signal leading the auditory signal.

The proportion of synchronous responses at each level of asynchrony was determined for each participant. The average proportions are plotted separately for the FF, PLD and NS conditions in Figure 1.

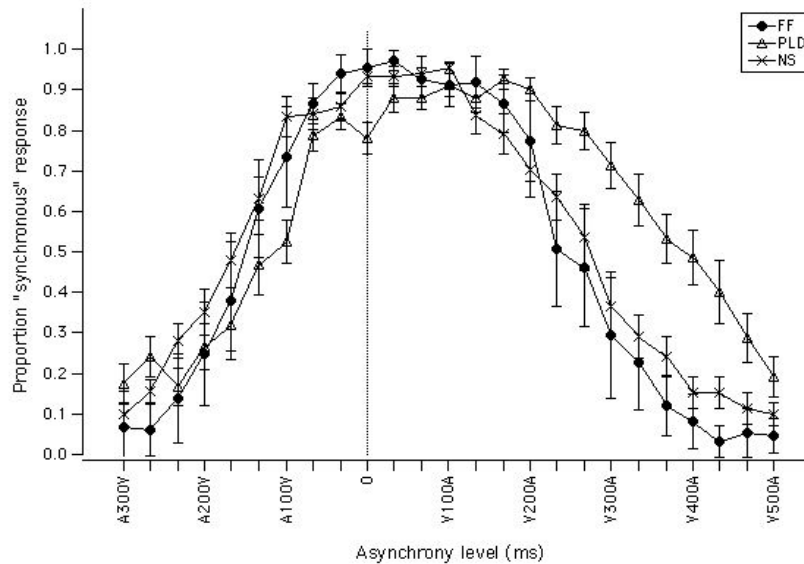


Figure 1. Average “in sync” response for all participants in Experiment 1, in FF, NS, and PLD conditions. The dotted vertical line is at 0-ms asynchrony.

In looking at the figure, two major features are apparent. First, the average range of asynchronies identified as synchronous was quite large, on the order of several hundred milliseconds. Second, this range was not centered at 0 ms, the synchronous condition, but was shifted to the right and centered on the visual-leading side of the continuum.

To quantify these findings, we fit each condition from the individual participants’ data with a Gaussian curve using Igor Pro 4.05A Carbon (copyright 1988-2002, WaveMetrics, Inc.). This resulted in a total of seven curves for each participant: FF, PLD, NS, and high and low visual-only intelligibility for FF and PLD. From these curves, we obtained estimates of the mean of the AV synchrony window as well as the low and high thresholds for asynchrony detection. For each subject we estimated the mean point of the synchrony window (MPS) as the mean of the Gaussian curve fit to the subject’s data. We operationalized the width of the AV synchrony window as the range of asynchronies over which subjects responded that the signals were synchronous more than half the time. For an estimate of this width, we calculated the full width at half maximum (FWHM) of the Gaussian curve. The auditory-leading

threshold for synchrony was calculated as the MPS minus the half width at half maximum, and the visual-leading threshold for synchrony was calculated as the MPS plus the half width at half maximum.

Table 1 presents a summary of the curve estimates obtained from a fit of the average response data weighted by the standard error. All statistical tests used estimates from fitting curves to individual subject data. Tukey HSD tests were used for all post-hoc analyses, and the familywise error rate was set to $\alpha=.05$.

Condition	MPS	FWHM	A-leading	V-leading
FF	47.405	388.39	-146.79	241.60
high VI	55.33	397.68	-143.51	254.17
low VI	47.749	373.42	-138.96	234.46
PLD	121.59	501.36	-129.09	372.27
high VI	116.42	483.63	-125.40	358.24
low VI	122.53	517.35	-136.14	381.20
NS	52.485	421.91	-158.47	263.44

Table 1. All numbers are in milliseconds. Negative numbers indicate that the auditory signal led the visual signal. VI = visual intelligibility; MPS = mean point of synchrony; FWHM = full width at half maximum; A-leading = auditory-leading threshold; V-leading = visual-leading threshold.

Mean Point of Synchrony (MPS)

The MPS was significantly larger than 0 in all three conditions of the experiment (FF: $t(14) = 7.119$; PLD: $t(14) = 15.656$; NS: $t(14) = 4.582$; all p 's < .001). Most of the participants had an MPS greater than 0 on all three conditions. One participant had an estimated MPS of A13V ms in the FF condition, and two had estimated MPSs of A20V and A24V ms in the NS condition.

A one-way repeated measures ANOVA revealed a significant effect of condition (NS, FF, PLD) on MPS ($F(2, 28) = 31.326$, $p < .001$). Post-hoc Tukey tests indicated that the MPS did not differ significantly for the NS and FF conditions. However, the MPS in the PLD condition was significantly larger than the MPS for either the NS or the FF conditions.

A two-way ANOVA on condition (FF, PLD) and visual-only intelligibility (high, low) revealed no significant overall effect of VI ($F(1, 14) < 1$). However, a significant interaction of Condition x VI ($F(1, 14) = 14.056$; $p < .05$) was found. Post-hoc Tukey tests revealed that both high and low VI words in the FF condition had a lower MPS than high or low VI words in the PLD condition. Also, in the FF condition, the high VI words had a significantly higher average MPS than the low VI words. The high VI words had an MPS that was on average 19.12 ms of visual lead higher than the low VI words. Although this difference was small, it was statistically significant and highly consistent across participants. Of the 15 participants, 12 showed the VI effect in the FF condition. Three participants had a lower MPS for the high VI FF words; in those cases the MPSs were 5.54 ms, 9.66 ms, and 12.67 ms smaller for high than for

low VI words. The difference between the high and low VI words in the PLD condition was in the opposite direction although it did not reach significance.

AV Synchrony Window

A one-way repeated measures ANOVA revealed a significant effect of condition on the size of the AV synchrony window ($F(2, 28) = 4.780, p < .05$). Post-hoc Tukey HSD tests showed that the synchrony window was larger in the PLD condition than in either the FF or NS conditions. A two-way repeated measures ANOVA on condition and VI revealed no additional significant effect of VI ($F(1, 14) < 1$) and no significant interaction of Condition x VI ($F(1, 14) < 1$).

Auditory-leading Thresholds

A one-way repeated measures ANOVA revealed no significant effect of condition on the auditory-leading threshold ($F(2, 28) = 1.850, p > .05$). Likewise, a two-way ANOVA on condition and VI revealed no significant effects (condition: $F(1, 14) = 1.233, p > .05$; VI: $F(1, 14) < 1$; Condition x VI: $F(1, 14) < 1$).

Visual-leading Threshold

The visual-leading threshold was significantly different across conditions ($F(2, 28) = 17.550, p < .001$). Post-hoc Tukey HSD analyses revealed that the PLD condition had a significantly higher visual-leading threshold than either the FF or the NS conditions. The FF and NS conditions did not differ in visual-leading threshold. A two-way ANOVA on condition and VI revealed no additional significant effect of VI ($F(1, 14) < 1$) and no significant interaction of Condition x Intelligibility ($F(1, 14) = 3.837, p > .05$).

Discussion

The present findings are consistent with previous studies and indicate that participants judged AV signals as subjectively synchronous for AV asynchronies that ranged over a window of several hundred milliseconds. In addition, participants judged larger asynchrony levels as subjectively synchronous with visual-leading stimuli than with auditory-leading stimuli. This AV processing asymmetry has also been reported in electrophysiological studies (King & Palmer, 1985; Meredith, 2002; Meredith, Nemitz, & Stein, 1987; B. Stein & Meredith, 1993); in behavioral studies using simple AV asynchronous stimuli (Dixon & Spitz, 1980; Lewald, Ehrenstein, & Guski, 2001; Lewkowicz, 1996); and in behavioral tasks involving AV speech (Dixon & Spitz, 1980; Grant & Greenberg, 2001; Grant & Seitz, 1998; Grant et al., 2003; McGrath & Summerfield, 1985; Pandey et al., 1986).

In contrast with a previous report by Dixon and Spitz (1980), however, the thresholds and means of the AV synchrony window were comparable for nonspeech (NS; average window: A159V to V263A ms; MPS = V52A ms) and full-face (FF) speech signals (average window: A147V to V242A ms; MPS = V47A ms). None of the differences obtained in this study between the NS and FF conditions were significant. However, we did find significant effects due to the visual characteristics of the speech stimuli. The size, MPS, and visual-leading thresholds in the PLD condition were all significantly higher than in the FF or NS conditions. Also, the MPS in the FF high VI condition was significantly higher than the MPS in the FF low VI condition; the effect of VI was small but highly consistent across subjects.

One issue raised by the results obtained in Experiment 1 is whether the duration chosen for the nonspeech signals, 100 ms, could have influenced the characteristics of the AV synchrony window for

simple AV signals, and potentially exaggerated the similarity between the windows for speech and nonspeech. For instance, it is possible that 100 ms may be a “special” number for AV interactions. Neurophysiological studies have indicated that multisensory enhancement may not occur if signals from multiple sensory modalities do not occur within 100 ms of each other (King & Palmer, 1985). Similarly, a recent behavioral study in humans has reported multisensory interactions only for AV asynchronies of up to 100 ms (Shams, Kamitani, & Shimojo, 2002). If 100 ms is a “special” duration, then the results for the nonspeech condition in Experiment 1 could be more similar to the results for the FF condition than would be expected if the stimuli used in the NS condition had been shorter or longer in duration.

More generally, Meredith et al. (1987) reported that multisensory neurons in the superior colliculus of the cat show multisensory enhancement when discharge trains from the unimodal stimuli overlap. This multisensory enhancement is greatest during overlap of the peak unimodal discharge trains. In our behavioral task, the AV synchrony window, which we took as a behavioral correlate of multisensory enhancement, could be similarly affected by the overlap of the peak neural response to a stimulus. Using data for NS stimuli of only one duration, it is difficult to assess whether the relevant stimulus information comes from the stimulus onset, offset, or overall duration, or some combination of these. To explore this issue further, we conducted a second experiment to investigate the effects of different durations of NS stimuli on the characteristics of the AV synchrony window.

Experiment 2

In Experiment 2, we examined the effect of nonspeech signal duration on AV synchrony judgments. We also used the same FF condition as in Experiment 1 as a baseline measure. If signal duration is an important cue in synchrony detection, then subjects might be better at detecting asynchronies in stimuli with shorter durations than in stimuli with longer durations. This might be the case because auditory and visual stimuli that are longer in duration have longer durations of overlap than shorter-duration AV stimuli at the same asynchrony level. For example, suppose we have a stimulus in which the auditory signal leads the visual signal by 300 ms (A300V). If the auditory and visual signals are each 33 ms in duration, then there is a “gap” of 267 ms between the offset of the auditory signal and the onset of the visual signal. However, if the auditory and visual signals are both 500 ms in duration, then the offset of the auditory signal will not occur until 200 ms after the onset of the visual signal. Note that in this example, stimulus offset also varies with stimulus duration. Thus, if the onset asynchrony was more important than the duration and/or offset of the signal for the size of the AV synchrony window, we would not expect to see any significant differences in the asynchrony judgments for nonspeech stimuli of different durations.

Methods

Participants

Participants were 23 undergraduate students at Indiana University (6 male and 17 female, mean age of 19.39 years). All were recruited from the Indiana University subject pool and 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.

Stimuli

Stimuli were created with the same methodology and auditory-visual onset asynchronies used in Experiment 1. The duration of the signals was manipulated so that in the first condition the auditory and

visual signals were both 33 ms (NS33); in the second they were 100 ms (as in Experiment 1; here, NS100); and in the third they were 500 ms (NS500). The 500-ms condition was used because the AV words in the FF condition were on average 500-ms long.

The stimuli used in the FF and NS100 conditions were identical to those used in Experiment 1. The NS33 and NS500 conditions used the same red circle and 3000-Hz tone as the NS100 condition, but differed in the duration of the auditory and visual signals, which were both 33 ms in the NS33 condition and both 500 ms in the NS500 condition.

Procedure

The procedure was identical to that described in Experiment 1, with the following exceptions. All four conditions (FF, NS33, NS100, and NS500) were tested, with 25 asynchrony levels \times 10 trials per level = 250 trials per condition. The three NS blocks were tested before the FF block, but the order of the three NS blocks was counterbalanced across participants. Response hand was also counterbalanced across participants.

Results

Average response data for the FF, NS33, NS100, and NS500 conditions are displayed in Figure 2. As in Experiment 1, individual subject data were fit with Gaussian curves. The mean of the curve was taken as the mean point of the AV synchrony window (MPS), and the auditory- and visual-leading thresholds were the low and high endpoints of the FWHM. Again, all statistical analyses were performed on individual subject data. Table 2 contains a summary of the curve estimates for the average subject data weighted by the standard error.

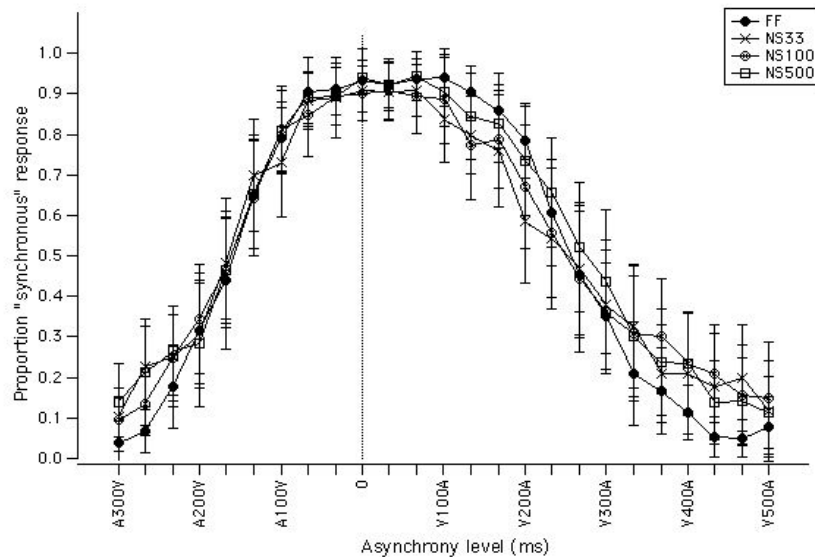


Figure 2. Average “in sync” response for all participants in Experiment 2, in FF, NS33, NS100, and NS500 conditions. The dotted vertical line is at 0-ms asynchrony.

Condition	MPS	FWHM	A-leading	V-leading
FF	53.8748	394.42	-143.34	251.09
high VI	62.8317	385.04	-129.69	255.35
low VI	40.192	407.50	-163.56	243.94
NS33	39.867	410.57	-165.42	245.15
NS100	52.757	425.02	-159.75	265.27
NS500	51.1082	448.20	-172.99	275.21

Table 2. All numbers are in milliseconds. Negative numbers indicate that the auditory signal led the visual signal. VI = visual intelligibility; MPS = mean point of synchrony; FWHM = full width at half maximum; A-leading = auditory-leading threshold; V-leading = visual-leading threshold.

Mean Point of Synchrony (MPS)

The MPS was significantly larger than zero in all four conditions (FF: $t(22) = 8.576, p < .05$; NS33: $t(22) = 4.517, p < .05$; NS100: $t(22) = 2.683, p < .05$; NS500: $t(22) = 4.973, p < .05$). Again, the majority of the participants (15 out of 23) showed an MPS greater than 0 for all conditions. Of the remaining eight participants, four had an MPS greater than 0 in one of the NS conditions only, three had an MPS greater than 0 in two of the NS conditions only, and one had an MPS greater than 0 in the FF condition and two of the NS conditions.

The MPS did not differ significantly across the conditions overall ($F(3, 66) = 1.049, p > .05$). However, as observed in Experiment 1, the high VI condition had a significantly higher MPS than the low VI condition ($t(22) = 4.428, p < .05$), with an average difference of 21.78 ms. The VI effect on the MPS was found in 20 of the 23 subjects, with the remaining three having low VI MPSs that were 1.33 ms, 3.63 ms, and 39.09 ms higher than their high VI MPSs. This VI effect on the MPS replicated the findings reported in Experiment 1.

AV Synchrony Window

A one-way ANOVA revealed no significant effect of condition on the size of the AV synchrony window ($F(3, 66) = 1.785, p > .05$). The size of the AV synchrony window was not significantly different for high versus low VI FF words ($t(22) = .730, p > .05$).

Auditory-leading Thresholds

The auditory-leading threshold did not differ significantly overall across conditions ($F(3, 66) = 1.171, p > .05$). The auditory-leading thresholds for the high and low VI FF words were not significantly different ($t(22) = .412, p > .05$).

Visual-leading Thresholds

A one-way ANOVA revealed no significant effect of condition on the visual-leading threshold ($F(3, 66) = 1.881, p > .05$). In addition, high and low VI FF words did not differ significantly for visual-leading threshold ($t(22) = 1.529, p > .05$).

Discussion

The results of Experiment 2 revealed no significant effects of the duration of the NS signals on the AV synchrony window for AV stimuli. In addition, the NS conditions did not differ overall from the FF condition, replicating the results found earlier in Experiment 1. Finally, the MPS for the FF high VI words was significantly higher than the MPS for the FF low VI words by about 20 ms, replicating the VI finding from Experiment 1.

The failure to find any effect of signal duration of the NS stimuli on AV synchrony detection in Experiment 2 suggests that the detection of asynchrony relies on processes related to stimulus onset rather than stimulus offset or duration. Of course, the case may differ for very short or very long stimuli; this is an empirical question that could be addressed in future research. Another explanation of our results is that the subjects were attending only to stimulus onset and ignoring other stimulus properties. Although this account cannot be completely ruled out based on our current data, the subjects were explicitly instructed to respond that the stimuli were synchronous only if they overlapped exactly. Further investigations, manipulating subjects' attentional strategies or response criteria, are needed to resolve this issue more definitively.

To begin to address these issues through a converging approach, we examined the individual words from the FF condition, in which significant effects of VI appeared in both Experiment 1 and Experiment 2. We reasoned that a more detailed analysis of the data for individual words might allow us to pinpoint what features of the particular utterances the subjects were relying on in making their asynchrony judgments. The speaker's face was visible throughout each speech movie, and we edited the movies so that global cues to asynchrony could not be used effectively (see Experiment 1, Methods). As a consequence, the physical onset of the auditory and visual stimuli would be a less reliable cue to asynchrony in the FF condition than in the NS conditions. However, word-internal articulatory events might have some influence on when a particular word was judged to be synchronous.

Word Item Analysis

The FF condition data obtained from 50 participants (15 from Experiment 1, 23 from Experiment 2, and 12 additional participants) were analyzed separately by word. Figure 3 shows the range of asynchronies over which 50% or more of participants responded "synchronous" for each word. Table 3 shows the VI of the word, size of the window over which 50% or more of participants responded "synchronous" for each word, the auditory-leading and visual-leading limits of that window, and the auditory duration of each word. The number of video frames presented per word varied according to synchrony level as described in the "Methods" section of Experiment 1, so video duration is not included in the table.

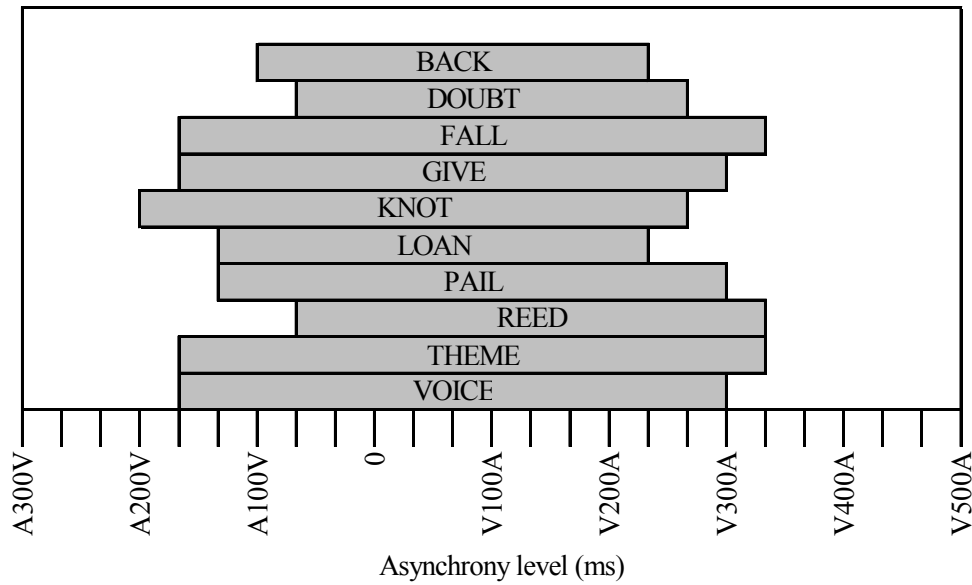


Figure 3. Asynchronies for which more than 50% ($\geq 25/50$) participants responded “in sync,” broken down by word.

Word	VI	Window	AV limit	VA limit	Dur(A)
BACK	low	300	A100V	V200A	397
DOUBT	high	300	A67V	V233A	467
LOAN	high	333	A133V	V200A	475
REED	high	367	A67V	V300A	522
PAIL	low	400	A133V	V267A	473
GIVE	low	433	A167V	V267A	388
KNOT	high	433	A200V	V233A	580
VOICE	low	433	A167V	V267A	647
FALL	high	500	A167V	V300A	490
THEME	low	500	A167V	V300A	656

Table 3. All numbers are in milliseconds. VI = visual intelligibility classification; Window = size of window including asynchrony levels for which 50% or more of participants responded "in sync"; AV limit = auditory-leading limit of window; VA limit = visual-leading limit of window; Dur(A) = duration of the auditory word.

The words in Table 3 are listed in order from those with the smallest asynchrony window to those with the largest window. Smaller windows were taken to indicate greater overall accuracy in judging when the words were synchronous. Table 3 indicates that the overall statistical difference observed in MPS between words defined as “high” or “low” VI based on whole-word scores may not provide the best description of the word item data. Likewise, the auditory duration does not seem to completely explain accuracy. However, the specific phonetic characteristics of the word, particularly the articulatory properties of the initial consonant, do appear to influence accuracy. The two words with the smallest synchrony windows, *back* and *doubt*, both begin with voiced stops articulated in the front of the mouth. The words with the largest synchrony windows, *fall* and *theme*, both begin with voiceless fricatives. The mid-range words begin with liquids (*loan* and *reed*), a voiceless bilabial stop (*pail*), a voiced velar stop (*give*), a nasal (*knot*), and a voiced fricative (*voice*).

The vowels and final consonants of the individual words appear to have a less regular relationship to AV synchrony judgments. For example, the same vowels are found in words with small and large windows (e.g., *reed* and *theme*) and short and long vowels are found at both ends of the spectrum as well. Also, *doubt* and *knot* have voiceless alveolar stops in final position, and *loan* and *theme* both have nasals in final position.

Taken together, this pattern of results suggests that the articulatory properties of the initial phonetic segment influence the accuracy with which words can be identified as synchronous or not. Stops that are voiced and articulated at the front of the mouth were identified most readily as synchronous. Segments articulated at the front of the mouth are easier to see than those at the back (e.g., velar stops; Summerfield, 1987), and stops also provide a discrete and relatively well-defined auditory and visual boundary. The shorter voice-onset time in voiced stops may also be linked more closely to the relevant aspects of the visual articulation than the longer voice-onset time in voiceless stops. Of course, further research will be necessary to determine whether the initial consonant is of primary importance in AV synchrony detection for a larger set of words that are specifically controlled for phonological contrasts, but the results of this initial analysis suggest that the phonetic properties of the initial consonant affect AV asynchrony judgments in this task.

General Discussion

Although the results of Experiments 1 and 2 are consistent with previous findings reported in the literature, they also provide several new insights into AV asynchrony detection. Both experiments demonstrated a similar AV synchrony window for speech and nonspeech sounds, in contrast to previous reports that suggested a larger window for speech sounds (Dixon & Spitz, 1980; McGrath & Summerfield, 1985). On the other hand, the PLD stimuli resulted in an AV synchrony window that was larger on the visual-leading side than the FF or NS windows. Finally, the onset of the relevant aspects of the stimulus, rather than the duration or offset of the stimulus, seemed to be important for judgments about asynchrony in both speech and nonspeech stimuli.

The Size and Shape of the AV Synchrony Window

The width of the AV synchrony window may reflect general information processing constraints (Munhall et al., 1996). For example, Guski and Troje (2003) have argued that events are linked at a perceptual level when they occur within around 200 ms of each other; they point out that the window for visual iconic processing is generally held to be around 250 to 300 ms and that the window for auditory echoic memory is around 250 ms. A multisensory interaction window that is several hundred milliseconds long is also consistent with the estimates of the temporal window for multisensory

enhancement and/or depression reported in electrophysiological studies in animals (King & Palmer, 1985; Meredith, 2002; Meredith et al., 1987; Stein & Meredith, 1993).

Several possible explanations have been proposed for the auditory-visual asymmetry observed in the intersensory temporal synchrony window. Some researchers have suggested that visual-leading asynchronies are tolerated more easily because they reflect long-term perceptual learning (Dixon & Spitz, 1980; McGrath & Summerfield, 1985). Specifically, perceivers might be able to more easily accommodate multimodal events in which the visual component begins before the auditory component because this type of event is common in their experience of the natural world (e.g., lightning preceding thunder). By contrast, events in which the auditory component comes before the visual component would not be expected based on prior experience and learning.

Another proposal is that the first modality to occur determines the timecourse of processing for asynchronous AV speech (Grant & Greenberg, 2001; Grant et al., 2003). This explanation is based on the hypothesis that visual speech cues from jaw and lip movements provide syllabic information relevant to the perception of place of articulation, whereas auditory speech cues provide information about voicing and manner of articulation (Summerfield, 1987). Visual syllabic information on the order of 200 to 250 ms is taken to be complementary to auditory information, which is hypothesized to be more important for phonological analysis and is conveyed at a faster rate of around 40 to 120 ms or less (Grant & Greenberg, 2001; Grant et al., 2003; van Wassenhove, Grant, & Poeppel, 2003). This explanation of AV interactions predicts a smaller integration window when the auditory speech signal leads than when the visual signal leads. However, this account is not consistent with the present results because we did not obtain any significant differences in asynchrony judgments between full-face speech and simple nonspeech signals, for which phonemic or syllabic considerations do not apply. In addition, recent findings indicate that phonetic information can be used in visual-only tasks, suggesting that meaningful visual information can be conveyed in speech below the syllabic level (Bernstein, Demorest, & Tucker, 2000; Lachs, 1999; Lachs & Pisoni, in press-a, in press-b; Mattys, Bernstein, & Auer, 2002).

Finally, another explanation for the auditory-visual asymmetry involves the timing of auditory and visual signals in the nervous system (Lewald et al., 2001). As Lewald and his colleagues (2001) and Schroeder and Foxe (2002) have noted, there are both physical and physiological differences in the transmission of light and sound. Light travels faster than sound in air, but stimulus transduction takes longer in the retina than in the cochlea (Lewald et al., 2001). Also, in discussing results from an AV asynchrony detection task in infants, Lewkowicz (1996) pointed out that the latency of the earliest evoked potentials are about 30-40 ms faster for auditory than visual signals.

In addition to these considerations, arrival time of auditory and visual signals to different subcortical and cortical regions differs as a function of which region is under consideration and the physical distance of the observer from the stimulus (Lewald & Guski, 2003; Schroeder & Foxe, 2002). For example, at distances of a meter or less, comparable to the observer's distance from the visual stimulus in our experiments (the participants were wearing headphones, so the distance of the auditory stimulus was essentially 0), auditory signals would be predicted to arrive around 40 ms before the visual signals in auditory association cortex. However, at the same distance, superior temporal polysensory areas would receive auditory and visual inputs at approximately the same latency of around 23-25 ms (Schroeder & Foxe, 2002). At further distances of around 40 feet, AV inputs to auditory association cortex would become synchronous while superior temporal polysensory areas would receive auditory and visual inputs asynchronously.

In a recent behavioral study, Sugita and Suzuki (2003) reported that the estimated time of arrival of an auditory stimulus increases with the viewing distance to the visual stimulus. In their study, stimuli

were judged as synchronous at a distance of 1 m when the auditory stimulus lagged by 5 ms, and at a distance of 20 m when the auditory stimulus lagged by 50 ms. The authors suggested that up until about 10 m of viewing distance, this increase was consistent with the brain's compensating for the slower velocity of sound. Such results point to the need for further experiments to clarify the role of viewing distance in auditory-visual interactions and perception of AV synchrony.

Although it is unclear at this time whether the relevant differences in auditory and visual processing times lie in transduction or occur later in neural information processing, the present findings and those reviewed from other studies suggest that at least under these presentation conditions auditory information is processed more quickly by the nervous system than visual information. The average MPS was around V50A ms for all but the PLD conditions, suggesting that visual leads of around 50 ms were most likely to be perceived as synchronous and that auditory stimuli may have been processed about 50 ms faster than visual stimuli. If we assume that the AV synchrony window is centered around V50A ms and extend the window 200 ms in either direction, as suggested by general information processing constraint explanations reviewed earlier (Guski & Troje, 2003; Munhall et al., 1996), we obtain a predicted AV synchrony window that extends from A150V to V250A ms. This is quite similar to the results we obtained for the FF and all the NS conditions in the present series of experiments.

Relationship to Neural Data

The average MPSs for the FF and the three NS conditions ranged from approximately V40A to V60A ms, indicating that the likelihood of a synchronous judgment was maximal when visual input led auditory input by approximately that time interval. Other recent behavioral studies of AV processing indicated that optimal performance on several perceptual tasks occurred with auditory delays of between 50 and 100 ms (Guski & Troje, 2003; Lewald & Guski, 2003). In general, these behavioral results imply that AV interactions relevant for the perception of synchrony occur early in neural processing.

Several neuroimaging studies have reported modulated activity in primary auditory and/or visual cortex during AV perception. For speech stimuli presented visual-only, enhanced auditory cortex activity has been reported in MEG (Sams et al., 1991) and fMRI studies (Calvert et al., 1997; Calvert & Campbell, 2003; MacSweeney et al., 2000) (but see Bernstein et al., 2002, for an exception). An EEG independent-components analysis of AV speech perception also suggested that enhanced auditory cortex activity was an important locus for the visual enhancement effect obtained for AV over auditory-alone presentation (Callan, Callan, Kroos, & Vatikiotis-Bateson, 2001).

Interestingly, Schroeder and Foxe (2002) reported that in the macaque, visual feedback reaches posterior auditory cortex at 50 ms poststimulus, while auditory feedforward input arrives at around 11 ms poststimulus. They suggest two possible origins for the visual signal—superior temporal polysensory areas and prefrontal cortex. In either case, they hypothesized that the earlier arriving auditory input could modulate responsiveness to the later arriving visual input. Based on the estimates of Schroeder and Foxe, visual and auditory inputs could be expected to arrive simultaneously at posterior auditory cortex in the macaque if the visual stimulus occurs approximately 40 ms before the auditory stimulus.

Recent data from event-related potential studies in humans suggest that the earliest audiovisual interaction in cortex can be detected over posterior cortex 40 to 50 ms after the presentation of a synchronous audiovisual stimulus (Giard & Peronnet, 1999; Molholm et al., 2002; Teder-Sälejärvi, McDonald, Di Russo, & Hillyard, 2002; van Wassenhove et al., 2003). However, because the ERP methodology has poor spatial resolution, the neural substrates of these early AV components remain unclear, and suggestions vary as to whether they are due to enhanced neural activity in auditory or in visual cortex. One recent study of nonspeech stimuli by Teder-Sälejärvi et al. (2002) concluded that the

early AV interaction was due to subjects' anticipation of stimulus presentation. Some researchers, especially those who used nonspeech stimuli, have suggested that the early interaction might be due to visual cortex activation (Fort, Delpuech, Pernier, & Giard, 2002; Giard & Peronnet, 1999), either from recently discovered feedforward projections from auditory cortex to early visual cortex or from auditory feedback from multisensory areas to early visual cortex (Fort et al., 2002; Molholm et al., 2002).

Researchers who used speech stimuli have argued instead for an early visual modulation of auditory processing (Pourtois, de Gelder, Vroomen, Rossion, & Crommelinck, 2000; van Wassenhove et al., 2003). Pourtois and colleagues (2000) found that presenting static expressive faces paired with voices resulted in modulation of auditory N1 by 90 to 130 ms poststimulus, the earliest significant interaction they observed. Interestingly, Giard and Peronnet (1999) reported an effect of sensory dominance of their participants on whether early enhancements occurred in visual or auditory cortex. Using nonspeech stimuli, they found that although visually dominant participants showed more enhanced early activity over auditory cortex, auditory-dominant participants showed more enhanced early activity over visual cortex. Further investigation is needed to clarify these issues.

Two recent imaging studies have specifically examined brain response to asynchronous AV stimuli and both have suggested subcortical rather than cortical involvement as the crucial factor. In a PET study, Bushara and colleagues (2001) measured detection of three auditory-leading and three visual-leading asynchronies using a circle paired with a tone. They reported that rCBF responses in the right insula increased with shorter asynchronies and that these responses were positively correlated with responses in the superior colliculus region of the posterior midbrain; the right posterior thalamus, precuneus, and prefrontal cortex; and the left insula. No correlations of right insular activity with superior temporal regions were observed.

In addition, an fMRI study by Olson, Gatenby, and Gore (2002) assessed the McGurk effect in synchronously and asynchronously presented speech, although only one asynchronous condition was used, in which the auditory signal was delayed by 1 second relative to the visual signal. They reported that although superior temporal regions were involved in both synchronous and asynchronous presentations, only the claustrum showed differential (increased) activity during the synchronous presentation condition. Olson et al. suggested that subcortical regions are more highly sensitive to timing of crossmodal stimuli than superior temporal regions. This hypothesis was also proposed in a review article by Calvert (2001), in which she suggested that the superior temporal sulcus is involved in stimulus identification and that the insula and superior colliculus are involved in stimulus timing. However, Calvert suggested that the left claustrum might be preferentially involved in crossmodal matching tasks. Because the Olson et al. study used McGurk stimuli that were incongruent across auditory and visual modalities, the mismatched nature of the stimuli could be responsible for the claustrum activity that was reported in their study.

Further neuroimaging research on auditory-visual asynchrony in both speech and nonspeech signals will be necessary to clarify whether the two types of signals show similar patterns of neural activation. In addition, future neuroimaging studies should be conducted with a wider range of AV asynchronies in order to examine more explicitly the neural responses to stimuli that fall both within and outside the AV synchrony window.

Context Effects: PLDs

In Experiment 1, we found that the AV synchrony window for PLD stimuli was centered on about V120A and was about 50 ms wider on both the auditory-leading and the visual-leading sides than the FF and NS windows. The PLDs had the same auditory-leading threshold as the FF and NS conditions,

but the visual-leading threshold was higher. There are several possible explanations for these results. One suggestion is that presenting the auditory signal first made matching to the unfamiliar PLD visual signal easier compared to presenting the PLD visual signal before the auditory signal. Recent findings by Lachs and Pisoni (in press-c) on crossmodal matching using isolated auditory words and PLDs do not support this idea. Participants in these studies were equally successful at matching auditory to visual and visual to auditory presentations. However, it is possible that other familiarity or “top-down” processing effects may have played a role in the PLD results. In addition, the physical characteristics of the PLD visual signals, such as low luminance and dispersion across the screen, might affect visual processing of these stimuli. Preliminary work in our lab has begun to examine this issue in more detail.

Context Effects on VI: Word Item Analysis

In both Experiments 1 and 2, the FF high VI words consistently had a higher MPS than the FF low VI words. However, the word-item analysis suggested that the VI results could be explained by participants’ sensitivity to the phonetic properties of the initial consonants rather than vowels or final consonants. Initial voiced stops articulated near the front of the mouth were identified as asynchronous most easily, whereas voiceless fricatives were the most difficult. These findings are consistent with the results of Experiment 2, which showed that nonspeech signals of different durations produced similar asynchrony detection results. Taken together, the results suggest that the onset of the two signals rather than the duration or offset is the critical factor controlling the perception of AV synchrony.

The importance of phonetic information in tasks involving AV speech has been reported previously (Bernstein et al., 2000; Lachs, 1999; Lachs & Pisoni, in press-a, in press-b; Mattys et al., 2002; Smeele, Sittig, & van Heuven, 1992). Smeele and her colleagues (1992) reported that for bimodal Dutch nonsense CVC words presented asynchronously in noise, the initial consonant was identified significantly more accurately with the auditory signal leading; conversely, the vowel and final consonant were identified more accurately with visual leads. Further research is needed using a larger inventory of words explicitly controlled for phonological inventory to determine whether the detection of AV asynchrony in speech is also affected by the phonetic and articulatory properties of the vowel and final consonant segments or only by the initial consonant segment as the results of the present study suggest.

Individual Differences in the AV Synchrony Window

The present study reports results on the perception of AV speech and nonspeech sounds in normal-hearing, typically developing adults. Additional findings have been reported suggesting that other atypical populations may have difficulties processing timing for unimodal and/or crossmodal stimuli. For example, children and adults with dyslexia have difficulties making judgments about the timing of crossmodal stimuli (Laasonen, Service, & Virsu, 2002; Laasonen, Tomma-Halme, Lahti-Nuutila, Service, & Virsu, 2000). It has been suggested that dyslexic individuals may have auditory, visual, and other sensory and motor deficits in processing transient stimuli that change quickly over time (J. Stein & Walsh, 1997).

In our own lab, we had an opportunity to examine the sensitivity to AV asynchrony in a postlingually deafened adult (“Mr. S”) who received a cochlear implant after two years of deafness (Goh, Pisoni, Kirk, & Remez, 2001). At the time of testing, our patient had used his implant for nine years. Mr. S has performed visual-only lipreading tasks at a consistently high level, with scores on the CUNY sentences presented visual-only of about 80% of words correct (baseline for cochlear implant patients who participated in another study (Goh et al., 2001) was around 24%). On our AV asynchrony judgment task, Mr. S was more accurate overall at detecting AV asynchrony than all but one of the 50 normal-hearing subjects tested, displaying a smaller AV synchrony window for all conditions and auditory- and

visual-leading thresholds closer to 0 for all but the NS100 condition. Figure 4 shows Mr. S's response data along with the data from the normal-hearing subject who performed most similarly to Mr. S. Data from two representative normal-hearing young adult subjects from Experiment 1 are also included for comparison.

In light of Mr. S's impressive lipreading abilities and previous reports that good lipreaders may be better at detecting AV asynchrony (McGrath & Summerfield, 1985; Pandey et al., 1986; but see Grant & Seitz, 1998), further investigation into the potential relationship between lipreading skills and sensitivity to temporal asynchrony between auditory and visual stimuli seems warranted and is currently underway in our lab (see Conrey, 2004, this volume).

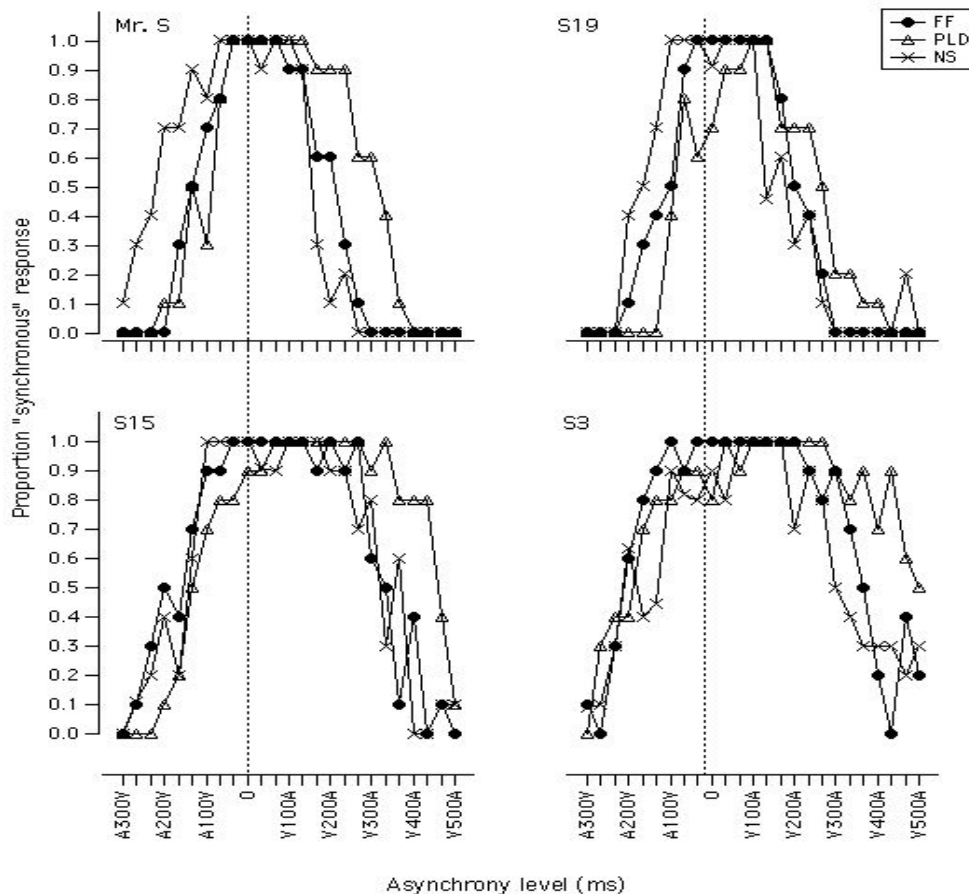


Figure 4. “In sync” response data for Mr. S, a cochlear implant patient and exceptionally good lipreader, and for three normal-hearing subjects from Experiment 1. Mr. S's performance on AV asynchrony detection was superior to that of 49 of 50 of our normal-hearing subjects. Top left panel: Mr. S. Top right panel: Subject 19, who performed comparably to Mr. S. Bottom panels: Subject 15 (left) and Subject 3 (right), whose performance was typical for normal-hearing subjects. The dotted vertical lines are at 0-ms asynchrony.

Conclusions

In summary, the results of the present experiments suggest a window of AV asynchronies several hundred milliseconds wide over which participants were unable to detect asynchronies above chance. FF speech signals and simple nonspeech signals did not differ statistically in terms of the mean, width, or auditory- or visual-leading thresholds of the AV synchrony window. Further research is needed on the characteristics of visual signals, such as PLDs, that significantly affect the size and shape of the AV synchrony window. In addition, the importance of the onset versus the duration or offset of the AV signal should be investigated with longer and shorter speech and nonspeech signals, with manipulations of subjects' attentional strategies, and with phonetically balanced word lists. Finally, future EEG and neuroimaging work should build on earlier studies such as Bushara et al. (2001) and Olson et al. (2002) in order to elucidate the neural mechanisms involved in the perception and detection of synchrony between auditory and visual signals. Such investigations should provide further insights into the timecourse and underlying neural mechanisms involved in auditory-visual multimodal processing.

References

- Bergeson, T.R., Reynolds, J.T., & Pisoni, D.B. (2003). Perception of point light displays of speech by normal-hearing adults and deaf adults with cochlear implants. Paper presented at the *AVSP 2003 International Conference on Auditory-Visual Speech Processing*.
- Bernstein, L.E., Auer, E.T., Jr., Moore, J.K., Ponton, C.W., Don, M., & Singh, M. (2002). Visual speech perception without primary auditory cortex activation. *NeuroReport*, *13*, 311-315.
- Bernstein, L.E., Demorest, M.E., & Tucker, P.E. (2000). Speech perception without hearing. *Perception & Psychophysics*, *62*, 233-252.
- Bushara, K.O., Grafman, J., & Hallett, M. (2001). Neural correlates of auditory-visual stimulus onset asynchrony detection. *Journal of Neuroscience*, *21*, 300-304.
- Callan, D.E., Callan, A.M., Kroos, C., & Vatakotis-Bateson, E. (2001). Multimodal contribution to speech perception revealed by independent component analysis: A single-sweep EEG case study. *Cognitive Brain Research*, *10*, 349-353.
- Calvert, G. (2001). Crossmodal processing in the human brain: Insights from functional neuroimaging studies. *Cerebral Cortex*, *11*, 1110-1123.
- Calvert, G., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C.R., McGuire, P.K., et al. (1997). Activation of auditory cortex during silent lipreading. *Science*, *276*, 593-596.
- Calvert, G., & Campbell, R. (2003). Reading speech from still and moving faces: The neural substrates of visible speech. *Journal of Cognitive Neuroscience*, *15*, 57-70.
- Cohen, J.D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, and Computers*, *25*, 257-271.
- Conrey, B.L. (2004). Multimodal sentence intelligibility and the detection of auditory-visual asynchrony in speech and nonspeech signals: A first report. In *Research on Spoken Language Processing Report No. 26* (pp. 345-356). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Dixon, N., & Spitz, L. (1980). The detection of audiovisual desynchrony. *Perception*, *9*, 719-721.
- Fort, A., Delpuech, C., Pernier, J., & Giard, M.-H. (2002). Early auditory-visual interactions in human cortex during nonredundant target identification. *Cognitive Brain Research*, *14*, 20-30.
- Giard, M.H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience*, *11*, 473-490.
- Goh, W.D., Pisoni, D.B., Kirk, K.I., & Remez, R.E. (2001). Audio-visual perception of sinewave speech in an adult cochlear implant user: A case study. *Ear & Hearing*, *22*, 412-419.

- Grant, K.W., & Greenberg, S. (2001). Speech intelligibility derived from asynchronous processing of auditory-visual information. Paper presented at the *AVSP International Conference on Auditory-Visual Speech Processing*.
- Grant, K.W., & Seitz, P.F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *Journal of the Acoustical Society of America*, *104*, 2438-2450.
- Grant, K.W., van Wassenhove, V., & Poeppel, D. (2003). Discrimination of auditory-visual synchrony. Paper presented at the *AVSP 2003 International Conference on Auditory-Visual Speech Processing*.
- Guski, R., & Troje, N. (2003). Audio-visual phenomenal causality. *Perception & Psychophysics*, *65*, 789-800.
- King, A.J., & Palmer, A.R. (1985). Integration of visual and auditory information in bimodal neurones in the guinea-pig superior colliculus. *Experimental Brain Research*, *60*, 492-500.
- Laasonen, M., Service, E., & Virsu, V. (2002). Crossmodal temporal order and processing acuity in developmentally dyslexic young adults. *Brain and Language*, *80*, 340-354.
- Laasonen, M., Tomma-Halme, J., Lahti-Nuutila, P., Service, E., & Virsu, V. (2000). Rate of information segregation in developmentally dyslexic children. *Brain and Language*, *75*, 66-81.
- Lachs, L. (1999). Use of partial stimulus information in spoken word recognition without auditory stimulation. In *Research on Spoken Language Processing Report No. 25* (pp. 82-114). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Lachs, L. (2002). Vocal tract kinematics and crossmodal speech information (*Research on Spoken Language Processing Technical Report No. 10*). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Lachs, L., & Hernandez, L.R. (1998). Update: The Hoosier Audiovisual Multitalker Database. In *Research on Spoken Language Processing Progress Report No. 22* (pp. 377-388). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Lachs, L., & Pisoni, D.B. (in press-a). Crossmodal source identification in speech perception. *Ecological Psychology*.
- Lachs, L., & Pisoni, D.B. (in press-b). Crossmodal source information and spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*.
- Lachs, L., & Pisoni, D.B. (in press-c). Specification of crossmodal source information in isolated kinematic displays of speech. *Journal of the Acoustical Society of America*.
- Lewald, J., Ehrenstein, W.H., & Guski, R. (2001). Spatio-temporal constraints for auditory-visual integration. *Behavioural Brain Research*, *121*, 69-79.
- Lewald, J., & Guski, R. (2003). Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli. *Cognitive Brain Research*, *16*, 468-478.
- Lewkowicz, D.J. (1996). Perception of auditory-visual temporal synchrony in human infants. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 1094-1106.
- MacSweeney, M., Amaro, E., Calvert, G., Campbell, R., David, A.S., McGuire, P.K., et al. (2000). Silent speechreading in the absence of scanner noise: An event-related fMRI study. *NeuroReport*, *11*, 1729-1733.
- Massaro, D., & Cohen, M. (1993). Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables. *Speech Communication*, *13*, 127-134.
- Massaro, D., Cohen, M.M., & Smeele, P.M.T. (1996). Perception of asynchronous and conflicting visual and auditory speech. *Journal of the Acoustical Society of America*, *100*, 1777-1786.
- Mattys, S.L., Bernstein, L.E., & Auer, E.T., Jr. (2002). Stimulus-based lexical distinctiveness as a general word-recognition mechanism. *Perception & Psychophysics*, *64*, 667-679.
- McGrath, M., & Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *Journal of the Acoustical Society of America*, *77*, 678-684.
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746-748.

- Meredith, M.A. (2002). On the neuronal basis for multisensory convergence: A brief overview. *Cognitive Brain Research*, *14*, 31-40.
- Meredith, M.A., Nemitz, J.W., & Stein, B.E. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *The Journal of Neuroscience*, *7*, 3215-3229.
- Molholm, S., Ritter, W., Murray, M.M., Javitt, D.C., Schroeder, C.E., & Foxe, J.J. (2002). Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Cognitive Brain Research*, *14*, 115-128.
- Munhall, K.G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics*, *58*, 351-362.
- Olson, I.R., Gatenby, J.C., & Gore, J.C. (2002). A comparison of bound and unbound audio-visual information processing in the human cerebral cortex. *Cognitive Brain Research*, *14*, 129-138.
- Pandey, C.P., Kunov, H., & Abel, M.S. (1986). Disruptive effects of auditory signal delay on speech perception with lip-reading. *The Journal of Auditory Research*, *26*, 27-41.
- Pourtois, G., de Gelder, B., Vroomen, J., Rossion, B., & Crommelinck, M. (2000). The time-course of intermodal binding between seeing and hearing affective information. *NeuroReport*, *11*, 1329-1333.
- Rosenblum, L.D., & Saldaña, H.M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 318-331.
- Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O.V., Lu, S.-T., et al. (1991). Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters*, *127*, 141-145.
- Schroeder, C.E., & Foxe, J.J. (2002). The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex. *Cognitive Brain Research*, *14*, 187-198.
- Shams, L., Kamitani, Y., & Shimojo, S. (2002). Visual illusion induced by sound. *Cognitive Brain Research*, *14*, 147-152.
- Sheffert, S.M., Lachs, L., & Hernandez, L.R. (1996). The Hoosier audiovisual multitalker database. In *Research on Spoken Language Processing No. 21* (pp. 578-583). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Smeele, P.M.T., Sittig, A.C., & van Heuven, V.J. (1992). Intelligibility of audio-visually desynchronised speech: Asymmetrical effect of phoneme position. Paper presented at the *International Conference on Spoken Language Processing*.
- Stein, B., & Meredith, M.A. (1993). *The merging of the senses*. Cambridge, MA: MIT Press.
- Stein, J., & Walsh, V. (1997). To see but not to read: The magnocellular theory of dyslexia. *Trends in Neurosciences*, *20*, 147-152.
- Sugita, Y., & Suzuki, Y. (2003). Implicit estimation of sound-arrival time. *Nature*, *421*, 911.
- Sumby, W.H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*, 212-215.
- Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by Eye: The Psychology of Lip-reading*.
- Teder-Sälejärvi, W.A., McDonald, J.J., Di Russo, F., & Hillyard, S.A. (2002). An analysis of audio-visual crossmodal integration by means of event-related potential (ERP) recordings. *Cognitive Brain Research*, *14*, 106-114.
- van Wassenhove, V., Grant, K.W., & Poeppel, D. (2002). Temporal integration in the McGurk effect. Paper presented at the Poster presented at the annual meeting of the *Society for Cognitive Neuroscience*, San Francisco.
- van Wassenhove, V., Grant, K.W., & Poeppel, D. (2003). Electrophysiology of auditory-visual speech integration. Paper presented at the *AVSP 2003 International Conference on Auditory-Visual Speech Processing*.

