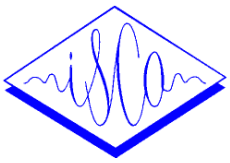

AVSP 2003 – International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4-7, 2003

Discrimination of Auditory-Visual Synchrony

Ken W. Grant[1], Virginie van Wassenhove[2], David Poeppel[2]

[1]Army Audiology and Speech Center, Walter Reed Army Medical Center, Washington, DC
[2]Neuroscience and Cognitive Science Program, Cognitive Neuroscience of Language Laboratory, University of Maryland, College Park, MD
grant@tidalwave.net

Abstract

Discrimination thresholds for temporal synchrony in auditory-visual sentence materials were obtained on a group of normal-hearing subjects. Thresholds were determined using an adaptive tracking procedure which controlled the degree of audio delay, both positive and negative in separate tracks, relative to a video image of a female speaker. Four different auditory filter conditions, as well as a broadband speech condition, were evaluated in order to determine whether discrimination thresholds were dependent on the spectral content of the acoustic speech signal. Consistent with previous studies of auditory-visual speech recognition which showed a broad, asymmetrical range of temporal synchrony (audio delays roughly between -40 ms and +240 ms) for which intelligibility was basically unaffected, synchrony discrimination thresholds also showed a broad, asymmetrical pattern of similar magnitude (audio delays roughly between -45 ms and 200 ms). No differences in synchrony thresholds were observed for the different filtered bands of speech, or for broadband speech. These results suggest a fairly tight coupling between a subject's ability to detect cross-modal asynchrony and the intelligibility of auditory-visual speech materials.

1. Introduction

Speech perception requires that listeners be able to combine information from many different parts of the audio spectrum in order to effectively decode the incoming message. This is not always possible for listeners in noisy or reverberant environments or for listeners with significant hearing loss because some parts of the speech spectrum, usually the high frequencies, are partially or completely inaudible, and most probably, distorted. Signal processing algorithms designed to remove some of the deleterious effects of noise and reverberation from speech signals often apply different processing strategies to low- or high-frequency portions of the spectrum. Thus, different parts of the speech spectrum are subjected to different amounts of signal processing depending on the goals of the processor and the listening environment. Ideally, none of these signal processing operations would entail any significant processing delays, however, this may not always be the case. Recent studies by Silipo et al. [1] and Stone and Moore [2] have shown that relatively small across-channel delays (< 20 ms) can result in significant decrements in speech intelligibility. Since it is imperative for listeners to combine information across spectral channels in order to understand speech, compensation for any frequency-specific signal-processing delays would seem appropriate.

But not all speech recognition takes place by hearing alone. In noisy and reverberant environments, speech recognition becomes difficult and sometimes impossible depending on the signal-to-noise ratio in the room or hall. Under these fairly common conditions, listeners make use of visual speech cues (i.e., via speechreading) to provide additional support to audition, and in most cases, are able to restore intelligibility back to what it would have been had the speech been presented in the quiet [3, 4]. Thus, in many listening situations, individuals not only have to integrate information across audio spectral bands, but also across sensory modalities [5]. As with audio-alone input, the relative timing of audio and visual input in auditory-visual speech perception can have a pronounced effect on intelligibility. And, because the bandwidth required for high fidelity video transmission is much broader than the bandwidth required for audio transmission (and therefore more difficult to transmit rapidly over traditional broadcast lines), there is more of an opportunity for the two sources of information to become mis-aligned. For example, in certain news broadcasts where foreign correspondents are shown as well as heard, it is often the case that the audio feed will proceed the video feed resulting in a combined transmission that is out of sync and difficult to understand. In fact, recent data reported by Grant and Greenberg [6] showed that in cases where the audio signal (comprised of a low- and high-frequency band of speech) leads the video signal, the intelligibility falls precipitously with very small degrees of audio-visual asynchrony. In contrast, when the video speech signal leads the audio signal, intelligibility remains high over a large range of asynchronies, out to about 240 ms. These results are shown below in Figure 1 and differ dramatically from those described in studies of across-channel spectral asynchrony in that when the video signal precedes the audio signal, intelligibility does not decline until the audio delay exceeds about 200 ms. However, when the audio signal precedes the video signal, intelligibility suffers immediately just as in the audio-alone experiments mentioned above [1, 2].

Another example of these unusually long temporal windows of integration can be found in the work of Van Wassenhove et al. [7]. In this study, the subjects' task was to identify consonants from stimuli composed of either a visual /gɑ/ paired with an audio /bɑ/, or a visual /kɑ/ paired with an audio /pɑ/. The stimulus onset asynchrony between audio and video portions of each stimulus were manipulated between -467 and +467 ms. When presented in synchrony, the most likely fusion responses for these pairs of incongruent auditory-visual stimuli are /dɑ/ (or /ðɑ/) and /tɑ/, respectively. However, when the audio and video components are made to be increasingly more asynchronous, fewer and

fewer fusion responses are given and the auditory response dominates. This pattern is shown in Figure 2 for the incongruent pair comprised of audio /pɑ/ and visual /kɑ/.

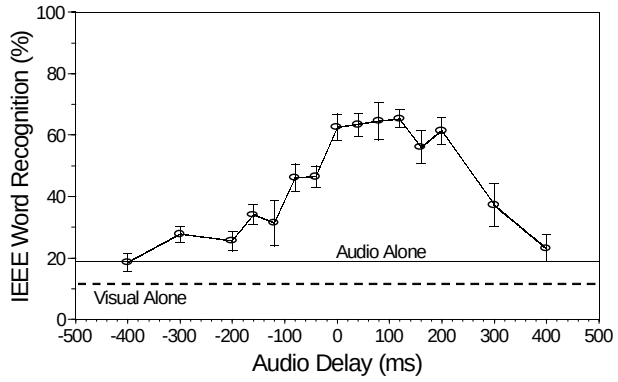

Figure 1. Average auditory-visual intelligibility of IEEE sentences as a function of audio-video asynchrony. Note the substantial plateau region between -50 ms audio lead to 200 ms audio delay where intelligibility scores are high relative to the audio-alone or video-alone conditions. Adapted from [6].

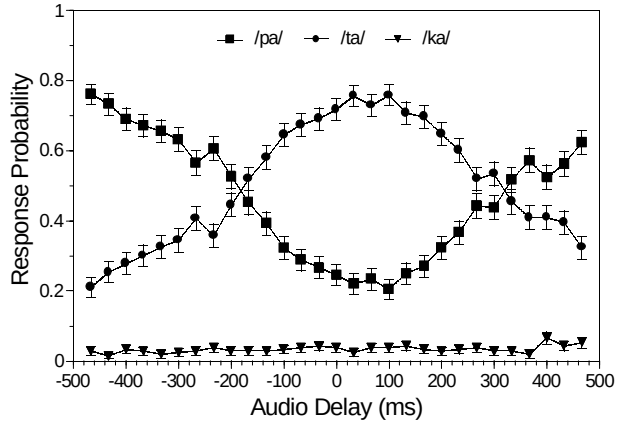

Figure 2. Labeling functions for the incongruent AV stimulus visual /kɑ/ and acoustic /pɑ/ as a function of audiovisual asynchrony (audio delay). Circles = probability of responding with the fusion response /tɑ/; squares = probability of responding with the acoustic stimulus /pɑ/; triangles = probability of responding with the visual response /kɑ/. Note the relatively long temporal window (-50 ms audio lead to +200 ms audio lag) where fusion responses are likely to occur. Adapted from [7].

One question that arises from these studies, and others like them [8, 9, 10], is whether subjects are even aware of the audio-video asynchrony inherent in the signals presented for audio delays corresponding to the plateau region where intelligibility is roughly unchanged. In other words, do the temporal windows of integration derived from studies of speech intelligibility correspond to the limits of synchrony perception? Or are subjects perceptually aware of small amounts of asynchrony that have no effect on intelligibility?

Another important question is whether the perception of auditory-visual synchrony depends on the spectral content of the acoustic speech signal? Previously, Grant and Seitz [11] and Grant [12] demonstrated that the cross-modal correlation between the visible movements of the lips (e.g., inter-lip distance or area of mouth) and the acoustic speech envelope depends on the spectral region from which the envelope is derived. In general, a significantly greater correlation has been observed for mid-to-high-frequency regions, typically associated with place-of-articulation cues, than for low-frequency regions or even broadband speech (Figure 3). Because the spectral content of the acoustic speech signal effects the degree of cross-modal correlation, it is possible that a similar relation might be found for the detection of cross-modal asynchrony. Specifically, we hypothesized that subjects would be better at detecting cross-modal asynchrony for speech bands in the F2-F3 formant regions (mid-to-high frequencies) than for speech filter bands in the F1 formant region (low frequencies).

The purpose of the current study was to address these various issues for auditory-visual speech perception in normal-hearing subjects using standard psychophysical methods that are likely to provide a more sensitive description of the limits of cross-modal temporal synchrony than those derived from speech identification experiments [6, 7, 8, 9, 10] or from subjective judgments of auditory-visual temporal asynchrony [7].

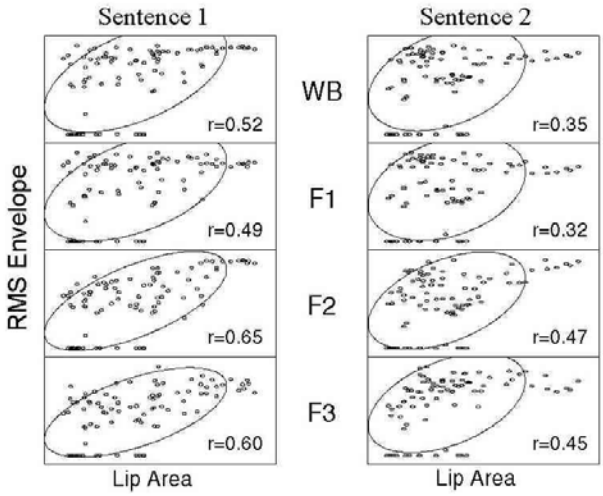

Figure 3. Scatter plots showing the degree of correlation between area of mouth opening and the rms amplitude envelope for two sentences ("Watch the log float in the wide river" and "Both brothers wear the same size"). The different panels represent the spectral region from which the speech envelopes were derived. WB = wide band, F1 = 100-800 Hz, F2 = 800-2200 Hz, F3 = 2200-6500 Hz. Adapted from [11].

2. Methods

This present study involved an adaptive, two-interval forced-choice discrimination task using band-pass filtered sentence materials presented under audio-video conditions. Normal-hearing subjects were asked to judge the synchrony between video and audio components of an audio-video speech stimulus. The video component was a movie of a female speaker producing one of two target sentences. The audio component was one of four different bandpass-filtered renditions of the target sentence. A fifth wide band speech

condition was also tested. The degree of audio-video synchrony was adaptively manipulated until the subject's discrimination performance (comparing an audio-video signal that is synchronized to one that is out of sync) converged on a level of approximately 71% correct. Audio-video synchronization threshold determinations were repeated several times and for several different audio speech bands representing low-, mid-, and high-frequency speech energy.

2.1. Subjects

Four adult listeners (35-49 years) with normal hearing participated in this study. The subject's hearing was measured by routine audiometric screening (i.e., audiogram) and quiet thresholds were determined to be no greater than 20 dB HL at frequencies between 250 and 6000 Hz. All subjects had normal or corrected-to-normal vision (static visual acuity equal to or better than 20/30 as measured with a Snellen chart). Written informed consent was obtained prior to the start of the study.

2.2. Stimuli:

The speech materials consisted of sentences drawn from the Institute of Electrical and Electronic Engineers (IEEE) sentence corpus [13] spoken by a female speaker. The full set contains 72 lists of ten phonetically balanced 'low-context' sentences each containing five key words (e.g., The *birch canoe slid* on the *smooth planks*). The sentences were recorded onto optical disc and the audio portions digitized and stored on computer. Two sentences from the IEEE corpus were selected for use as test stimuli. These were "The birch canoe slid on the smooth planks" and "Four hours of steady work faced us". The sentences were processed through a Matlab© software routine to create four filtered speech versions each comprised of one spectrally distinct 1/3-octave band (band 1: 298-375 Hz; band 2: 750-945 Hz; band 3: 1890-2381 Hz ; and band 4: 4762-6000 Hz). Finite impulse response (FIR) filters were used with attenuation rates exceeding 100 dB/octave. A fifth condition, comprised of the unfiltered wide band sentences, was also used.

2.3. Procedures

Subjects were seated comfortably in a sound-treated booth facing a computer touch screen. The speech materials were presented diotically (same signal to both ears) over headphones at a comfortable listening level. A 21" video monitor positioned 5 feet from the subject displayed films of the female talker speaking the target sentence. An adaptive two-interval, forced-choice procedure was used in which one stimulus interval contained a synchronized audio-visual presentation and the other stimulus interval contained an asynchronous audio-visual presentation. The assignment of the standard and comparison stimuli to interval one or interval two was randomized. Subjects were instructed to choose the interval containing the speech signal that appeared to be "out of sync". The subject's trial-by-trial responses were recorded in a computer log. Correct-answer feedback was provided to the subject after each trial.

The degree of audio-video asynchrony was controlled adaptively according to a two-down, one-up adjustment rule. Two consecutive correct responses led to a decrease in audio-video asynchrony (task gets harder), whereas an incorrect response led to an increase in audio-video asynchrony (task gets easier). At the beginning of each adaptive block of trials, the amount of asynchrony was 390 ms which was obvious to the subjects. The initial step size was a factor of 2.0, doubling and halving the amount of asynchrony depending on the subject's responses. After three reversals in the direction of the adaptive track, the step size decreased to a factor of 1.2, representing a 20% change in asynchrony. The track continued in this manner until a total of six reversals were obtained using the smaller step size. Thresholds for synchrony discrimination were computed as the mean of these last six reversals. A total of four to six adaptive blocks per filter condition were run representing both audio leading conditions and audio lagging conditions. Two different sentences per condition were used to improve the generalizability of the results.

3. Results and Discussion

The results, averaged across four subjects and two sentences, are displayed in Figure 4. A three-way repeated measures ANOVA with sentence, temporal-order (auditory lead and auditory lag), and filter-band condition as within subjects factors, showed a significant effect for temporal order ($F(1,3) = 109$, $p = 0.002$), but no effect for sentence or filter-band condition or any of the interactions.

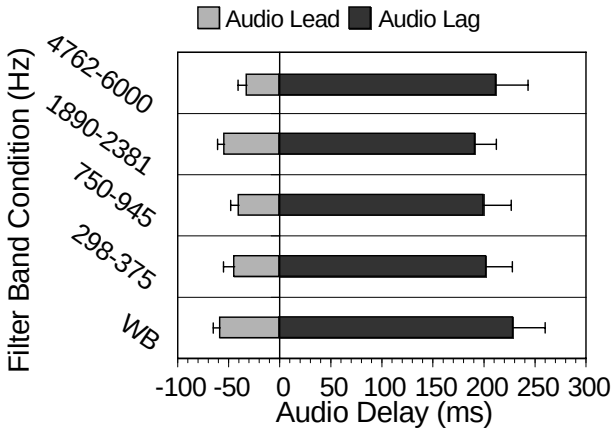

Figure 4. Average synchrony discrimination thresholds for unfiltered (wide band) speech and for four different bandpass-filtered speech conditions.

The fact that there was no significant difference in discrimination thresholds for the various filter conditions was somewhat unexpected given previous data [11, 12] showing that the correlation between lip kinematics and audio envelope tends to be best in the mid-to-high spectral regions (bands 3 and 4). Our initial expectation was that audio signals that are more coherent with the visible movements of the speech articulators would produce the most sensitive discrimination thresholds. However, because the correlation between lip kinematics and acoustic envelope are modest at best and are sensitive to the particular speaker and phonetic makeup of the sentence, differences in the degree of audio-video coherence across filter conditions may have been too subtle to allow for threshold differences to emerge (see Figure 3). Although not significant, it is interesting that the thresholds for the mid-frequency band between 1890-2381 Hz

were consistently smaller than those for the other frequency bands when the audio signal lagged the visual signal. Additional work with a larger number of sentences and subjects will be required to explore this issue further.

The two most compelling aspects of the data shown in Figure 4 are the overall size of the temporal window for which asynchronous audio-video speech input is perceived as synchronous and the highly asymmetric shape to the window. As discussed earlier (cf. Figures 1 and 2), the temporal window for auditory-visual speech recognition, where intelligibility is roughly constant, is about 250 ms (~50 ms audio lead to ~200 ms visual lead). This corresponds roughly to the resolution needed for temporally fine-grained phonemic analysis on the one hand (< 50 ms) and course-grain syllabic analysis on the other (roughly 250 ms), which we interpret as reflecting the different roles played by auditory and auditory-visual speech processing. When speech is processed by eye (i.e., speechreading), it is advantageous to integrate over long time windows of roughly syllabic lengths (200-250 ms) because visual speech cues are rather course [14]. At the segmental level, visual recognition of voicing and manner-of-articulation is generally poor [15], and while some prosodic cues are decoded at better-than-chance levels (e.g., syllabic stress, and phrase boundary location) accuracy is not very high [16]. In contrast, acoustic processing of speech is much more robust and capable of fine-grained analyses using temporal window intervals between 10-40 ms [17, 18]. What is interesting is that when acoustic and visual cues are combined asynchronously, the data suggest that whichever modality is presented first seems to determine the operating characteristics of the speech processor. That is, when visual cues lead acoustic cues, a long temporal window seems to dominate whereas when acoustic cues lead visual cues, a short temporal window dominates.

For the simple task used in the present study, one that does not require speech recognition, but simply discriminating synchronous from asynchronous auditory-visual speech inputs, the results are essentially unchanged from that observed earlier in recognition tasks. Audio lags up to approximately 200 ms are indistinguishable from the synchronous condition, at least for these speech materials. For audio-leading stimuli, asynchronies less than approximately 45 ms went unnoticed, giving a result more consistent with audio-alone experiments. Thus, unlike many psychophysical tests comparing discrimination on the one hand to identification on the other (where discrimination thresholds are far better than identification), cross-modal synchrony discrimination and speech recognition of asynchronous auditory-visual input appear to be highly related and similar in magnitude.

The asymmetry (auditory lags being less noticeable than auditory leads) appears to be an essential property of auditory-visual integration. One possible explanation makes note of the natural timing relations between audio and visual events in the real world, especially when it comes to speech. In nature, visible byproducts of speech articulation, including posturing and breath, almost always occur before acoustic output. This is also true for many non-speech events where visible movement precedes sound (e.g., a hammer moving and then striking a nail). It is reasonable to assume that any learning network (such as our brains) exposed to repeated occurrences of visually leading events would adapt its processing to anticipate and tolerate multisensory events where visual input leads auditory input while maintaining the perception that the two events are bound together. Conversely, because acoustic cues rarely precede visual cues in the real world, the learning network might become fairly intolerant and unlikely to bind acoustic and visual input where acoustic cues lead visual cues. Thus, precise alignment of audio and visual stimuli are not required for successful auditory-visual integration, but attention to the temporal order between audio and video components *is* critical.

4. Conclusions

Auditory-visual integration of speech is highly tolerant of audio and visual asynchrony, but only when the visual stimulus precedes the audio stimulus. When the audio stimulus precedes the visual stimulus, asynchrony across modality is readily perceived. This is true regardless of whether the subject's task is to recognize and identify words or syllables, or to simply discriminate which of two auditory-visual speech inputs is synchronized. This suggests that as soon as auditory-visual asynchrony is detected, the ability to integrate the two sources of information declines. We note that this is a fairly unusual outcome in psychological tests where the limits of sensitivity to a particular stimulus property (as measured by detection and discrimination) coincides with its use in higher-order decisions (e.g., recognition and identification).

The range of auditory-visual temporal asynchronies which go apparently unnoticed in speech is fairly broad (roughly -45 ms to +200 ms). It is suggested that tolerance to such a broad range arises from the distribution of naturally occurring events in the real world where visual motion typically precedes acoustic output. The functional significance of such constant exposure is to create perceptual processes that are capable of grouping auditory and visual events into a coherent, single object in spite significant temporal misalignments. For speech processing, it is not surprising, and probably fortunate, that the extent of the temporal window is roughly that of a syllable.

5. Acknowledgements

This research was supported by the Clinical Investigation Service, Walter Reed Army Medical Center, under Work Unit #00-2501 and by grant numbers DC 000792-01A1 from the National Institute on Deafness and Other Communication Disorders to Walter Reed Army Medical Center, SBR 9720398 from the Learning and Intelligent Systems Initiative of the National Science Foundation to the International Computer Science Institute, and DC 004638-01 and DC 005660-01 from the National Institute on Deafness and Other Communication Disorders to the University of Maryland. We would like to thank Dr. Steven Greenberg for his support and many fruitful discussions concerning this work. The opinions or assertions contained herein are the private views of the authors and should not be construed as official or as reflecting the views of the Department of the Army or the Department of Defense.

6. References

[1] Silipo, R., Greenberg, S., and Arai, T. (1999). "Temporal constraints on speech intelligibility as deduced from exceedingly sparse spectral representations," in

Proceedings of Eurospech 1999. Budepest, pp. 2687-2690.

[2] Stone, M.A., and Moore, B.C.J. (2003). "Tolerable hearing aid delays. III. Effects of speech production and perception of across-frequency variation in delay," Ear Hear. 24, 175-183.

[3] Sumby, W.H., and Pollack, I. (1954). "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am. 26, 212-215.

[4] Grant, K.W., and Braida, L.D. (1991). "Evaluating the Articulation Index for audiovisual input," J. Acoust. Soc. Am. 89, 2952-2960.

[5] Grant, K.W., and Seitz, P.F. (1998). "Measures of auditory-visual integration in nonsense syllables and sentences," J. Acoust. Soc. Am. 104, 2438-2450.

[6] Grant, K.W., and Greenberg, S. (2001). "Speech intelligibility derived from asynchronous processing of auditory-visual information" in *Proceedings Auditory-Visual Speech Processing (AVSP 2001)*, Scheelsminde, Denmark, September 7-9, 2001.

[7] van Wassenhove, V., Grant, K.W., and Poeppel, D. (2001). Timing of Auditory-Visual Integration in the McGurk Effect. Presented at the Society of Neuroscience Annual Meeting, San Diego, CA, November, 488.

[8] McGrath, M., and Summerfield, Q. (1985). "Intermodal timing relations and audio-visual speech recognition by normal-hearing adults," J. Acoust. Soc. Am. 77, 678-685.

[9] Pandey, P.C., Kunov, H., and Abel, S.M. (1986). "Disruptive effects of auditory signal delay on speech perception with lipreading," J. Aud. Res. 26, 27-41.

[10] Massaro, D.W., Cohen, M.M., and Smeele, P.M. (1996). "Perception of asynchronous and conflicting visual and auditory speech.," J. Acoust. Soc. Am. 100, 1777-1786.

[11] Grant, K.W., and Seitz, P.F. (2000). "The use of visible speech cues for improving auditory detection of spoken sentences," J. Acoust. Soc. Am. 108, 1197-1208.

[12] Grant, K.W. (2001). "The effect of speechreading on masked detection thresholds for filtered speech," J. Acoust. Soc. Am. 109, 2272-2275.

[13] Institute of Electrical and Electronic Engineers (1969). *IEEE recommended practice for speech quality measures*. IEEE, New York.

[14] Seitz, P.F., and K.W. Grant, K.W. (1999). "Modality, perceptual encoding speed, and time-course of phonetic information," AVSP'99 Proceedings, Aug, 7-9, 1999, Santa Cruz, CA.

[15] Grant, K.W., Walden, B.E., and Seitz, P.F. (1998). "Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition , sentence recognition, and auditory-visual integration," J. Acoust. Soc. Am. 103, 2677-2690.

[16] Grant, K.W., and Walden, B.E. (1996). "The spectral distribution of prosodic information," J. Speech Hear. Res. 39, 228-238.

[17] Stevens, K.N., and Blumstein, S.E. (1978). "Invariant cues for place of articulation in stop consonants," J. Acoust. Soc. Am., 64, 1358-1368.

[18] Greenberg, S., and Arai, T. (2001). The relation between speech intelligibility and the complex modulation spectrum," in Proceedings of the 7th European Conference on Speech Communication and Technology (*Eurospeech-2001*), Aalborg, Denmark, September, 473-476.