SPEECH INTELLIGIBILITY DERIVED FROM ASYNCHRONOUS PROCESSING OF AUDITORY-VISUAL INFORMATION

Ken W. Grant[1] and Steven Greenberg[2]

[1]Army Audiology and Speech Center
Walter Reed Army Medical Center, Washington, D.C. 20307

[2]1nternational Computer Science Institute
1947 Center Street, Berkeley, CA 94704 USA

ABSTRACT

The current study examines the temporal parameters associated with cross-modal integration of auditory-visual information for sentential material (Harvard/IEEE sentences). The speech signal was filtered into 1/3-octave channels, all of which were discarded (in the primary experiment) save for a low-frequency (298-375 Hz) and a high-frequency (4762-6000 Hz) band. The intelligibility of this audio-only signal ranged (depending on the listener) between 9% and 31% (of the key words correct) for nine normal-hearing subjects. Visual-alone presentation of the same material ranged between 1% and 22% intelligibility. When the audio and video signals are combined and presented in perfect synchrony, intelligibility climbs to an *average* of 63%. The audio and video signals were systematically desynchronized in symmetrical fashion up to a maximum onset asynchrony of 400 ms. When the audio signal leads the video, intelligibility declines appreciably for even the shortest asynchrony of 40 ms, falling to an asymptotic level of performance for asynchronies of ca. 120 ms and longer for most subjects. In contrast, when the video signal leads the audio, intelligibility remains relatively stable for onset asynchronies up to 160-200 ms, and *in some instances* may actually *improve*. Hence, there is a marked asymmetry in the integration of audio and visual information that has important implications for sensory-based models of auditory-visual speech processing.

1. INTRODUCTION

Speechreading has been shown to improve speech understanding in difficult communication environments such as those associated with background noise, reverberation and hearing loss. The benefit provided by the integration of auditory and visual speech cues relative to audio information alone arises from at least two separate mechanisms. First, independently derived speech information from the eye and the ear are combined early in the decision process so that what is ambiguous within a single modality is rendered much less so in concert with the other [14][21]. Acoustic cues for place of articulation (e.g., [p] vs. [t] vs. [k]) are often weakly represented or entirely absent in the presence of noise or reverberation, while they are robustly encoded in the visual modality. In contrast, voicing cues (e.g., [p] vs. [b]) are robust acoustically, but are difficult to discern in the visual signal. This complementary nature of bimodal speech cues is an important factor underlying the superiority of auditory-visual speech processing relative to auditory-only recognition [8][14].

A second mechanism underlying the robust nature of auditory-visual speech recognition robust is the observer's ability to track the low-frequency (< 30 Hz) modulation pattern shared in common between the visible portion of articulatory motion (derived from the lips, jaw, tongue tip, eye brows, etc.) and the energy envelope of the acoustic signal (particularly in the mid-to-high-frequency region of the spectrum [5][7]), permitting the observer to group the acoustic and visual streams into a single, coherent object.

When full-spectral-bandwidth speech is presented in background noise, auditory-recognition performance declines in direct proportion to the signal-to-noise ratio [1]. When visual speech cues are presented concurrently (and in synchrony) with the noisy acoustic speech recognition dramatically improves. It has been shown by Grant and colleagues that much of this improvement can be accounted for by the complementary nature of low-frequency acoustic cues and speechreading [8][9]. Analysis of phonetic information contained in band-limited speech from different regions of the spectrum suggests that speechreading is largely redundant with the portion of the speech spectrum above 800 Hz.

Corroboration of an association between speechreading and mid-frequency acoustic cues is found in separate studies pertaining to the correlation between the area of mouth opening and the acoustic amplitude envelope derived from low-, mid- and high-frequency speech bands [5][7]. The correlation between lip-area movement and the acoustic amplitude envelope is greatest for mid-frequency bands (800-2200 Hz) and least for low-frequency channels (below 800 Hz).

Because speechreading and audio-speech cues are processed in highly integrated fashion by the brain [4][18][20] it is of interest to ascertain whether the time constraints characteristic of speech processing derived from concurrent auditory and visual information are similar to those pertaining when only acoustic cues are present.

Greenberg and associates have recently shown that the auditory system appears to be exquisitely sensitive to even small amounts of asynchrony across the spectrum when the speech signal is composed exclusively of acoustic information [11][19]. In their studies the speech signal (composed of sentences from the TIMIT corpus) was par-

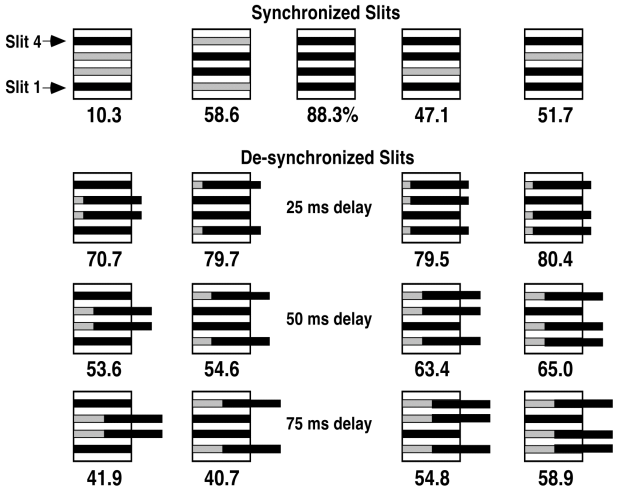

Figure 1. Intelligibility (percent words correct) of sparse spectral sentences containing four narrow-band (1/3 octave) channels as a function of slit asynchrony. Note the relatively symmetrical decline in intelligibility associated the central slits leading or lagging the lateral slits. 16 subjects. Adapted from [12].

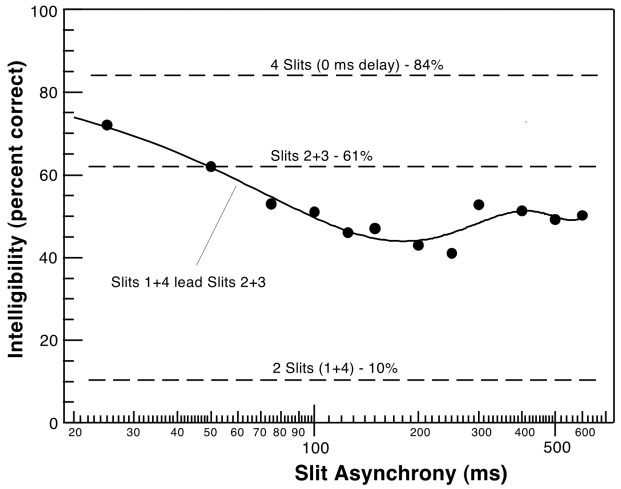

Figure 2. Intelligibility (percent words correct) of sparse spectral sentences containing four narrow-band (1/3 octave) channels as a function of slit asynchrony. Note that intelligibility goes below baseline (slits 2+3_ when the slit asynchrony exceeds 50 ms). 27 subjects. Adapted from [19].

titioned into 1/3-octave channels ("slits") and most of the spectrum discarded. In the conditions of greatest relevance to the current study, four slits were concurrently presented, (1) 298-375 Hz, (2) 750-945 Hz, (3) 1890-2381 Hz and (4) 4762-6000 Hz, either synchronously (84% or 88% intelligibility, depending on the specific set of listeners), or asynchronously. The conditions of primary interest (for the current study) are when the central slits, as a pair (bands 2 *and* 3) or singly (band 2 *or* 3), lead or lag the other channels (cf. Figure 1). Desynchronization of the central slits relative to the lateral bands has a dramatic effect on intelligibility when onset asynchrony is as short as 50 ms.Word recognition declines to a baseline associated with presentation of the central slits by themselves (Figure 1). In most instances the order of asynchrony (i.e., whether the central slits led or lagged the lateral slits) had a symmetrical impact on intelligibility (Figure 1).

Further increases in onset asynchrony among slits, up to ca. 250 ms, resulted in a progressive decline in intelligibility (Figure 2) to a level well below baseline (i.e., there was substantial interference between the lateral and central bands), but above performance associated with the lateral bands alone (10%). For onset asynchronies longer than 250 ms there appears to have been a slight *improvement* in intelligibility for most listeners.

The integration of audio-visual speech information appears to obey different time constraints than those pertaining to audio-only cues. It has been demonstrated that when speechreading cues are presented asynchronously with noisy, full-bandwidth speech (0.1-8.5 kHz), most subjects are relatively insensitive to *acoustic delays* up to ca. 200 ms [6]. This result is in seeming contrast to the much greater sensitivity to cross-channel *acoustic* asynchronies observed by Greenberg and colleagues [11][19]. One possible explanation for the different effects observed in cross-modality asynchrony experiments relative to audio-only, cross-channel asynchrony experiments is that the cross-modality experiments used full-spectral-bandwidth speech as opposed to spectrally sparse signals used in the experiments described in [12][19]. Greenberg and colleagues argue that when broadband speech is presented in quiet, intelligibility is effectively greater than 100% due to the high degree of redundancy in the signal (i.e., there is a ceiling effect). In support of this conjecture, they have demonstrated that full-bandwidth, speech does, in fact, appear to be relatively impervious to spectral desynchronization [3][11]. Thus, to delineate the "true" sensitivity of the brain to spectral (and auditory-visual) asynchrony it may be necessary to use spectrally sparse speech.

In summary, the studies reviewed above suggest the following scenario. Speechreading provides information comparable to that of the mid-frequency portion of the acoustic speech spectrum. Lip movement appears to be highly correlated with the modulation pattern of acoustic energy in circumscribed regions of the spectrum. The highest correlations are observed between lip movement and acoustic-amplitude envelopes from spectral bands above 800 Hz. Experiments using spectrally sparse speech signals indicate that intelligibility declines monotonically (over a range between 25 and 250 ms) when mid-frequency bands are desynchronized relative to low- and high-frequency bands.

The present study employs an experimental paradigm similar to that used by Greenberg and colleagues in their study of spectrally sparse speech [12][19]. Instead of desynchronzing four spectral channels relative to each other, the visual component of the speech signal was delayed or advanced relative to the lateral acoustic bands (298-375 Hz and 4762-6000 Hz) in order to ascertain the sensitivity of auditory-visual speech integration under conditions where there is a minimum amount of redundant information with which to derive intelligibility.

2. METHODS

2.1 Subjects

Nine normal-hearing subjects between the ages of 22 and 57 years old (mean = 38.1 years) participated in the study. Subjects were recruited from eligible staff, patients, and dependents associated with the Army Audiology and Speech Center, Walter Reed Army Medical Center, and were screened binaurally to assure pure-tone air conduction thresholds of 20 dB HL (or better) at audiometric test frequencies between 0.25 and 4 kHz and 30 dB HL (or better) at frequencies between 6 and 8 kHz [2]. All subjects were native speakers of American English, with normal or corrected-to-normal vision (i.e., visual acuity equal to or better than 20/30 as measured with a Snellen chart). Subjects were paid $10.00 per hour for their participation. The total amount of time required for each subject to complete the experimental protocol was approximately ten hours, partitioned into five, two-hour sessions.

2.2 Procedures

The stimulus presentation structure was based on a randomized block design with repeated measures. Each experimental condition involved a different speech recognition situation, including listening-alone (audio signal only), speechreading alone (visual signal only) and listening while speechreading (audio and visual signals presented in tandem). Each subject was tested under all conditions. The subject's task was to identify the sequence of words in sentences presented under these different situations.

Speech materials consisted of sentences from the IEEE/ Harvard sentence corpus [13] spoken by a female speaker of general American English. The IEEE sentences were chosen because they are constructed to have roughly equal intelligibility across lists and have approximately the same duration, number of syllables, grammatical structure and intonation. Each sentence is composed of ca. 7 to 12 words, with five key words identified for the purposes of scoring. Four of the key words are monosyllabic, while the fifth is bisyllabic in form.

The sentences were video-taped, with the audio and visual signals transferred to an optical disk recorder (Panasonic TQ-3031F). The audio portion of each sentence was digitized (16 bit A/D, 20 kHz sampling rate), normalized in amplitude and stored on a personal computer for subsequent processing and presentation.

Each of the sentences was filtered into two narrow spectral bands (approximately 1/3-octave in bandwidth) with attenuation slopes exceeding 100 dB/octave. The passband of the low-frequency channel was 298-375 Hz and that of the high-frequency channel was 4762-6000 Hz. In a previous study using acoustic-only signals [11], such stimuli resulted in ca. 10% intelligibility for sentential material of slightly greater difficulty (TIMIT corpus) than that used in the current experiment. The low intelligibility of the *audio-alone* condition (denoted as A_{iooi}) was important to insure that the *audio-visual* intelligibility did not exceed 90% correct (in order to preclude the possibility of "ceiling" effects). Two additional spectral slits were used in two of the audio-alone conditions. In one condition the mid-frequency bands (750-945 Hz and 1890-2381 Hz) were presented by themselves (A_{oiio}), and in the other condition they were presented in tandem with the lateral slits (A_{iiii}). These two audio-alone conditions were used to ascertain intelligibility under conditions comparable to those employed in the earlier audio-only experiment of Greenberg and colleagues [11] and as control conditions for the present study.

Subjects were tested binaurally using headphones (Beyer Dynamic DT770) in a sound-treated booth. Processed sentences were presented for identification under nineteen combinations of auditory, visual (speechreading) and audio-visual conditions. Three conditions were of the audio-alone variety, one was visual-alone (speechreading only) and fifteen were audio-visual with varying amounts of asynchrony between audio and visual signals. For each of these nineteen conditions three lists of ten sentences (150 key words), each, were used. No sentence was repeated to the subject at any point during the experiment and no feedback was provided. The three audio-alone conditions consisted of listening to the sentences presented through the spectral slits (either A_{iooi}, A_{oiio} or A_{iiii}). The video-alone (speechreading) condition consisted of well-lighted motion views of the talker's face and head displayed on a 19-inch color monitor (SONY PVM 2030) positioned approximately 5 feet from the subject. At this distance, the video image of the talker's face was life-size. The 15 auditory-visual conditions consisted of synchronous and asynchronous presentations of audio (A_{iooi}) and video (speechreading) signals. Acoustic delays ranging between - 400 ms (audio signal leads) and + 400 ms (video signal leads) were tested. In the proximity of the synchronous audio-visual condition (0 ms delay), a fine-grained exploration of temporal asynchrony was performed using 40-ms steps. For long delays (200, 300 and 400 ms) a coarser granularity of time steps was used. The conditions in which the audio component *leads* the video are denoted in terms of *negative* delay for purposes of illustration.

The subject's task was to identify the words contained in the sentences. Subjects responded both verbally and in writing. Subjects were informed that all of the speech material was composed of meaningful sentences with proper grammar and syntax. Guessing was encouraged in cases where the sentences were perceptually ambiguous. The experimenter monitored each subject's verbal responses and scored the intelligibility in terms of the number of correctly repeated key words per sentence. At the end of each 10-sentence list, the total number of key words correctly repeated was tallied and entered into a computer. All subjects received the same three lists per condition. The order of conditions (3 audio, 1 visual, and 15 audio-visual) was randomized separately for each subject.

3. RESULTS

Table 1 summarizes the intelligibility data associated with the audio-visual, audio-alone and visual-alone conditions

	Audio Signal Leading (ms)							Sync	Video Signal Leading (ms)							Audio-Alone			Video
Sub	-400	-300	-200	-160	-120	-80	-40	0	40	80	120	160	200	300	400	iooi	oiio	iiii	V
JJL	23.3	48.0	41.3	43.3	50.7	62.7	70.7	72.7	84.0	84.0	80.7	77.3	74.0	61.3	30.7	30.0	69.3	98.0	12.0
LDH	26.0	25.3	19.3	26.7	28.7	50.0	54.0	77.3	74.7	66.7	65.3	62.7	65.3	36.7	18.7	12.0	50.0	83.3	17.0
SMT	25.3	37.3	39.3	53.3	37.3	59.3	55.3	72.0	71.3	70.7	66.0	64.7	79.3	40.0	40.7	13.3	48.7	88.0	19.0
DWC	24.0	38.0	24.0	48.7	41.3	52.0	54.7	64.0	70.7	70.7	67.3	64.7	68.7	44.7	24.0	12.0	52.7	82.7	22.0
MRM	10.0	15.3	18.7	44.0	35.3	46.7	41.3	66.0	56.7	66.7	74.0	62.7	66.0	29.3	16.7	31.3	80.7	96.7	4.0
JEN	18.0	15.3	21.3	18.7	21.3	37.3	31.3	64.7	60.0	63.3	67.3	46.0	57.3	29.3	20.0	8.7	45.3	93.3	6.0
DGK	10.7	22.7	22.7	28.7	32.0	39.3	44.7	62.0	56.0	60.7	66.7	54.7	55.3	38.7	19.3	12.7	46.7	92.7	10.0
MPR	14.7	28.0	20.0	30.7	19.3	41.3	42.0	46.7	52.7	52.7	55.3	43.3	50.7	37.3	26.7	26.7	57.3	95.3	1.0
KEG	15.3	19.3	23.3	12.7	17.3	26.0	24.7	38.0	44.7	46.0	44.7	28.7	35.3	18.7	11.1	20.7	62.0	88.7	7.0
Mean	18.6	27.7	25.6	34.1	31.5	46.1	46.5	62.6	63.4	64.4	65.3	56.1	61.3	37.3	23.1	18.6	57.0	91.0	10.9
S.D.	6.3	11.3	8.6	14.0	11.0	11.5	13.9	12.7	12.5	11.0	10.3	14.6	13.3	11.8	8.7	8.7	11.8	5.6	7.2

Table 1. Intelligibility (in terms of the percent of key words correctly recognized) associated with integration of acoustic and visual signals of IEEE/Harvard sentence material for nine subjects, along with intelligibility of audio-only and video-alone signals.

for all nine subjects. The intelligibility data for the audio-visual integration conditions (only) are plotted in Figure 3.

For the audio-alone conditions, intelligibility ranges between 9% and 31% for the lateral (A_{iooi}) slits ($\overline{X}$=18.6%), between 45% and 81% for the central (A_{oiio}) slits ($\overline{X}$=57%) and between 83% and 98% for all four (A_{iiii}) slits combined ($\overline{X}$=91%). The mean intelligibility is comparable to that reported by Greenberg and colleagues [12][19] for a different corpus of sentence materials. The lateral slits alone (A_{iooi}) are recognized with somewhat higher accuracy compared to the TIMIT sentences, perhaps reflecting the slightly more predictable nature of the word sequences in the IEEE/Harvard sentences. Otherwise, the mean intelligibility of the two sets of material are similar for comparable conditions. Intelligibility for the video-alone condition varies between 1% and 22% ($\overline{X}$=11%).

Intelligibility associated with the audio-visual conditions varies a great deal, both as a function of A/V onset asynchrony and across subjects. However, in spite of large individual differences the pattern of performance is quite similar across subjects. In general, integration of auditory

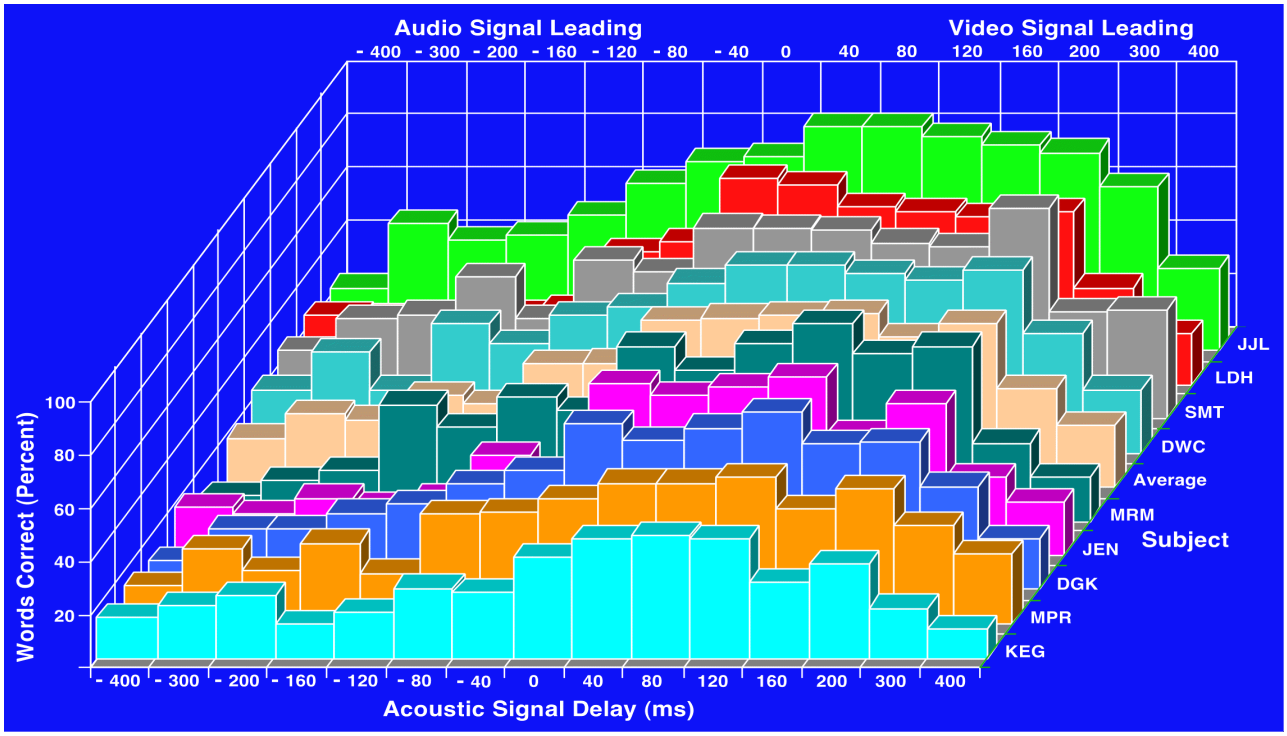

Figure 3. Intelligibility (percent of key words correct) of sentential material using concurrent acoustic and visual presentation as a function of intermodal onset asynchrony for each of nine subjects (plus the mean). The acoustic signal consists of two 1/3-octave slits, one with energy between 298 and 375 Hz, the other with energy between 4792 and 6000 Hz.

and visual speech scores is quite good when the visual signal leads the audio signal but is less efficient when the audio signal leads the visual signal (Figure 3).

For the A/V synchronous condition (0-ms delay) intelligibility ranges between 38% and 77% ($\overline{X}$=63%). For all but two subjects intelligibility is greater than 60%.

When the audio signal leads the video all subjects exhibit an appreciable decline in recognition performance relative to the synchronous condition for asynchronies as short as 40 or 80 ms (Table 1, Figures 3 and 4). For most subjects there is a relatively progressive decline in intelligibility as the audio signal increasingly leads the video. When the audio signal leads the video by 400 ms the average intelligibility is approximately 20% and equals that of the audio-alone condition (Figure 4). This progressive decline in word recognition as a function of onset asynchrony is comparable in magnitude to that associated with spectral asynchrony for audio-alone signals (Figure 2), particularly if the higher intelligibility scores of the latter condition are taken into account (i.e., the pattern of *relative* decline in intelligibility is similar).

When the video signal leads the audio, a different pattern emerges. In contrast to the progressive decline in intelligibility observed when the audio signal leads the video, word recognition appears to be stable (and in some instances may actually improve) across a wide range of bimodal asynchrony (Figure 4). Average intelligibility does not appreciably decline below the level associated with the synchronous A/V condition until the video leads the audio signal by more than 200 ms. The stability of A/V integration (as assessed by intelligibility) appears manifest for all 9 subjects, despite significant differences in the *absolute* level of word-recognition performance (Figure 3).

For both auditory integration across spectral regions (Figure 2) and A/V integration (Figure 4), word recognition performance often exceeds what one might expect from the simple addition of multiple sources of *independent* information. In terms of acoustic, spectral integration, the predicted performance for the *average* four-slit (A_{iiii}) condition is 64% based on the "product of errors" rule, given the *average* intelligibility associated with the A_{iooi} (18.6%) and A_{oiio} (57%) conditions. This predicted performance level was exceeded (for the studies described in [12] and [19]) only when the slits were desynchronized by less than 50 ms (Figures 1 and 2). For A/V integration the predicted level of *average* intelligibility is 27.5% (given an *average* A_{iooi} recognition score of 18.6% and *average* video-alone performance of 10.9%). In contrast to audio-alone spectral integration, A/V integration performance exceeds the predicted level of intelligibility for a broad range of A/V asynchronies (-160 ms through 300 ms).

4. DISCUSSION

The current study demonstrates that the ability to integrate auditory and visual information under conditions of bimodal asynchrony is highly dependent on whether the audio signal leads or lags the video. When the audio leads, A/V integration declines precipitously for even small

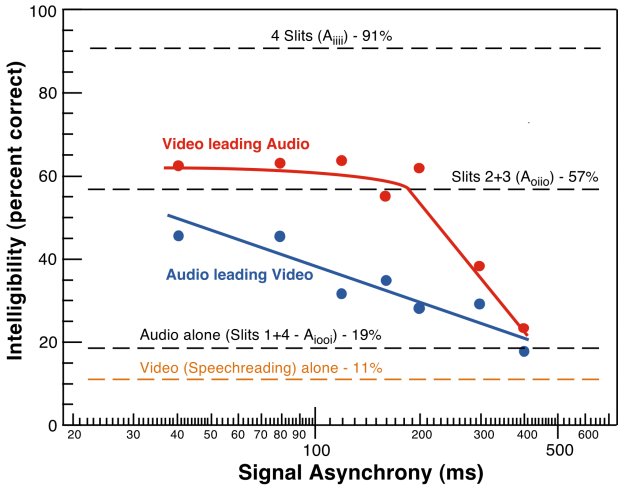

Figure 4. *Average* intelligibility (for 9 subjects) associated with audio-visual speech recognition as a function of bimodal signal asynchrony. The audio-leading-video conditions are marked in blue, the video-leading-audio conditions shown in red. Baseline audio-only conditions are marked in black, dashed lines, and the video-alone condition is shown in orange.

onset asynchronies, similar in pattern to intelligibility under conditions of spectral asynchrony for acoustic-only signals [12][19]. When the video leads the audio, A/V integration remains relatively stable for onset asynchronies as long as 200 ms, similar in pattern to those observed for a variety of speech materials [6][17]. Our data also suggest that certain subjects experience a slight, but pronounced *gain* in intelligibility when the video signal leads the audio by 40-200 ms relative to the A/V-synchronous condition (Figure 3).

Previous studies of AV integration and auditory-visual asynchrony have focused only on conditions where the visual signal precedes the audio signal [6][16][17]. The one previous study which explored conditions where the visual signal lagged the audio signal [15] used nonsense syllables of limited entropic capacity (there were only four A/V-congruent stimuli) which may have potentially masked the full extent of audio-visual integration. Such a limited set of stimuli may explain the relatively small effect of audio-leading asynchrony on speech recognition when the two modalities are desynchronized by as much as 267 ms.

The current study also differs from previous work in using spectral slits as the audio signal. In the past, the audio component consisted of either full-spectrum speech embedded in background noise [6][17], a rectangular pulse train whose rate was modulated by the talker's fundamental frequency [16], or nonsense CV syllables (embedded within a McGurk-effect paradigm) [15]. The advantage of using spectral slits for assessment of A/V integration pertains to the complementary nature of the audio and video signals used. Neither form of signal presented by itself results in more than 31% of the key words recognized (for the best subject) and the average intelligibility is far lower (11% for the video-alone and 19% for

audio-alone conditions). When the audio and video signals are combined, integration increases dramatically (to 84% for the best subject), thus providing a large dynamic range with which to assess the effects of bi-modal asynchrony on A/V integration.

It has been shown in the past that speechreading cues are primarily associated with place-of-articulation features [8][14][21] and that it is the mid-frequency portion of the audio spectrum (ca. 2 kHz) that pertains to such video-derived information [5][7][8]. The current study is consistent with these earlier findings in that combining audio information derived from the low- and high-frequency channels with the video signal results in an intelligibility gain that is similar in pattern (if not quite in magnitude) to that observed when all four spectral slits are combined (cf. Figures 1, 2 and 4). In this sense the audio and video signals used in the current experiment may be of a more complementary form than signals used in earlier studies.

The current data do not allow us to determine with precision why A/V integration is more robust for conditions where the video signal leads the audio (relative to the converse). It is possible that this asymmetry in bi-modal integration pertains to the level of analytical abstraction associated with audio- and video-alone signals. For the former, spectral integration germane to speech processing is likely to focus on deriving phonetic-segment information with time constants between 40 and 120 ms [10]. In contrast, video-alone cues are more likely to pertain to syllabic units (whose average duration, in American English, is 200 ms [10]), given the coarse nature of the phonetic information associated with this modality (presented in isolation). The modality that leads in the asynchronous conditions may thus determine the level of abstraction (and hence the time constant) over which bi-modal processing proceeds.

5. ACKNOWLEDGEMENTS

This research was supported by a grant from the Learning and Intelligent Systems Initiative of the National Science Foundation. We thank Rosario Silipo for assistance in creating the audio stimuli.

We thank the subjects participating in this research, all of whom provided written informed consent prior to beginning the study. The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the Department of the Army or the Department of Defense.

6. REFERENCES

[1] American National Standards Institute "Methods for the Calculation of the Articulation Index". ANSI S3.5-1969, American National Standards Institute, New York, 1969.

[2] American National Standards Institute "Specifications for Audiometers". ANSI S3.6-1989, American National Standards Institute, New York, 1989.

[3] Arai, T. and Greenberg, S. "Speech intelligibility in the presence of cross-channel spectral asynchrony," *Proc. IEEE ICASSP*, pp. 933-936, 1998.

[4] Calvert, G. A., Bullmore, E. T., Brammer, M.J., Campbell, R., Williams, S. C., McGuire, P. K., Woodruff, P. W., Iversen, S. D., and David, A.S. "Activation of auditory cortex during silent speechreading," *Science*, 276: 593-596, 1997.

[5] Grant, K.W. "Effect of speechreading on masked thresholds for filtered speech," *J. Acoust. Soc. Am.*, 109: 2272-2275, 2001.

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

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

[8] Grant, K.W., and Walden, B.E. "Evaluating the Articulation Index for auditory-visual consonant recognition," *J. Acoust. Soc. Am.*, 100: 2415-2424, 1996.

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

[10] Greenberg, S. "Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, 29: 159-176, 1999.

[11] Greenberg, S. and Arai, T. "Speech intelligibility is highly tolerant of cross-channel spectral asynchrony," *Proc. Joint Meeting Acoust. Soc. Am. and Int. Cong. Acoust.*, pp. 2677-2678, 1998.

[12] Greenberg, S., Arai, T. and Silipo, R. "Speech intelligibility derived from exceedingly sparse spectral information," *Proc. Int. Conf. Spoken Lang. Proc.*, pp. 2803-2806, 1998.

[13] IEEE "IEEE recommended practice for speech quality measurements," Institute of Electrical and Electronic Engineers, New York, 1969.

[14] Massaro, D. *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry.* Hillsdale, NJ: Lawrence Earlbaum Associates, 1987.

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

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

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

[18] Sams, M., Aulanko, R., Hamalainen, M., Hari, R., Lounasmaa, O.V., Lu, S.T., and Simola, J. (1991). "Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex," *Neuroscience Letters*, 127: 141-145.

[19] Silipo R., Greenberg, S. and Arai, T. "Temporal constraints on speech intelligibility as deduced from exceedingly sparse spectral representations," *Proc. Eurospeech*, pp. 2687-2690, 1999.

[20] Stein, B.E., and Meredith, M.A. *The Merging of the Senses*. Cambridge, MA: MIT Press, 1993.

[21] Summerfield, Q. "Some preliminaries to a comprehensive account of audio-visual speech perception," In *Hearing by Eye: The Psychology of Lip-Reading*, B. Dodd and R. Campbell (eds.). Hillsdale NJ: Lawrence Erlbaum Associates, pp. 3-52, 1987.