Spectro-Temporal Interactions in Auditory and Auditory-Visual Speech Processing

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

[1]Army Audiology and Speech Center, Walter Reed Army Medical Center, Washington, DC
[2]Speech Institute, Oakland, CA

grant@tidalwave.net

Abstract

Speech recognition often involves the face-to-face communication between two or more individuals. The combined influences of auditory and visual speech information leads to a remarkably robust signal that is greatly resistant to noise, reverberation, hearing loss, and other forms of signal distortion. Studies of auditory-visual speech processing have revealed that speechreading interacts with audition in both the spectral and temporal domain. For example, not all speech frequencies are equal in their ability to supplement speechreading, with low-frequency speech cues providing more benefit than high-frequency speech cues. Additionally, in contrast to auditory speech processing which integrates information across frequency over relatively short time windows (20-40 ms), auditory-visual speech processing appears to use relatively long time windows of integration (roughly 250 ms). In this paper, some of the basic spectral and temporal interactions between auditory and visual speech channels are enumerated and discussed.

1. Spectral Interactions in Auditory-Visual Speech Processing

Under most conditions, auditory-visual speech recognition has been shown to be superior to either auditory or visual speech recognition alone, even for listeners with average to below-average speechreading skills. Moreover, the superior intelligibility scores obtained under auditory-visual speech conditions are seldom achieved with the use of hearing aids or through signal processing techniques aimed at improving the clarity or speech-to-noise ratio of the speech signal [21]. Over the years, attempts have been made to understand the various factors that enable listeners to combine optimally auditory and visual modalities to enhance their speech comprehension [5,8-11]. Much of this work has focused on an abstract level of linguistic information. That is, "What kind of information could subjects' extract from visual-alone and auditory-alone speech cues?" "How well do subjects integrate this information?" "Are semantic and morpho-syntactic constraints used to varying degrees by subjects in resolving ambiguities in the message?" More recently [7,13,18], studies have investigated the temporal window of integration for both cross-spectral (auditory-only) and cross-modal (auditory-visual) speech perception. In the sections below, we review some of this work and discuss implications for more comprehensive models of auditory-visual speech integration.

1.1. Spectral weighting in auditory-visual speech processing

It is well established that speech recognition can be dramatically improved by the addition of visual cues. For example, typical, untrained, normal-hearing listeners can recognize nonsense syllables at roughly 90% correct by listening alone at a speech-to-noise ratio of 0 dB, whereas this same level of performance is achieved at a speech-to-noise ratio of roughly –8 dB under auditory-visual conditions [5,9]. Such "real-world" advantages are difficult to achieve using noise-suppression, signal-processing algorithms or microphone arrays in wearable devices.

Because of the obvious advantages of auditory-visual speech input for improved speech communication, it has been an important goal to try and predict speech recognition performance in situations where visual cues are available. The ANSI (1969) Standard for Calculating the Articulation Index [1] represents one way to accomplish this goal. However, the visual "correction" curve provided in the standard does not take into account the possibility that frequency bands of equal auditory intelligibility may not be equivalent for audiovisual speech perception. For example, a high-frequency band of speech (Figure 1, filter condition 6) can have significantly *greater* auditory intelligibility than a low-frequency band of speech (e.g., filter condition 1), and yet result in significantly *poorer* auditory-visual intelligibility.

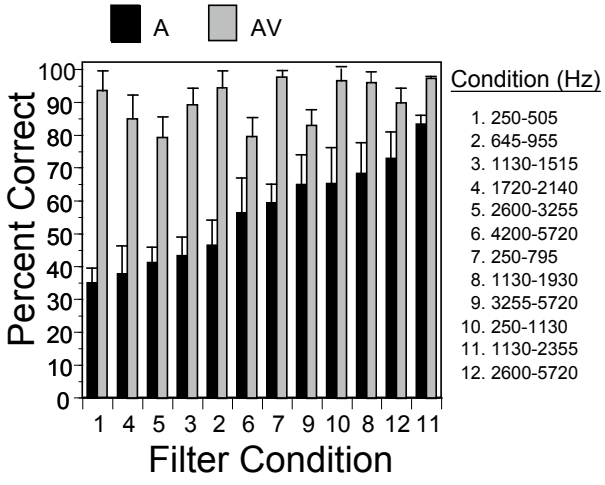

Figure 1. Auditory and auditory-visual nonsense syllable recognition under different filtered-speech conditions. Reprinted from [9].

1.2. "Designer" auditory signals and speechreading

The finding that low-frequency speech signals can be a particularly effective supplement to speechreading has been demonstrated in a number of studies [4,17]. For example, amplitude-envelope signals derived from low-frequency speech bands, or frequency-modulated signals matched to the voice fundamental frequency of the talker, have been shown to dramatically improve speech communication (over speechreading alone) to approximately 70-80% of normal-listening performance in quiet, even though these acoustic signals, when presented without speechreading, had near zero intelligibility [4].

Interest in delineating the various factors that enabled such high levels of speech communication with acoustic signals that were themselves incomprehensible was fostered by the development of cochlear implants. The input to the electrode arrays in these devices were primarily amplitude envelopes derived from different spectral bands of speech. This led to a host of questions, such as, (1) "Which combination of envelope signals were the best for understanding speech?", (2) "How should these envelopes be extracted?", and (3) "What is the best carrier wave to use?" Additionally, the modulated carrier wave can be "smoothed" to varying degrees to limit the range of modulation frequencies conveyed by the envelope. Figure 2 shows that, for a fixed smoothing filter of 100 Hz, envelopes derived from an octave band of speech centered at 500 Hz provided significantly greater benefit to speechreading than did envelopes derived from octave bands at 1600 or 3150 Hz, or even from the wideband speech signal [6]. Further experiments showed amplitude-envelope signals derived from high-frequency regions of speech were useful only when the envelope was "smoothed" with lowpass filters in excess of 100 Hz, whereas envelopes derived from lower frequency analysis bands were less sensitive to changes in smoothing filters down to 30 Hz. One possible explanation for this result is that the relatively broad lowpass filters used to smooth the high-frequency amplitude-envelope signals allowed fundamental frequency information to be perceived in the form of envelope ripple, or periodicity pitch. In other words, envelope cues derived from high-frequency speech bands became valuable supplements to speechreading only when low-frequency information (i.e., fundamental frequency) became available.

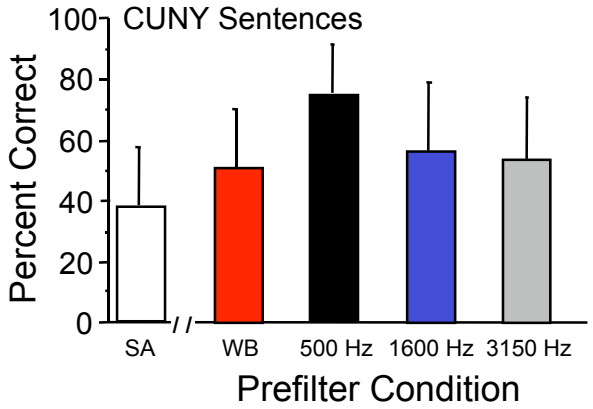

Figure 2. Speech intelligibility for speechreading alone (SA) and for speechreading supplemented by single-band envelopes derived from octave bands centered at 500, 1600, or 3150 Hz (Prefilter Condition) or from wideband (WB) speech. Carrier signals were tones centered in the analysis band or noise in the case of WB. Adapted from [6].

1.3. The "Information Redundancy" hypothesis

One way to explain the auditory-visual benefits in intelligibility observed with so-called "designer" acoustic signals (signals with near zero auditory intelligibility) is through an analysis and comparison of speech features conveyed by the two unimodal sources. In general, if the auditory and visual inputs convey redundant information, the benefit will be substantially smaller than if the two sources convey complementary information [3,11,15]. For example, consider two acoustic signals, one which transmits place-of-articulation information perfectly and one which transmits voicing information perfectly. For consonant recognition, the signal that conveys only place cues will be far more intelligible than the signal that conveys only voicing, and yet, auditory-visual intelligibility will show the reverse pattern, with the voicing signal providing much more benefit (relative to speechreading alone) than the place signal [3,15].

Although this general approach to understanding auditory-visual benefit has been demonstrated in a number of studies to be essentially correct, it is severely limited in its ability to *predict* auditory-visual intelligibility without first making extensive speech recognition measures under auditory-alone and visual-alone conditions. Further, it is not so simple to identify the range of speech features that should be considered in determining the degree of redundancy between auditory and visual channels. In connected speech, for example, *suprasegmental* cues can play a significant role in determining speech intelligibility along with *segmental* cues to consonant and vowel recognition. Because the spectral distribution of suprasegmental cues across the speech frequency range can be substantially different from segmental cues [10], it becomes necessary to consider the relative weight of each cue as it apples to overall intelligibility. To our knowledge, this would require a much more comprehensive model of speech intelligibility than exists today. But more importantly, even if a set of relevant features for connected speech could be identified and the degree of redundancy across modalities determined, the concept of informational redundancy is phenomenological at best and offers little or no guidance as to the mechanisms involved in auditory-visual integration. Specifically, how does speechreading interact with audition to "turn down" the noise or reverberation, or to minimize the effects of a hearing loss? These questions are addressed briefly in the remaining sections.

2. Spectro-Temporal Interactions in Auditory and Auditory-Visual Speech Processing

Studies aimed at determining the temporal window for integration of auditory and visual speech cues, or for integration of auditory cues from different spectral regions help to set limits on models of speech intelligibility (auditory and auditory-visual) in terms of where and how information from multiple sources come together [2,7,13,16,18]. A detailed understanding of the temporal integration window for speech can have practical implications as well, such as determining the maximum amount of time available to perform advanced signal processing on the acoustic signal (e.g., noise reduction, feature extraction, automatic speech recognition) before constraints on cross-spectral or cross modal alignment harm intelligibility [16,20].

2.1. Spectro-temporal asynchrony in auditory speech processing

In recent studies of the effects of spectro-temporal asynchrony on auditory speech intelligibility, Greenberg and colleagues [2,13,18] have shown that the auditory system is extremely sensitive to changes made in the relative timing among different spectral bands of speech. The basic paradigm involved filtering speech into four discrete non-overlapping 1/3-octave bands (298-375 Hz, 750-945 Hz, 1890-2381 Hz, and 4762-6000 Hz) and presenting the bands in various combinations, either synchronously or asynchronously. Individually, the filtered speech bands were of very low intelligibility (roughly between 2-9% correct sentence recognition). When presented in combination however, recognition scores were significantly higher, often exceeding what one might expect from the simple addition of two independent bands. For example, combining bands 2 and 3 resulted in an average recognition score of 60% (individual bands recognition score of 9% each). When all four bands were combined synchronously, the score was 89%, showing that in quiet, as few as four non-contiguous (but widely spaced) frequency channels are sufficient for near perfect speech recognition (despite the omission of nearly 80% of the spectrum). In one experiment, the effects of spectral asynchrony was assessed by displacing the two middle bands in time with respect to the two fringe bands. The impact of both leading and lagging temporal displacements is illustrated in Figure 3 for a broad range of channel delays. Intelligibility is highest when all bands are presented synchronously and progressively declines as the degree of asynchrony increases. The effect of asynchrony on intelligibility is relatively *symmetric* in that performance is roughly similar for conditions in which the middle bands lead the fringe bands and vice versa.

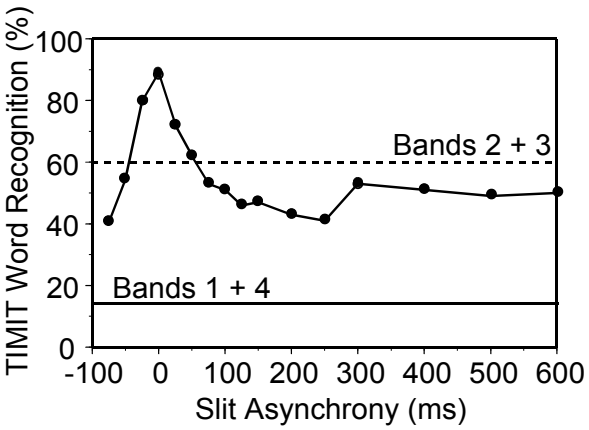

Figure 3. The effect of spectral slit asynchrony on the intelligibility of sentences. Adapted from [13].

Intelligibility is not always so sensitive to cross-spectral asynchrony, however. In instances where the spectral bandwidth of the speech signal is broad and continuous (e.g., no spectral gaps), listeners are fairly tolerant to cross-spectral asynchrony up to about 140 ms [2,12] before intelligibility appreciably declines. One interpretation of these data is that there appear to be multiple time intervals over which speech is decoded in the auditory system. These range from short analysis windows (1-40 ms), possibly reflecting various aspects of phonetic detail at the articulatory feature level (e.g., voicing), mid-range analysis windows (40-120 ms) possibly reflecting segmental processing, to long analysis windows (beyond 120 ms), possibly reflecting the importance of prosodic cues, such as stress accent and syllable number, in the perception of running speech [12,14].

2.2. Spectro-temporal asynchrony in auditory-visual speech processing

Analogous experiments to those described above were conducted recently by Grant and Greenberg [7] to assess the effects of temporal asynchrony on auditory-visual speech recognition. The audio signal consisted of the same two mid-frequency bands evaluated by Greenberg et al [13]. These were presented along with the video image of a talker across a range of asynchronous conditions where the audio signal lagged and led the video signal. Results are shown in Figure 4. The two most compelling aspects of these data are the overall size of the temporal window for which asynchronous audio-video speech input is recognized as well as synchronous audio-video speech and the highly *asymmetric* shape of the window. Unlike the temporal window for auditory speech recognition, 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 and course-grain syllabic analysis on the other, which may be interpreted as reflecting the different roles played by auditory and visual speech processing.

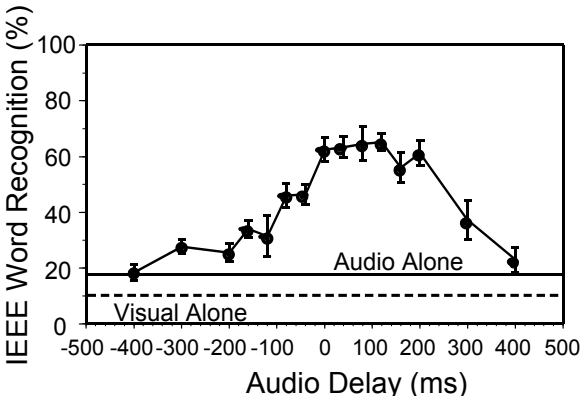

Figure 4. Sentence intelligibility as a function of audio-video asynchrony. Note the substantial plateau region between -50 ms audio lead to 200 ms audio delay where intelligibility scores are high relative to the audio-alone or video-alone conditions. [Adapted from 7].

When speech is processed by eye (i.e., speechreading) it is likely to be advantageous to integrate over long time windows of roughly syllabic lengths (200-250 ms) because visual speech cues are rather course. At the segmental level, visual recognition of voicing and manner-of-articulation is generally poor [9,11], 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 [10]. In contrast, acoustic processing of speech is much more robust and capable of much more fine-grained analyses using temporal window intervals between 10 and 40 ms [19]. 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.

3. Conclusions

Auditory and visual speech cues interact in a number of ways involving both the spectral and temporal domains. Not all spectral bands of speech have equal value in combination with speechreading. Low-frequency bands appear more beneficial than high-frequency bands. This is most likely due to the fact that low-frequency speech cues are more complementary to speechreading than are high-frequency speech cues. A second important finding is that auditory-visual speech processing, unlike audio-only processing, is highly tolerant of cross-modal asynchrony, but only when the visual stimulus precedes the audio stimulus. The range of auditory-visual temporal asynchronies which have little effect on speech intelligibility is fairly broad (roughly -50 ms of audio lead to +200 ms of audio lag). These spectro-temporal interactions provide a glimpse into how speechreading and audition may come together to form a remarkably robust signal that is greatly resistant to environmental disturbances such as noise and reverberation. On the one hand, speechreading may serve as a rough approximation to a mid-frequency acoustic speech signal. In other words, the visual channel *acts* like an additional auditory channel that is rich in place information but little else. Further, the long temporal integration window suggests that the influence of the visual channel is trans-phonemic, operating over syllabic length units of roughly 250 ms.

4. Acknowledgements

This research was supported by grant numbers DC 000792-01A1 from the National Institute on Deafness and Other Communication Disorders to Walter Reed Army Medical Center and SBR 9720398 from the Learning and Intelligent Systems Initiative of the National Science Foundation to the International Computer Science Institute. 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.

5. References

[1] ANSI S3.5-1969 (American National Standards Institute, New York).

[2] Arai, T., and Greenberg, S. (1998). "Speech intelligibility in the presence of cross-channel spectral asynchrony," *IEEE International Conference on Acoustics, Speech and Signal Processing*, Seattle, 933-936.

[3] Braida, L.D. (1991). "Crossmodal integration in the identification of consonant segments," Quart. J. Exp. Psych. **43**, 647-677.

[4] Grant, K.W., Ardell, L.H., Kuhl, P.K., and Sparks, D.W. (1985). "The contribution of fundamental frequency, amplitude envelope, and voicing duration cues to speechreading in normal-hearing subjects", J. Acoust. Soc. Am. **77**, 671-677.

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

[6] Grant, K.W., Braida, L.D., and Renn, R.J. (1991). "Single band amplitude envelope cues as an aid to speechreading," Quart. J. Exp. Psych. **43**, 621-645.

[7] 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, 132-137.

[8] 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.

[9] Grant, K.W., and Walden, B.E. (1996a). "Evaluating the articulation index for auditory-visual consonant recognition," J. Acoust. Soc. Am. **100**, 2415-2424.

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

[11] 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.

[12] Greenberg, S. (1996). "Understanding speech understanding: Towards a unified theory of speech perception," in Proceedings of the ESCA Workshop on the "Auditory Basis of Speech Perception", Keele University, 1-8.

[13] Greenberg, S., Arai, T., and Silipo, R. (1998). "Speech intelligibility derived from exceedingly sparse spectral information," in *Proceedings of the International Conference of Spoken Language Processing*. Sydney, Australia, December, 74-77.

[14] Huggins, A.W.F. (1972). "On the perception of temporal phenomena in speech," J. Acoust. Soc. Am. **51**, 1279-1290.

[15] Massaro, D.W. (1987). Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry. Hillsdale, NJ: Lawrence Earlbaum Assoc.

[16] 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.

[17] Rosen, S.M., Fourcin, A.J., and Moore, B.C.J. (1981). Voice pitch as an aid to speechreading," Nature **291**, 150-152.

[18] 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, 2687-2690.

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

[20] 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.

[21] Walden, B.E., Grant, K.W., and Cord, M.T. (2001). "Effects of amplification and speechreading on consonant recognition in persons with impaired hearing," Ear Hear. **22**, 333-341.