Auditory Supplements to Speechreading

Ken W. Grant

Army Audiology and Speech Center, Walter Reed Army Medical Center, Washington DC 20307-5001
E-mail: grant@tidalwave.net

Auditory-visual speech recognition is far more accurate and robust than speech recognition by hearing alone. Yet, in spite of the benefits and obvious importance of auditory-visual speech for everyday communication, little is known about the mechanisms involved in auditory-visual speech integration. As a preliminary step toward the development of a generalized model of speech communication that incorporates visual speech cues, it is necessary to delineate the spectral and temporal interactions that occur when visual speech cues are used in tandem with acoustic cues. It will be shown that this interaction is both highly synergistic and non-linear. Further, it is suggested that visual speech cues may serve as a guide for auditory speech processing by informing the listener of spectral and temporal landmarks that can be used to decode the speech message.

Keywords Speechreading, Auditory-Visual Speech Perception, Auditory-Visual Integration, Intelligibility Model

1. Introduction

1.1. Acoustic cues that supplement speechreading

The importance of visual cues for understanding spoken language has been known for some time [1]. Much of the early work focused on the specific needs of profoundly hearing-impaired patients who rely on speechreading as the primary means for decoding spoken language. When speechreading is used as the sole channel for receiving speech, a number of important segmental and suprasegmental speech features are lost (e.g., voicing, nasality, and intonation), thus restricting the rate and accuracy of communication to roughly 40% of that of a normal-hearing individual [2]. On the other hand, when speechreading is combined with information from other sensory channels (auditory or tactile), lost information can often be recovered, especially if the information supplied by the other sensory channels complements the information provided visually. An example of such a combination is shown in Table 1. The eleven visual categories or "visemes" shown in the top row of Table 1 were obtained from an analysis of error patterns made by trained normal-hearing speechreaders [3]. Consonants belonging to different categories were seldom confused (e.g., /b/ vs /t/), while consonants belonging to the same category were frequently confused (e.g., /b/ vs /p/). The subsequent three rows in the Table show what would be expected to happen if information about voicing, nasality, and affrication were provided from some other sensory channel and combined with speechreading. As can be seen, the additional information completely resolves all remaining ambiguities, thus leading, in theory, to perfect recognition.

1.2. The search for minimal acoustic supplements

In order to transmit information that complements speechreading to profoundly deaf individuals, researchers had to come to grips with a perplexing problem, namely, that the information-handling capacity of residual auditory function in deaf patients, or of the tactile system, is greatly reduced compared to normal hearing [4]. As a result, natural speech signals could not be directly transmitted to these channels and some form of coding was required. In other words, it became necessary to find acoustic and/or tactile supplements to speechreading that were 1) capable of conveying information about voicing, nasality, affrication, and other speech features that were not readily transmitted via speechreading, and 2) were simple enough to be processed effectively by the receiving modality. This approach resulted in a number of demonstrations which showed that certain acoustic signals, which by themselves were mostly unintelligible, could nevertheless lead to very high intelligibility scores when combined with speechreading [2, 5, 6]. For example, Grant, et al. [2] measured the contribution of auditory sinewave analogs representing various speech features, such as amplitude-envelope and fundamental-frequency information, to speechreading. The acoustic signals were pure tones modulated in frequency (FM), amplitude (AM), or both (AMFM) based on an analysis of the

Table 1. Linguistic feature contributions to visual speech recognition. The top row represents typical feature classifications for speechreading alone (visemes). Each subsequent row represents the effects of adding information about another linguistic feature via an additional input channel (in this case auditory). Note that as additional features are added, consonant confusions associated with speechreading are resolved to a greater and greater extent. Adapted from [3].

Speechreading	p,b,m	t,d,n	g,k	f,v	θ,ð,s,z	ʃ,tʃ,dʒ,ʒ	l	r	w	j			
+													
Voicing	p	b,m	t	d,n	g	k	f	v	θ	ð	s	z	ʃ,tʃ	dʒ,ʒ	l	r	w	j
+													
Nasality	p	b	m	t	d	n	g	k	f	v	θ	ð	s	z	ʃ,tʃ	dʒ,ʒ	l	r	w	j
+													
Affrication	p	b	m	t	d	n	g	k	f	v	θ	ð	s	z	ʃ	tʃ	dʒ	ʒ	l	r	w	j

amplitude and frequency of the voice fundamental. Speech understanding was evaluated using the connected discourse tracking procedure [7] which involves a speaker reading aloud from text and the receiver repeating verbatim what the speaker has said. The sessions are timed and the results expressed as the number of correctly reproduced words per minute (WPM), or as a percent of the normal-hearing tracking rate (roughly 110 WPM). Results are displayed in Figure 1, and show clearly that the reception of connected speech is improved dramatically with acoustic signals that by themselves have almost zero intelligibility. For example, the tracking rate increased from roughly 37% for speechreading alone (SA) to almost 68% for either AM or FM tones. A further increase to nearly 80% of the normal tracking rate was observed when amplitude-envelope and fundamental-frequency information were combined, as in the AMFM condition or a lowpass filtered speech condition (LPF) with a cutoff frequency of 300 Hz.

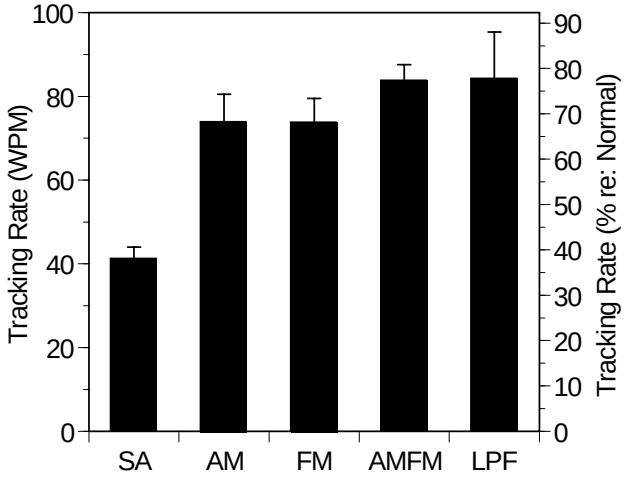

Figure 1. Connected discourse tracking rates for speechreading alone (SA), and for speechreading plus amplitude modulated tone (AM), frequency-modulated tone (FM), amplitude- and frequency-modulated tone (AMFM), and lowpass-filtered speech (LPF). Reprinted from [2].

1.3. A challenge for models of speech intelligibility

These data, along with those from similar studies that use non-traditional acoustic signals as supplements to speechreading [5, 6, 8, 9], pose a serious challenge for models of speech intelligibility which base their predictions solely on physical attributes of the signal, speaker, listener, and the listening environment [10, 11, 12]. Predictors of speech intelligibility such as the Articulation Index (AI) or the Speech Transmission Index (STI), either ignore the role of visual speech cues altogether (STI), or treat the visual channel as an *independent* source of speech information that simply adds to the auditory information (AI). In the case of the AI, this relatively simplistic view is most likely incorrect in that it does not allow for the possibility of auditory-visual interactions. For example, in the 1969 ANSI standard for calculating the Articulation Index [13], a graphical correction to the auditory AI was used when visual cues were present. This correction curve is shown in Figure 2. As indicated by the figure, the auditory-visual AI is simply a function of the auditory AI, regardless of any differences (spectral or temporal) that might exist among acoustic signals. Thus, for example, a calculated auditory AI of 0.2 would always be equivalent to an effective auditory-visual AI of 0.35. For low-context sentence materials, this *effective* increase in AI translates to an increase in intelligibility from roughly 50% words correct to 90% words correct [14].

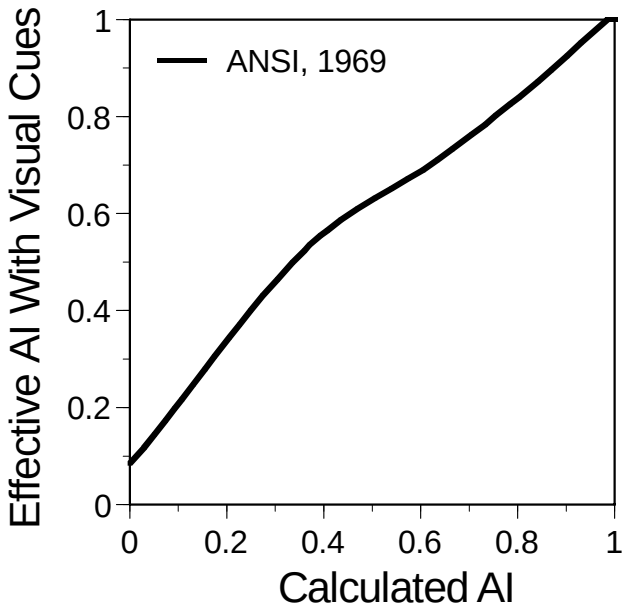

Figure 2. ANSI 1969 correction curve for estimating the effective Articulation Index (AI) with visual cues. Reprinted from [13].

The overarching assumption of the ANSI 1969 AI standard vis à vis auditory-visual speech recognition is that visual cues benefit speech intelligibility equally regardless of the spectral content of the acoustic speech signal. This assumption was tested directly by Grant and Walden [15]. The auditory conditions consisted of /ɑ/-consonant-/ɑ/ (aCa) tokens processed through twelve different bandpass filters of varying bandwidth and center frequencies. The results showed that there was little relationship between overall auditory intelligibility and overall auditory-visual intelligibility (Figure 3). Further inspection of these data revealed that low-frequency bands (e.g., 250-505 Hz) tended to provide much more benefit to speechreading than mid-frequency (e.g., 2800-3255 Hz) or high-frequency (e.g., 4200-5720 Hz) bands. An information analysis [16] of the consonant error

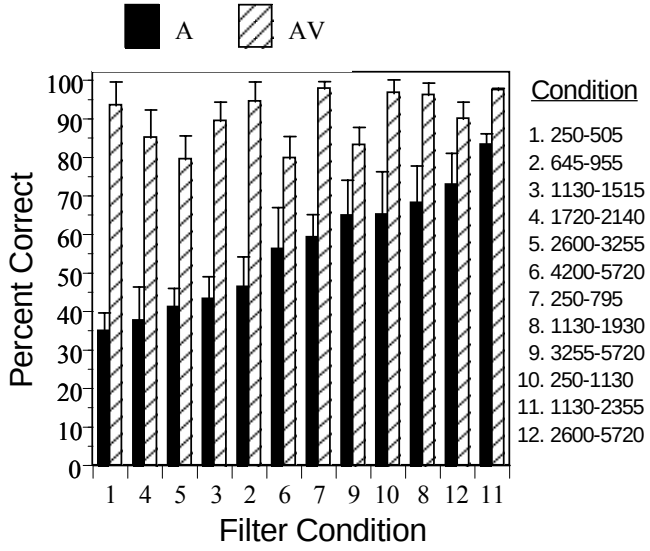

Figure 3. Auditory and auditory-visual speech intelligibility as a function of filter band. Note, that even though some bands have greater auditory intelligibility than others (e.g., band 6 versus band 1), their auditory-visual intelligibility is not as great. Adapted from [15].

patterns showed that the information conveyed by speechreading-alone was almost completely restricted to place of articulation (i.e., little or no transmission of voicing or manner-of-articulation information). Furthermore, auditory bands which conveyed a relatively high degree of place-of-articulation information provided the least amount of benefit when combined with speechreading. In other words, when the auditory and visual channels contained similar articulatory feature information (i.e., the two channels were mostly ***redundant*** with respect to each other), little auditory-visual benefit was obtained. In contrast, when the auditory channel conveyed a relatively high degree of information about consonant voicing and consonant manner of articulation (i.e., ***complementary*** information relative to speechreading), the auditory-visual benefit was very high. Thus, in direct contradiction to the assumptions made in the ANSI Standard, different auditory conditions with the same AI need not have the same auditory-visual AI. These findings are consistent with various models of auditory-visual integration [17, 18, 19] and demonstrated that, at least for consonants, the amount of benefit provided by combining auditory and visual speech cues is determined primarily by the degree of articulatory-feature redundancy between the two channels.

The data from Grant and Walden [14] strongly suggest that visual speech cues have a ***weighted*** influence on the perception of auditory cues, depending on the spectral content of the acoustic speech signal. Moreover, the amount of benefit provided by visual speech cues for nonsense syllable recognition can be predicted fairly accurately by determining the degree of complementarity between the auditory and visual channels [14, 17, 18, 19]. In response to these observations, the revised standard for calculating the Articulation Index (referred to as the Speech Intelligibility Index, or SII) has removed the graphical auditory-visual correction procedure, and limited the scope to conditions that do not include multiple, sharply filtered bands of speech, sharply filtered noise, or acoustic signals which are not typical of normal speech (e.g., sine wave speech).

1.4. Auditory-visual interactions in time and frequency

The studies described above have helped improve our understanding of some of the various perceptual factors involved in auditory-visual speech recognition, while at the same time exposing certain weaknesses and limitations in models of speech intelligibility in general. However, they do not speak directly to the mechanisms or processes that are used during auditory-visual speech recognition. Specifically, how do listeners relate what they see on the lips to what they hear? Summerfield [20] hypothesized three possible roles for visual cues in improving speech understanding in noise. The two most apparent of these are to provide segmental (e.g., consonants and vowels) and suprasegmental (e.g., intonation, stress, rhythmic patterning, etc.) information which is 1) redundant to cues provided acoustically, and 2) complementary to cues provided acoustically. As already discussed, the greatest benefits occur when speechreading and audition provide complementary feature information. In noisy and reverberant environments, or for individuals with hearing impairment, many of the relevant acoustic attributes that lead to the identification of phonetic units may be very weak, absent, or distorted [21]. Under these conditions, there is significant ambiguity in the auditory channel, in particular with regard to place-of articulation. When audition and speechreading are combined, however, a substantial proportion of place cues are restored through speechreading and the integrated auditory-visual percept is far more complete than that obtained from either of the unimodal sources alone.

The third role of speechreading hypothesized by Summerfield (1987) pertains to the spectro-temporal relations that exist between visible movements of a speaker's articulators and the acoustic speech signal. When a listener watches a talker speak, the acoustic signal and the visible movements of the talker's lips share common spatial, temporal, and spectral properties which help segregate the speech signal of interest from the surrounding background noise. Direct measurements of the displacements of the upper and lower inner margins of the lips at midline, or of the area of lip opening, have been shown to be related to the overall amplitude contour of the speech signal [22]. Further measures have shown that the correlation between the area of lip opening and acoustic envelope dynamics also depends on the spectral region of the acoustic signal, with the highest correlation observed for acoustic signals with energy concentrated in the region of the second and third formant frequencies [23].

One psychophysical consequence of this relation between speech kinematics and spectrally-specific acoustic speech envelopes is that when visual speech cues are present there is a reduction in the spectral and temporal uncertainty associated with the onset of syllables and words. Recent studies [22, 23] have shown that this reduction in uncertainty can lead to improved speech *detection* thresholds in noise, through a process called bimodal comodulation masking protection (BCMP). In other words, watching the movements of the lips during speech production can inform the listener not only *where* in space and *when* in time to listen to prominent acoustic events, but also *where* in the acoustic spectrum to expect the events to occur. The experimental paradigm used in these studies was a variant of the comodulation masking release paradigm [24]. The primary goal was to determine if comodulated activity between orofacial kinematics and acoustic amplitude envelope led to an improvement in speech detection thresholds. Thresholds for detecting sentences in noise were determined under a variety of conditions: auditory alone, auditory-visual with matching (congruent) video, filtered auditory-visual speech with congruent video, and auditory-visual with unmatched (incongruent) video. For each condition, the degree of correlation between area of mouth opening and auditory envelope fluctuations was determined. In addition, a control condition using visual orthography to indicate the text of the target audio sentence prior to each test trial. was tested

As seen in Figure 4, a significant masking release for detecting spoken sentences (1-3 dB depending on the specific target audio sentence) was observed when simultaneous visual speech information was provided. There was no effect on auditory masking when mismatched (incongruent) visual speech information was provided (not shown). Results of informing subjects as to the identity of the target audio sentence using an orthographic display resulted in a small release from masking (0.5 dB) that was independent of the target sentence, probably reflecting a general reduction in stimulus uncertainty. Results for filtered-speech targets corresponding roughly to the first (100-800 Hz) and second (800-2200 Hz) formant-frequency regions showed that mid-frequency speech targets produced a masking release equivalent to that of broadband unprocessed speech, and low-frequency speech targets produced significantly smaller amounts of masking release.

These results suggest that the visible modulations of the lips and jaw during speechreading make auditory detection of speech easier by informing listeners about the probable spectro-temporal structure of a near-threshold acoustic speech signal. A correlation analysis of the area of opening of the lips and the acoustic envelope modulations of the co-occurring sentence revealed a predictive relationship between the degree of masking release and the strength of the area function/acoustic envelope correlation. Specifically, sentences with a high correlation between area function and acoustic envelope showed more

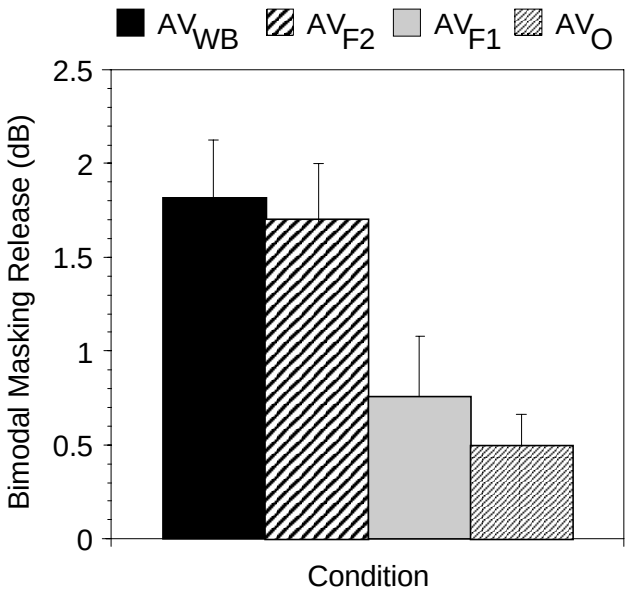

Figure 4. Masked threshold differences, or masking release (in dB) for auditory and matching auditory-visual conditions. AV_{WB} = wideband speech; AV_{F2} and AV_{F1}= bandpass-filtered speech (see text); AV_O = orthographically cued speech. Adapted from [22, 23].

masking release than sentences with a lower correlation. Furthermore, measures between the area of lip opening and acoustic envelope modulations extracted from specific spectral regions of speech showed that a higher correlation can be expected for acoustic energy modulations in the second (F2) and third (F3) formant regions. This correspondence is exactly what one might predict given speechreaders' abilities to extract primarily place-of-articulation information. It is well established that auditory place-of-articulation information is conveyed by cues contained primarily in the mid-to-high frequency regions (i.e., F2 and F3 regions of speech).

2. Discussion and Conclusion

Viewed collectively, studies describing the benefits of auditory-visual speech recognition using minimally intelligible acoustic signals, and studies of the relations between speech acoustic and articulatory dynamics suggest a two-tiered approach towards modelling auditory-visual integration. On the one hand, linguistic information derived from auditory and visual speech processes combine synergistically, such that information disrupted or lost in one channel may be recovered by the other channel [17, 18, 19]. The best example of this pertains to place-of-articulation information, which is extremely vulnerable acoustically to noise, reverberation, and hearing loss. Place cues, however, are fairly robust visually and are minimally affected, if at all, by noise, reverberation, and hearing loss. Naturally, visual place cues are affected by other environmental factors such as lighting and viewing angle, but these factors are relatively unimportant to auditory processing. This complementary arrangement of acoustic and visual speech cues makes for a remarkably robust signal and forms the basis of most current models of auditory-visual integration.

A second important way that visual speech interacts with auditory speech processing is through the correlated activity between movements of the face during speech production and various aspects of the speech amplitude envelope. This information can be used by listeners to influence low-level auditory processing of speech. The visible movements of orofacial structures during speech production inform listeners about when (in time) to expect peak amplitudes in the acoustic waveform, and where (in the acoustic frequency spectrum) to expect these peaks to occur. Thus, by watching the face while listening to speech, there is a significant reduction in signal uncertainty that enables listeners to extract signals from noise at S/N ratios that otherwise would be below threshold.

By considering the physical coherence between the two sources of information, in addition to their respective linguistic content, it becomes possible to couch some of the benefits of speechreading in auditory-visual speech processing in terms of the activity of populations of multisensory neurons having particular optimal stimulus onset asynchronies. For example, enhanced physiologic responses of multisensory neurons presumably translate to increased reaction speeds of superior colliculus-mediated attentive and orientation responses [25]. These enhanced levels of neurologic activity may provide greater overall drive to higher-level auditory neurons, which in turn allow for reduced speech detection thresholds. Of course, at this point, these are only speculations. But the latest psychophysical results on bimodal coherence masking protection are at least consistent with recent physiological findings. In addition, this somewhat more physical interpretation of how speechread cues can be used to guide auditory analysis of speech suggests new strategies for signal processing in the areas of automatic speech recognition and automatic noise reduction in hearing aid design. For instance, it may be possible to use the correlated activity between optics and acoustics in speech production to fashion temporal filters that can be used to effectively segregate target speech components from interfering background noise or other talkers. Further, the fact that the visible movements of the lips are operating on a time frame consistent with slow-rate acoustic modulations in the range between 0-30 Hz, suggests a more unified approach to modelling auditory-visual speech processing. The basic idea of this approach is to treat the visual speech signal as an additional channel of amplitude modulation that can augment and guide auditory modulation analysis of speech. In other words, an auditory-visual modulation spectrum could be derived and interpreted in much the same way as auditory modulation patterns are currently interpreted within models such as the Speech Transmission Index [12]. There is a growing literature demonstrating that speech intelligibility is critically dependent on the preservation of these slow-rate, spectro-temporal amplitude modulations, reflecting the dynamic movement of the speech articulators [26, 27] as well as variations in syllable and phonetic duration observed in conversational speech [28]. Because the visual channel can serve as another source of this critical information, one that is relatively immune to environmental noise and reverberation, it should prove invaluable in a host of applications from models of speech intelligibility to automatic speech recognition. Exactly how to use this information best will require further work aimed at delineating the precise relations between acoustic and visual modulation spectra and the extent to which this information is spectrally specific. Work along these lines is currently underway.

3. Acknowledgments

This research was supported by the Clinical Investigation Service, Walter Reed Army Medical Center, under Work Unit #2508 and 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. All subjects participating in this research provided

written informed consent prior to beginning any of the described studies. I would like to thank Dr. Jennifer Tufts for her helpful comments on an earlier draft of this paper. The opinions or assertions contained herein are the private views of the author and are not to be construed as official or as reflecting the views of the Department of the Army or the Department of Defense.

References

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

[2] 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.

[3] Grant, K.W., Ardell, L.H., Kuhl, P.K., and Sparks, D.W. (1986). "The transmission of prosodic information via an electrotactile speechreading aid," Ear and Hearing, 7, 328-335.

[4] Mazeas, R. (1968). "Hearing capacity, its measurement and calculation," Amer. Annals of the Deaf 113, 268-274.

[5] Rosen, S.M., Fourcin, A.J., and Moore, B.C.J. (1981). "Voice pitch as an aid to lipreading," Nature 291 (5811), 150-152.

[6] Breeuer, M., and Plomp, R. (1984). "Speechreading supplemented with frequency-selective sound-pressure information," J. Acoust. Soc. Am. 76, 686-691.

[7] DeFilippo, C.L., and Scott, B.L. (1978). "A method for training and evaluating the reception of ongoing speech," J. Acoust. Soc. Am. 63, 1186-1192.

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

[9] Grant, K.W., Braida, L.D., and Renn, R.J. (1994). "Auditory supplements to speechreading: Combining amplitude envelope cues from different spectral regions of speech," J. Acoust. Soc. Am. 95, 1065-1073.

[10] French, N.R., and Steinberg, J.C. (1947). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am., 19, 90-119.

[11] Fletcher, H., and Gault, R.H. (1950). "The perception of speech and its relation to telephony," J. Acoust. Soc. Am., 22, 89-150.

[12] Houtgast, T., and Steeneken, H.J.M. (1980). "Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics," Acoustica 46, 60-72.

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

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

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

[16] Miller, G.A., and Nicely, P.E. (1955). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. 27, 338-352.

[17] Massaro, D.W. (1987). *Speech Perception by Ear and Eye: a Paradigm for Psychological Inquiry*. Hillsdale, NJ: Lawrence Erlbaum Assoc.

[18] Massaro, D.W. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle.* Cambridge, MA: MIT Press.

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

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

[21] Lindbloom, B. (1996). "Role of articulation in speech perception: Clues from production," J. Acoust. Soc. Am. 99, 1683-1692.

[22] 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.

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

[24] Hall, J.W., Haggard, M.P., and Fernandes, M.A. (1984). "Detection in noise by spectro-temporal pattern analysis," J. Acoust. Soc. Am., 76, 50-56.

[25] Meredith, M.A., and Stein, B.E. (1996). "Spatial determinants of multisensory integration in cat superior colliculus," J Neurophysiol 75, 1843-1857.

[26] Drullman, R., Festen, J., and Plomp, R. (1994). Effect f envelope smearing on speech perception," J. Acoust. Soc. Am. 95, 1053-1064.

[27] Arai, T., Pavel, M., Hermansky, H., and Avendano, C. (1996). "Intelligibility of speech with filtered time trajectories of spectral envelopes," Proc. ICSLP, 2490-2492.

[28] Greenberg, S. (1997). "On the origins of speech intelligibility in the real world," Proc. ESCA Workshop on *Robust Speech Recognition for Unknown Communication Channels*, 23-32.