Modeling auditory and auditory-visual speech intelligibility: Challenges and possible solutions

Ken W. Grant, Joshua G.W. Bernstein, and Elena Grassi

Army Audiology and Speech Center, Walter Reed Army Medical Center, Washington, DC, USA

Models of speech intelligibility (e.g., Speech Intelligibility Index and Speech Transmission Index) have proven useful in a number of applied (e.g., algorithm development) and theoretical applications (e.g., theories of speech perception). However, in many real-world situations, these models fail to predict accurately speech intelligibility due to the complex nature of the soundscape (e.g., competing talkers), particular attributes of the listener/talker combination (e.g., speaking rate, age, and hearing loss), and presentation modality (auditory or auditory-visual). This paper discusses several of these challenges and recent efforts to address them. Particular attention is paid towards our efforts to model auditory-visual speech intelligibility. Current models of speech intelligibility base their predictions on characteristics of the acoustic speech signal, background noise, and reverberation. However, because visual speech cues are not included in these models, they provide a poor prediction of speech intelligibility in many everyday environments. To address this particular challenge, we describe a method for integrating visual and acoustic speech cues into a unified model of speech intelligibility. Kinematic motion from a talkers' face during speech production is combined with the acoustic speech signal processed by a computational multi-channel model of peripheral auditory analysis. The outputs of the peripheral model are integrated with the visual signal in a weighted fashion based on the degree of to which the visual kinematics are predictive of the acoustic envelopes derived from each frequency channel, yielding an enhanced acoustic signal, especially in the mid-to-high frequencies. Enhanced and unmodified noisy speech signals are then processed through a cortical model which extracts critical speech modulations to compute a spectro-temporal modulation index (STMI), yielding predictions for auditory and auditory-visual speech presented in steady-state noise.

INTRODUCTION

Models of speech intelligibility, such as the Articulation Index (AI; ANSI, 1969), Speech Intelligibility Index (SII; ANSI, 1997), and the Speech Transmission Index (STI; Steeneken and Houtgast, 1980) serve at least two broad purposes. First, they highlight various properties of speech and background signals thought to be important for speech intelligibility. For example, when calculating the STI, intelligibility is thought to be determined largely by the ability to preserve the slow rate time-intensity information (i.e., amplitude envelope) extracted from separate channels along the tonotopic frequency axis. In the SII, intelligibility is thought to be determined by the

Auditory signal processing in hearing-impaired listeners. 1st International Symposium on Auditory and Audiological Research (ISAAR 2007). T. Dau, J. M. Buchholz, J. M. Harte, T. U. Christiansen (Eds.).
ISBN: 87-990013-1-4. Print: Centertryk A/S.

speech-to-noise ratio in different spectral bands, weighted in importance according to the type of materials used for testing (e.g., nonsense syllables, words, or sentences). Each of these models represents a view of speech recognition that has proven useful in predicting average speech intelligibility under a variety of listening situations.

In addition to validating choices concerning the important determinants of speech intelligibility, models of intelligibility provide a practical tool that can be used to assess new devices and algorithms for transmitting speech information without the time and cost associated with behavioral testing. Based primarily on physical measures of the speech signal and background environment, these models predict the transmission efficiency and accuracy of different signal processing schemes prior to testing any listeners.

Over the years, speech intelligibility models have undergone refinements intended to improve accuracy and robustness across different background environments and populations of listeners. Nevertheless, several fundamental challenges remain in generating a robust model of speech intelligibility applicable to the range of listening situations encountered in everyday life. These challenges include (a) the availability of non-acoustic cues for speech understanding, (b) complex effects associated with fluctuating noise and informational masking, (c) individual differences in speech understanding across listeners, and (d) the availability of visual speech information. First, a brief overview of each of these challenges is presented, along with a discussion of attempts made to address each problem. Then, a potential solution for the fourth challenge is described in more detail, whereby auditory and visual signals are integrated in a stimulus-based model of speech intelligibility.

CHALLENGES TO SPEECH INTELLIGIBILITY MODELS

Non-acoustic speech information

Importance functions used to weight the model outputs along the frequency axis appear to shift with speech material (syllable, word, and sentence, ANSI, 1997) and with input modality (auditory or auditory-visual, Grant, 2006; Grant and Bernstein, 2007). The confounding of "hearing speech" (access to acoustic cues) with "understanding speech" (use of language processes including lexical knowledge, context, and memory) continues to be debated. The question of whether non-acoustic cues and cognitive linguistic processes should be included in the computation of intelligibility indexes has not been resolved.

As an example, the ANSI (1969) standard for estimating the articulation index (AI) shows significant variation in the recognition scores for different speech materials at the same AI (see Figure 15, ANSI, 1969). This is because speech intelligibility models were originally developed to account for the effects of acoustic speech cues and not the many non-acoustic processes involved in speech perception. Researchers have attempted to extend these models to integrate contributions of non-speech cues (French and Steinberg, 1947). For example, Boothroyd and Nittrouer (1988) characterized the relation between phoneme scores, word scores and sentence scores in terms of scalar exponents, *j*- and *k*-factors. However, this characterization is based on empirical curve fitting and therefore cannot be generalized beyond the psychoacoustically measured speech databases and particular listening conditions. Furthermore, questions remain as to how the use of context may interact with the available acoustic cues. Semantic and morpho-syntactic context may affect consonant and vowel recognition differently. The optimal use of context may depend on the acoustic environment (i.e., reverberation, environmental noise, etc.), or on the frequency content of the speech. The amount and type of context available to the listener may also affect the integration of bottom-up and top-down speech processes.

Fluctuating noise and informational maskers

Speech intelligibility varies greatly depending on the temporal and linguistic characteristics of the background maskers. Compared to steady-state maskers, temporally modulated maskers typically result in less speech interference for normally-hearing subjects owing to the listeners' ability to extract information about the target signal during quiet portions of the masker (Festen and Plomp, 1990). However, when the masker includes competing speech from other speakers, the amount of interference can be very large due to the informational content of the masker (Brungart, 2001). Thus, for speech intelligibility models to predict performance accurately under these complex masking conditions, temporal and informational properties of the target and background signals must be considered. Recent work by Rhebergen *et al.* (2006) has addressed the issue of fluctuating maskers by implementing a series of short-time calculations that enables the intelligibility index to reflect the benefit engendered by temporal dip listening in normal-hearing listeners. However, this modification does not address the well-known result whereby hearing-impaired listeners generally do not benefit from fluctuating maskers (Festen and Plomp, 1990; George *et al.* 2006). It remains an open question whether this deficit reflects a simple temporal resolution problem, or a suprathreshold distortion that limits the cues available for segregating the target speech from a speech-like fluctuating masker (Bregman, 1990). For example, given that hearing-impaired listeners are more likely to rely on temporal rather than spectral cues for understanding speech (due to reduced frequency resolution), a temporally modulated noise, which is easily segregated from speech by normally-hearing listeners, may appear to be more "speech like" to someone with a hearing loss, thus causing significant interference between target and background streams. This suggests similarities to modulation detection interference (MDI) whereby a modulated masker, remote in frequency from a modulated target, can nevertheless exert significant disruption on the detection of target modulation (Yost and Sheft, 1994; Oxenham and Dau, 2001). Thus, when the target and masker become perceptually similar, *informational* masking (defined as a reduction in performance that cannot be accounted for by peripheral energetic masking) becomes a factor which can negatively impact speech understanding (Brungart, 2001). Difficulty in understanding the target speech in presence of competing talker(s) is a very commonly encountered scenario in everyday life. However, as informational masking involves linguistic processing and memory, it requires an elaboration of central auditory processing which is typically not part of current intelligibility models.

Individual differences across listeners

Speech recognition involves the extraction, integration, and interpretation of cues contained in the speech signal and environment. As discussed earlier, if the speech materials are meaningful words, phrases, or sentences, the listener's vocabulary, knowledge of the particular situation, and general knowledge about the world also play an important role in the recognition process. As a result of the many different stages involved in speech recognition, performance on a given speech task can vary widely across listeners. This is especially true for elderly and hearing-impaired listeners tested in noisy environments even when audiometric thresholds are quite similar. Therefore, models of intelligibility that include only the audiogram as a characterization of the listeners hearing cannot fully account for observed individual differences. Plomp (1978) suggested that suprathreshold differences in hearing capacity across listeners may relate better to speech recognition scores in noise. Recent efforts to model suprathreshold deficits have mostly been directed towards peripheral distortion such as reduced frequency and temporal selectivity (Glasberg and Moore, 1989), and not central processing. Individualized characterizations of the auditory periphery, when incorporated into a model of intelligibility, have the potential to track more accurately individual performance differences. Efforts are underway to model the internal representations of speech in noise for these listeners (Zilany and Bruce, 2007; Summers *et al.*, this volume). These efforts combine a peripheral model of the auditory system (e.g., Lopez-Poveda and Meddis, 2001) with a cortical model (e.g., Chi *et al.*, 1999; Elhilali *et al.*, 2003). The output of the cortical model highlights spectral and temporal modulations in the target (noisy) speech signal and compares these to the outputs of a "clean" speech signal through a normal-hearing model. By fitting model parameters for the peripheral stages of processing according to individual data obtained through psychophysical tests, predictions of speech recognition in noise can more accurately reflect individual suprathreshold distortion than current models using only hearing thresholds.

Effect of visual speech information

Under typical communication settings between two or more people, speech recognition is determined by both auditory and visual cues. Research has shown that auditory-visual speech is more robust, easier to recognize in noise and reverberation, and faster to encode than auditory alone speech. Furthermore, how a listener performs with audio-alone speech is not necessarily predictive of how they will perform in AV conditions (Grant and Walden, 1996). For example, if a listener has only high-frequency auditory information available, the visual benefit will not be as great as if they only have low-frequency information. This suggests a need for a model that can make predictions about a listener's AV speech intelligibility.

A number of models have been proposed to account for the benefits of visual cues to speech perception. However, these models require that recognition performance in auditory, visual, and auditory-visual conditions be measured separately (Braida, 1991; Massaro, 1998). Thus, although predictions of audiovisual (AV) speech recognition and AV integration are possible with these models, one of the key practical advantages afforded by an intelligibility index such as the STI or SII is lost, namely that predictions can only be made after extensive behavioral data on speech recognition have been gathered. What is required is some means for combining auditory and visual sources of information based on the physical characteristics of the signals themselves. A potential solution for integrating auditory and visual speech cues for speech intelligibility models is presented below.

A SIGNAL-BASED MODEL OF AUDIO-VISUAL SPEECH INTELLIGIBILITY

Background

Elhilali *et al.* (2003) proposed a model of speech intelligibility based on the degree to which the temporal and spectral modulations of speech are preserved after being degraded by the external environment and processed by the auditory system (Chi *et al.*, 1999). Their model of auditory processing consists of a middle ear filter, a peripheral filterbank stage, a lateral inhibition stage providing spectral sharpening, speech envelope extraction, and a cortical stage that highlights slowly changing spectral and temporal modulations. A spectrotemporal modulation index (STMI) then provides an estimate of speech intelligibility by comparing the modulations contained in a clean-speech template (derived by passing clean speech through the peripheral and cortical models) to those contained in the target speech signal. The STMI has successfully accounted for degraded speech intelligibility due to background noise and reverberation (Elihilali *et al.*, 2003) and the speech recognition differences associated with directional and omnidirectional microphones (Grant et al, in press).

Based on the work of Grant and Seitz (2000) and Grant (2001) showing that speech envelopes and visual lip movements are correlated, and that listeners exploit this correlation to improve speech detection in noise, Girin *et al.* (2001) and Berthommier (2004) showed that the physical lip movements can be used to enhance acoustic speech signals that have been degraded by noise. We used a similar approach whereby the outputs of the peripheral stage are enhanced by the visual input signal before being passed to the cortical stage.

Stimuli

Visual inputs to the model were extracted using an Optotrak 3D system (Northern Digital Inc., Waterloo, Ontario, Canada). An array of small infrared-emitting units were attached to a talker's face and head at 14 positions on the face (eight around the lips, two on the chin, and two on each cheek). Five additional units were attached to a crown fixed to the talker's head, which served as a reference point to subtract out head movements from the facial movement data. Three infrared cameras recorded the movements of the sensors during speech production, allowing the extraction of the three dimensional locations of each sensor over time (sampling rate of 130 Hz). Recordings of sixty IEEE (1969) sentences spoken by a single male talker formed the primary database used for this study. Although the Optotrak captured movements in three dimensions, only two dimensions (vertical and horizontal) were used, yielding a total of 28 channels of visual information (14 sensors x two dimensions). For the purposes of this work, the depth dimension (i.e. distance from the camera) was discarded.

The corresponding audio speech signal was digitally stored simultaneously at a sampling rate of 22050 Hz. The linear peripheral filterbank of the Elhilali *et al.* model was replaced by the level-dependent, dual-resonance nonlinear filterbank (Lopez-Poveda and Meddis, 2001) parametrized for a normal-hearing listener. There were 136 peripheral channels, with logarithmically-spaced best frequencies (BFs) ranging from 125 to 6000 Hz. A lateral inhibition stage (Elhilali *et al.*, 2003) reduced the number of channels by one to 135. Envelopes for the output of each channel were derived by half-wave rectifying and low-pass filtering (3-dB cutoff = 20 Hz, slope 6 dB/oct). The audio envelopes were then downsampled to the 130-Hz rate associated with the visual data.

Combining auditory and visual information

Based on the work of Girin *et al.*, visual and degraded auditory envelopes were combined to make the best possible estimate of the clean auditory envelopes. The resulting "enhanced" auditory envelopes then became the input to the next stage of the model. Formally, the clean auditory envelope ($A_{ch,clean}$) for each of 135 auditory channels ch was assumed to be a linear combination of the degraded auditory envelope for that channel ($A_{ch,degr}$) and the visual motion (V_i) from each of the 28 visual channels i, plus noise (ε):

$$A_{ch,clean}(t) = b_{a,ch} A_{ch,\deg r}(t) + b_{0,ch} + \sum_{i=1}^{28} b_{i,ch} V_i(t) + \varepsilon(t) \quad \text{(Eq. 1)}$$

where $b_{a,ch}$ and $b_{1,ch} \ldots b_{28,ch}$ represent the weights applied to the degraded auditory envelope and the visual envelopes respectively, and $b_{0,ch}$ is a constant for channel ch.

A training phase based on thirty IEEE sentences produced estimates ($\hat{b}_{a,ch}$, $\hat{b}_{0,ch} \ldots \hat{b}_{28,ch}$) of the 4050 coefficients in equation 1 (135 channels times 30 coefficients/channel) that would yield the best prediction of the clean auditory envelopes ($\hat{A}_{ch,clean}$), i.e. minimizing the root-mean-square of the error term ε. Because we did not expect the visual signals to yield information about the absolute magnitude of the audio signal, the auditory envelopes were normalized by the maximum envelope value across the 135 channels and 30 sentences.

It was assumed that the weights ba would vary depending on the degree to which the auditory envelopes were degraded. For example, in the case of no background noise, $A_{ch,degr}$ would yield a perfect estimate of $A_{ch,clean}$, and therefore $b_{a,ch}$=1 and $b_{0,ch}$ ... $b_{28,ch}$ would all equal zero. At the other extreme, at a poor SNR where the audio signal is inaudible, $b_{0,ch}$ ... $b_{28,c}$ would be maximal and $b_{a,ch}$=0. Therefore, a different set of coefficients were estimated for each of a range of signal-to-noise ratios (SNRs). The spoken sentences were mixed with stationary Gaussian speech-shaped noise (SSN, shaped to have the same long-term amplitude spectrum as the speech) at SNRs of +∞ (no noise), +18, +12, +6, +3, 0 -3, -6, -12 and -18 dB. The level of the speech wa fixed at 65 dB SPL, and the level of the noise varied to produce each desired SNR. The envelopes derived from the combined speech and noise formed the degraded signals used to produce estimates of the coefficients in equation 1 at each SNR.

Figure 1A shows the correlation between the clean ($A_{ch,clean}$) and degraded auditory envelopes ($A_{ch,degr}$) as a function of the auditory periphery channel BF, for the various SNRs tested. The envelope information available in the degraded auditory signals is clearly affected by the presence of noise. Figure 1B shows the correlation between the $A_{ch,clean}$ and the $\hat{A}_{ch,clean}$ calculated from the $A_{ch,degr}$ and the V_i's according to equation 1. The correlations in Fig. 1B are larger than for the corresponding plots in Fig. 1A, indicating that the visual channels have restored some of the auditory information that had been degraded by noise. There is little decrease in the magnitude of the correlations below an SNR of -12 dB SPL, suggesting that the -12 and -18 dB plots represent the proportion of the variation in the $A_{ch,clean}$ accounted for by the visual signals alone. At these low SNRs, the correlation tends to increase in magnitude with increasing BF, indicating that the visual inputs are redundant and provide more information about high than low frequencies. This is consistent with previous results indicating that visual inputs tend to provide more perceptual benefit when only low-frequency auditory information is available than when only high-frequency audio content is present (e.g. Grant and Walden, 1996; Bernstein and Grant, this volume). Figure 1B demonstrates that introducing the availability of low-frequency auditory information by adjusting the SNR would have a greater effect on intelligibility than altering the high-frequency SNR.

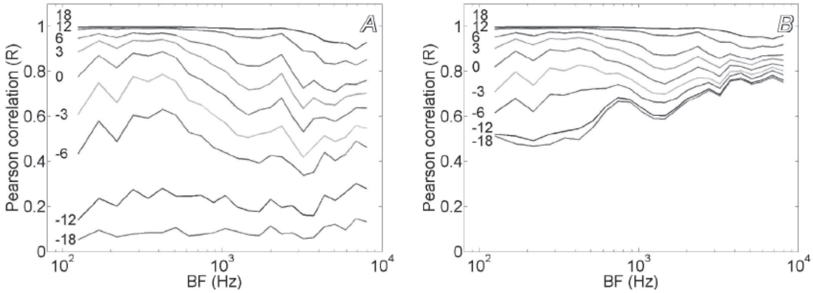

Fig. 1: (A) Correlations between degraded audio envelopes and clean audio envelopes as function of cochlear channel best frequency (BF). The parameter shows the SNR from -18 dB to +18 dB. (B) Correlations between visually enhanced audio envelopes and clean audio envelopes as a function of cochlear channel.

Psychophysical results

The correlation plots (Fig. 1) indicate that the visual enhancement process described above rendered degraded auditory envelopes to be more similar to their clean versions. A pilot experiment tested whether this enhancement process improved those aspects of the envelopes that are important for speech understanding. A vocoder scheme was used (Shannon, 1995), whereby the 135 envelopes – output directly from the peripheral model (audio-alone condition) or following enhancement using the visual channels (visually-enhanced condition) as described above – excited 135 noise bands produced by passing stationary Gaussian white noise through a linear filterbank of sixth order Butterworth filters with center frequencies the same as the peripheral model's

BFs. Auditory signals were passed to the peripheral model in noise at a +6-dB SNR. For each listener, thirty IEEE (1969) sentences each were presented with the vocoder signal created from the noise-degraded audio envelopes, and with the audio envelopes enhanced using the visual inputs as described above. Note that all stimuli were presented to the listeners in the auditory domain. The enhancement provided a significant ($p<0.05$) improvement in mean performance from 42% to 82% of keywords correctly identified, indicating that the AV integration procedure successfully enhanced aspects of the auditory signal that are important for speech understanding.

STMI predictions

The intelligibility of the noise-degraded and visually-enhanced envelopes was estimated using the STMI. Following the procedure of Elhilali *et al.* (2003), a clean-speech template was generated by processing 250 seconds of speech from the TIMIT (Garofolo, 1988) database at 65 dB SPL though the peripheral stage of the model, creating a cochlear spectrogram. The spectrogram was then passed through the cortical spectro-temporal model (Chi *et al.*, 1999), which generates a four-dimensional representation of the input stimulus in terms of the spectral and temporal modulation content as a function of time and tonotopic frequency. A similar procedure was then followed for the IEEE test sentences from the AV integration model described above, with the audio-alone envelope outputs of the peripheral stage, or their visually-enhanced counterparts, passed though the cortical model. This was done for three sentences each at SNRs ranging from -12 to +12 dB in 6-dB steps, plus the quiet condition.

To produce an estimate of speech intelligibility, the cortical representations for the template and each test item were collapsed across tonotopic frequency and time, yielding a two-dimensional representation describing the spectral and temporal modulation content of each stimulus. For each combination of SNR and modality (audio or visually-enhanced audio), the STMI was calculated based on the normalized distance between the template and the test signals in this two dimensional space:

$$\mathrm{STMI} = 1 - \frac{\|\mathrm{T} - \mathrm{X}\|^2}{\|\mathrm{T}\|^2} \qquad \text{(Eq. 2)}$$

where T and X represent the spectrotemporal content of the template and test stimulus, respectively, and $\|\cdot\|$ represents the maximal singular value of the matrix.

Figure 2 shows the results of the STMI analysis (solid and dashed lines), along with intelligibility data for audio and AV IEEE sentences in steady-state noise averaged across three normal-hearing listeners (filled and unfilled circles). The model generally accounts for the main trends in the perceptual data. First, performance generally improves with SNR, which has been previously shown by Elhilali *et al.* (2003). More importantly for the current study, the inclusion of visual information provides a boost to the STMI score. Moreover, the visual benefit diminishes with increasing SNR, as intelligibility scores approach ceiling.

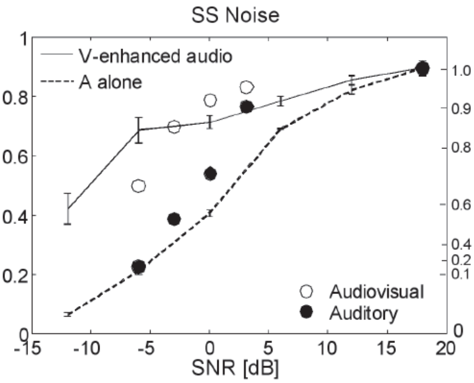

Fig. 2: The STMI outputs of the cortical model in response to audio-alone and visually-enhanced speech envelopes (solid and dashed curves) provide a resoanable qualitative fit to the mean intelligibility data measured in three normal-hearing listeners (circles).

DISCUSSION

The model presented here forms a first attempt at a general model of AV speech intelligibility that makes predictions based on the audio and visual signals alone, akin to the auditory-only AI, SII and STI models, and not on phoneme recognition like previous AV speech models (Massaro, 1998; Braida, 1991). Based on the statistics of the signals alone, the model predictions yield a good qualitative fit to the perceptual data. Further adjustments will be required to improve quantitative fit. For example, tonotopic weights from the SII could be added, or weighting functions could be implemented to place more emphasis on particular spectrotemporal modulation rates and scales. Ultimately, the goal of this modeling effort is to produce individual-specific models of speech intelligibility under audio-alone and AV conditions by applying the AV approach described here to the effort described by Summers *et al.* (2007) to make predictions of speech intelligibility by parameterizing the model of the auditory periphery for individual hearing-impaired listeners.

CONCLUSIONS

Considerable challenges exist for models of speech intelligibility to account for a range of known aspects of speech intelligibility in everyday listening conditions. We have described a model that makes predictions of AV speech intelligibility based on the auditory and visual time waveforms alone, and is able to qualitatively account for the visual benefit to speech understanding. This forms one important step toward a comprehensive model of speech intelligibility.

ACKNOWLEDGMENTS

This research was partially supported by Cooperative Research and Development Agreements between the Clinical Investigation Regulatory Office, U.S. Army Medical Department and the Oticon Foundation, Copenhagen, Denmark. We are grateful to

Dr. Khalil Iskarous and Dr. Doug Whalen of Haskins Laboratory, New Haven, CT, for their assistance with the Optotrak recordings. The authors would also like to thank Dr. Sandeep Phatak and Dr. Jennifer Tufts for helpful comments on an earlier draft of this paper. The opinions and assertions presented are the private views of the authors and are not to be construed as official or as necessarily reflecting the views of the Department of the Army, or the Department of Defense.

REFERENCES

ANSI (**1969**). Methods for the calculation of the articulation index, S3.5 (American National Standards Institute, New York).

ANSI (**1997**). Methods for calculation of the speech intelligibility index, S3.5 (American National Standards Institute, New York).

Bernstein, J. G. W, and Grant, K. W. (this volume). “Frequency importance functions for audiovisual speech and complex noise backgrounds,” in Proceedings of the International Conference of Auditory and Audiological Research, Helsingör, Denmark, August, 2007.

Berthommier, F. (**2004**). “A phonetically neutral model of the low-level audio-visual interaction," Speech Comm. **44**, 31-41.

Boothroyd, A., and Nittrouer, S. (**1988**) “Mathematical treatment of context effects in phoneme and word recognition,” J. Acoust. Soc. Am., **84**, 101-114.

Braida, L. D. (**1991**). “Crossmodal integration in the identification of consonant segments,” Quarterly J. Exp. Psych. **43**, 647-677.

Bregman, A. S. (**1990**). “Auditory scene analysis: The perceptual organization of sound,” MIT Press, Cambridge, MA.

Brungart, D. S. (**2001**). “Informational and energetic masking effects in the perception of two simultaneous talkers,” J. Acoust. Soc. Am. **109**, 1101-1109.

Chi, T., Gao, Y., Guyton, M. C., Ru, P., and Shamma, S. (**1999**). “Spectro-temporal modulation transfer functions and speech intelligibility,” J. Acoust. Soc. Am. **106**, 2719-2732.

Elihilali, M., Taishih, C., and Shamma, S. A. (**2003**). “A spectro-temporal modulation index (STMI) for assessment of speech intelligibility,” Speech Comm. **41**, 331-348.

Festen, J. M., and Plomp, R. (**1990**). “Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing,” J. Acoust. Soc. Am. **88**, 1725-1736.

French, N. R., and Steinberg, J. C. (**1947**). “Factors governing the intelligibility of speech sounds,” J Acoust. Soc. Am. **19**, 90-119.

Garofolo, J. (**1988**). “Getting Started with the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database,” National Institute of Standards and Technology NIST, Gaithersburg, MD.

George, E. L. J., Festen, J. M., and Houtgast, T. (**2006**). “Factors affecting masking release for speech in modulated noise for normal-hearing and hearing-impaired listeners,” J. Acoust. Soc. Am., **120**(4), 2295-2311.

Girin, L., Schwartz, J.-L., and Feng, G. (**2001**). "Audio-visual enhancement of speech in noise," J. Acoust. Soc. Am. **109**, 3007-3020.

Glasberg, B. R., and Moore, B. C. J. (**1989**). “Psychoacoustic abilities of subjects with unilateral and bilateral cochlear impairments and their relationship to the ability to understand speech,” Scand. Audiol. Suppl., **32**, 1-25.

Grant, K. W. (**2001**). “The effect of speechreading on masked detection thresholds for filtered speech,” J. Acoust. Soc. Am. **109**, 2272-2275.

Grant, K. W. (**2006**). “Frequency-band importance functions for auditory and auditory-visual speech by hearing-impaired adults,” J. Acoust. Soc. Am. **119**, 3416-3417.

Grant, K. W., and Bernstein, J.G.W. (**2007**). “Frequency band-importance functions for auditory and auditory-visual sentence recognition,” J. Acoust. Soc. Am. **121**, 3044 (A).

Grant, K. W., Elhilali, M., Shamma, S. A., Walden, B. E., Surr; R. K., Cord, M. T., and Summers, V. (in press). “An objective measure for selecting microphone modes in OMNI/DIR hearing-aid circuits,” Ear Hear.

Grant, K. W., and Seitz, P. (**2000**). “The use of visible speech cues for improving auditory detection of spoken sentences,” J. Acoust. Soc. Am. **108**, 1197-1208.

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

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

Lopez-Poveda, E. A., and Meddis, R. (**2001**). “A human nonlinear cochlear filterbank,” J. Acoust. Soc. Am. **110**, 3107-3118.

Massaro, D. W. (**1998**). ”Perceiving talking faces: From speech perception to a behavioral principle,” (MIT Press, Cambridge).

Oxenham, A. J., and Dau, T. (**2001**). “Modulation detection interference: Effect of concurrent and sequential streaming,” J. Acoust. Soc. Am., **110**, 402-408.

Plomp, R. (**1978**). “Auditory handicap of hearing impairment and the limited benefit of hearing aids,” J. Acoust. Soc. Am. **63**, 533-549.

Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (**2006**). “Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise,” J. Acoust. Soc. Am. **120**, 3988-3997.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303-304.

Steeneken, H. J. M., and Houtgast, T. (**1980**). “A physical method for measuring speech-transmission quality,” J. Acoust. Soc. Am. **67**, 318-326.

Summers, V., Makashay, M. J., Grassi, E., Grant, K. W., Bernstein, J. G. W, Walden, B. E., Leek, M. R., and Molis, M. R. (this volume). “Toward an individual-specific model of impaired speech intelligibility,” in Proceedings of the International Conference of Auditory and Audiological Research, Helsingör, Denmark, August, 2007.

Yost, W. A., and Sheft, S. (**1994**). “Modulation detection interference: Across-frequency processing and auditory grouping,” Hear. Res., **79**, 48-58.

Zilany, M. S. A., and Bruce, I. C. (**2007**). “Predictions of speech intelligibility with a model of the normal and impaired auditory-periphery,” in Proceedings of the 3rd International IEEE/EMBS Conference on Neural Engineering, Kohala Coast, HI, 481-485.