The Effect of Speechreading on Masked Detection Thresholds for Filtered Speech

by

Ken W. Grant
Walter Reed Army Medical Center
Army Audiology and Speech Center
Washington, DC 20307

Address all correspondence to: grant@tidalwave.net

ABSTRACT

Detection thresholds for spoken sentences in steady-state noise are reduced by 1-3 dB when synchronized video images of movements of the lips and other surface features of the face are provided. In an earlier study [K.W. Grant and P.F. Seitz, J. Acoust. Soc. Am. **108**, 1197-1208 (2000)], we showed that the amount of masked threshold reduction, or *bimodal coherence masking protection* (BCMP), was related to the degree of correlation between the rms amplitude envelope of the target sentence and the area of lip opening, especially in the mid-to-high frequencies typically associated with the second (F2) and third (F3) speech formants. In the present study, we extend these results by manipulating the cross-modality correlation through bandpass filtering. Two filter conditions were tested corresponding roughly to the first and second speech formants: F1 (100-800 Hz) and F2 (800-2200 Hz). Results for F2-filtered target sentences were comparable to those of unfiltered speech, yielding a BCMP of roughly 2-3 dB. Results for F1-filtered target sentences showed a significantly smaller BCMP of approximately 0.7 dB. These results suggest that the magnitude of the BCMP depends on both the spectral and temporal properties of the target speech signal in the mid frequencies.

PAC numbers: 43.66.Dc, 43.66.Mk

I. INTRODUCTION

Auditory-visual speech recognition is the most robust form of speech communication used by humans or machines, especially in noisy environments. Further understanding of the processes involved in human auditory-visual speech recognition will likely assist in the development of hearing aids that are more resistant to environmental noise and reverberation, as well as automatic speech recognition systems that are capable of achieving high rates of accuracy with speech spoken conversationally. Advancements in these two areas would have benefits for hearing-impaired individuals as well as normal-hearing persons who need to communicate in noisy or reverberant environments. This study addresses some fundamental questions regarding how vision and audition interact during speech recognition, and seeks to understand further the mechanisms by which watching a talker's lips and face during speech production helps to reduce interference from background noise. For automatic speech recognition, being able to segregate the speech signal from other extraneous sound sources, including speech from other talkers, is an essential first step in decoding the target speech signal into meaningful units. Understanding the perceptual mechanisms that allow human observers to accomplish this task through the use of speechreading should be of obvious benefit in transferring this knowledge to machine applications.

Recently, we have demonstrated that watching the movement of the lips and face can also improve the *detection* of speech (Grant and Seitz, 2000). Depending on the specific sentence, the improvement in detection thresholds can be between 1-3 dB. For all three sentences tested, a significant difference in thresholds was obtained between auditory-visual and auditory conditions, with greater sensitivity demonstrated for the auditory-visual conditions. There were also

significant differences in thresholds across sentences. While the overall amount of masking protection due to visual speech cues may seem relatively small, a 1-3 dB improvement in masked thresholds may nevertheless lead to recognition of specific speech cues and understanding of a spoken message that might otherwise have been inaudible, and therefore, incomprehensible (Sumby and Pollack, 1954).

To account for their results, Grant and Seitz (2000) assumed that subjects have the ability to correlate the visible movements of the speech articulators (e.g., variation in the area of lip opening during speech production) and the acoustic speech envelope computed over time intervals corresponding roughly to the average duration of a syllable (333 ms). When this correlation is high (e.g., greater than 0.9) and the amplitude envelope is at a maximum relative to other peaks in the speech sample, there will be a positive effect of speechreading on detection thresholds. Grant and Seitz called this effect on speech detection thresholds *bimodal coherence masking protection* (BCMP), adapted from the earlier work of Gordon (1997a,b). The label *BCMP* is used to denote the fact that the information from one modality (in this case, visual) partially protects the target speech signal from the deleterious effects of noise. In the present study, the effects of speechreading on speech detection were explored further by manipulating the degree of correlation between the peak amplitude locations in the speech waveform and the lip-area function. This was accomplished by filtering target sentences using bandpass filters with different center frequencies and bandwidths. Filtering the target signals alters the relative energies of specific frequency regions of the speech. The purpose of the present experiment was to test the hypothesis that speechreading aids auditory detection of spoken sentences when the amplitude-

envelope peaks of the sentence coincides with a temporal location of high cross-modal correlation between area of lip opening and amplitude envelope. Since the correlation between area of mouth opening and speech amplitude envelope tends to be greatest for mid- to high-frequency speech signals (Grant and Seitz, 2000), we predicted that the magnitude of the BCMP would be greater for bandpass-filtered speech targets containing mid-frequency energy than for bandpass-filtered speech targets containing low-frequency energy.

II. METHODS

A. Subjects

Six normally hearing subjects (mean age = 37.7 years) participated in the study. Subjects were screened to assure pure-tone air-conduction thresholds bilaterally of ≤ 20 dB HL at audiometric test frequencies 0.25-4.0 kHz and ≤ 30 dB HL at 6.0 kHz (ANSI, 1989). All subjects were native English speakers with normal or corrected-to-normal vision (visual acuity equal to or better than 20/30 as measured with a Snellen chart). Eligible subjects were paid $10.00 per hour to compensate them for their for their participation.

B. Stimuli

Speech materials consisted of video-recorded spoken sentences from the IEEE/Harvard sentence corpus (IEEE, 1969). The visual portion of each sentence was 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), filtered (8.5 kHz), normalized in level, and stored on a personal computer. Two sentences and their variants (as described below) were used. These were “Both brothers wear the same size” and “Watch the log float in the wide river”. In the previous study

(Grant and Seitz, 2000), these two sentences (denoted here as sentence 2 and 3 to be consistent with the previous study) provided approximately 1 dB and 2.5 dB of BCMP, respectively.

The target sentences were digitally bandpass filtered using FIR filters with greater than 100 dB/oct attenuation outside the passband. Two bandpass filters were applied to each target sentence. One filter was centered primarily on the F1 speech region (100-800 Hz) whereas the second filter was centered primarily on the F2 speech region (800-2200 Hz). After filtering, the target audio sentences were scaled in amplitude so that the average rms levels were equivalent.

The amplitude peaks in the filtered speech spectrum are the most likely locations in time where signal detection will take place. In order for the visual speech signal to aid the listener in detecting these filtered speech waveforms, the movements of the visible articulators must provide cues to the temporal and possibly spectral locations of the most prominent envelope peaks. It has been shown previously (Grant and Seitz, 2000) that the correlation between the area of mouth opening and speech amplitude envelope is greatest in the F2 region and lowest in the F1 region. By filtering speech using bandpass filters centered on each of these two formant regions, it is predicted that a large BCMP would result for F2-filtered speech whereas small BCMP would result for F1-filtered speech.

C. Procedure

Subjects were tested binaurally under headphones (Beyer Dynamic DT770) in a sound-treated booth using an adaptive two-interval forced-choice (2IFC) tracking procedure. Masked thresholds for detecting speech were obtained under both auditory alone (A) and auditory-visual (AV) conditions. Each test block consisted of multiple interleaved tracks corresponding to the two

different filtered target sentences and different test conditions (auditory and auditory-visual). The masking noise consisted of a white noise lowpass filtered at 8.5 kHz whose duration was equal to the target sentence plus a random amount (additional 200-800 ms). The target sentence was temporally centered in the noise. For the AV conditions, video speech information of the talker saying the target sentence was presented in both observation intervals. In other words, the identical visual speech information was provided during the noise alone and the noise-plus-speech intervals. The filtered audio signals were manually realigned with the corresponding unprocessed audio signals on the optical disk to account for any audio delays that might have been imposed by the filtering process. Video signals were displayed on a 19-inch color monitor (SONY PVM 2030) positioned approximately 5 feet from the subject.

The subject's task was to identify the interval containing the target auditory sentence. The speech signal level was held constant at approximately 50 dB SPL. The intensity of the noise masker varied independently for each track according to a 3-up, 1-down adaptive tracking procedure targeting the 79% point on the psychometric function (Levitt, 1971). Each track was controlled independently and selected randomly on each trial. The initial step size of the digital noise attenuator (TDT PA4) was 3 dB during the first three reversals in the direction of the track. At that point the step size changed to 1 dB for an additional six track reversals. Threshold estimates for each track were computed as the mean of the noise levels obtained on the last six reversal points. Final threshold values for each of the target sentences in each of the conditions were the average of three separate threshold estimates. If the standard error of the three estimates

exceeded 1 dB, a fourth estimate was obtained and the final threshold value was the average of all four estimates.

Because there were two target sentences, two filter conditions (F1 and F2), and two modalities (A and AV), a total of eight interleaved tracks were run on each test block. Each test block took approximately 1.5 hours to complete, allowing for frequent rest periods.

III. RESULTS

The mean speech-detection threshold, expressed in terms of the speech-to-noise ratio, was -22.36 dB under auditory-visual conditions and -21.13 dB under auditory conditions demonstrating a small but consistent BCMP. Figure 1 shows the magnitude of the BCMP (i.e., the difference in dB between auditory and auditory-visual detection thresholds) for each target sentence as well as the average BCMP across sentences (bars labeled AV_{WB} and AV_{O} conditions will be described later). As observed in the figure, both F1- and F2-filtered sentences resulted in a BCMP, but the magnitude of the BCMP for F2-filtered sentences was larger than that for the F1-filtered sentences (average AV_{F1} = 0.76, average AV_{F2} = 1.70) and the magnitude of the F2-filtered BCMP for sentence 3 was greater than that for sentence 2 (AV_{F2} = 0.98 dB for sentence 2 and 2.42 for sentence 3). Multiple t-tests with Bonferroni adjusted probabilities were conducted to test whether the magnitude of the BCMP for each sentence and each filter condition was significantly greater than zero (i.e., no difference in masked threshold for auditory and auditory-visual conditions). Results showed that only the F2-filtered targets yielded significant BCMPs (t = 4.45, p = 0.027 and t = 5.9, p = 0.008 for sentence 2 and 3, respectively).

To further analyze these effects, a repeated measures ANOVA with BCMP as the dependent variable and filter condition and sentence as factors revealed a significant effect for filter [$F(1,5) = 6.9$, $p = 0.047$]. Neither sentence nor the interaction between sentence and filter was significant, although the difference across sentences approached significance ($p = 0.077$).

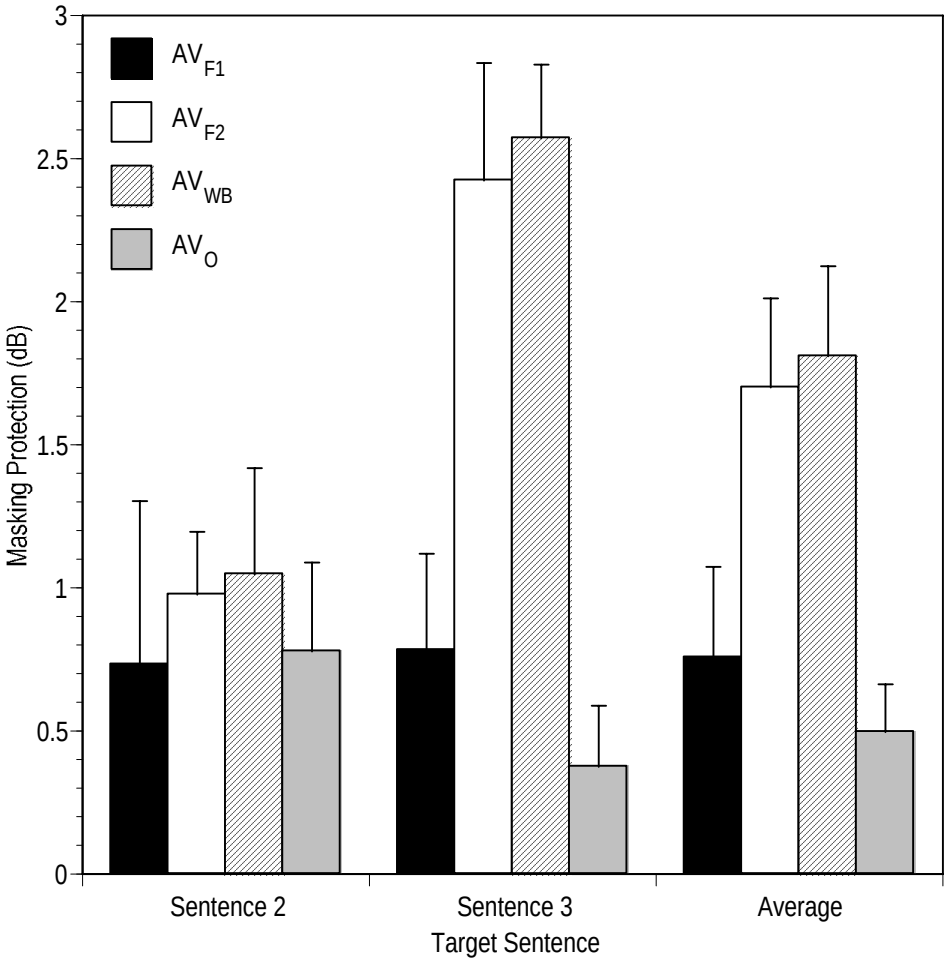

Figure 1. Difference in auditory and auditory-visual masked detection thresholds (masking protection) for spoken filtered sentences. AV_{F1} = Auditory-visual presentation of speech filtered between 100-800 Hz; AV_{F2} = Auditory-visual presentation of speech filtered between 800-2200 Hz; AV_{WB} = Auditory-visual presentation of wideband speech (100-8500 Hz); AV_O = Auditory presentation of wideband speech preceded by visual orthography. Error bars show 1 standard deviation.

Figure 1 also compares the present BCMP results with those reported earlier by Grant and Seitz (2000) obtained by the same six subjects for wideband target sentences (AV_{WB}) and for sentences presented auditorily with the aid of an orthographic display (AV_O) informing the subject of the exact text of the target sentence presented on each trial. The orthographic display was presented just prior to each test trial and lasted for 0.5 seconds. Knowing the text of the target sentence resulted in a small BCMP (approximately 0.5 dB), probably due to a slight reduction in informational masking and stimulus uncertainty (Watson and Kelly, 1981). The BCMP for low-frequency speech (F1 filtering) was only slightly greater than that for orthography, and like orthography, appears to be roughly independent of the target sentence. The BCMP for wideband speech and for F2-filtered speech signals were nearly identical, suggesting that when presented with wideband speech targets in conjunction with speechreading, the listener/observer extracts the cross-modality coherence between variations in visible facial kinematics and the acoustic amplitude envelope derived mostly from mid-frequency spectral channels.[1]

Grant and Seitz (2000) suggested that the presence of significant BCMP for all target sentences and observed differences in BCMP magnitude across target sentences may be explained by the degree of correlation between the temporal envelope of the target sentence and the kinematic variation of the area of mouth opening during the production of the target sentence. Correlations between amplitude envelope and area functions over the course of a whole sentence known to produce a significant BCMP were previously observed to be only about 0.5. However, Grant and Seitz also noted that the detection of a speech event only requires a very brief moment of the target signal to be audible. They suggested that the whole sentence correlation may not be

as relevant as a "local" correlation using a sliding window of approximately syllable length (333 ms) updated every 33 ms (i.e., the duration of each video frame). Figure 2 shows these local correlations (thick line) along with the speech amplitude envelope (thin line) for the four target signals. The top two panels show the results for sentence 2 whereas the bottom two panels show results for sentence 3. On the left side of the figure are F1-filtered targets and on the right are F2-filtered targets. According to Grant and Seitz (2000), the relevant information in each panel necessary to explain the presence of significant BCMP is the temporal location of the greatest amplitude envelope peaks comprising the top 2% of the amplitude range (indicated by arrows) along with the local correlation at these moments in time. Note that for both sentences, the correlations are higher in the vicinity of amplitude-envelope peaks for the F2-filtered targets than for F1-filtered targets, and that for sentence 3 in particular, all F2-filtered peak-amplitude locations are associated with relatively high correlations. Focusing on the six peak locations identified for the F1-filtered targets, the average correlation between lip-area function and acoustic amplitude envelope was 0.34 (sd = 0.32), whereas the average correlation between lip-area function and acoustic amplitude envelope correlation for the five peak locations identified for the F2-filtered targets was 0.82 (sd = 0.11).

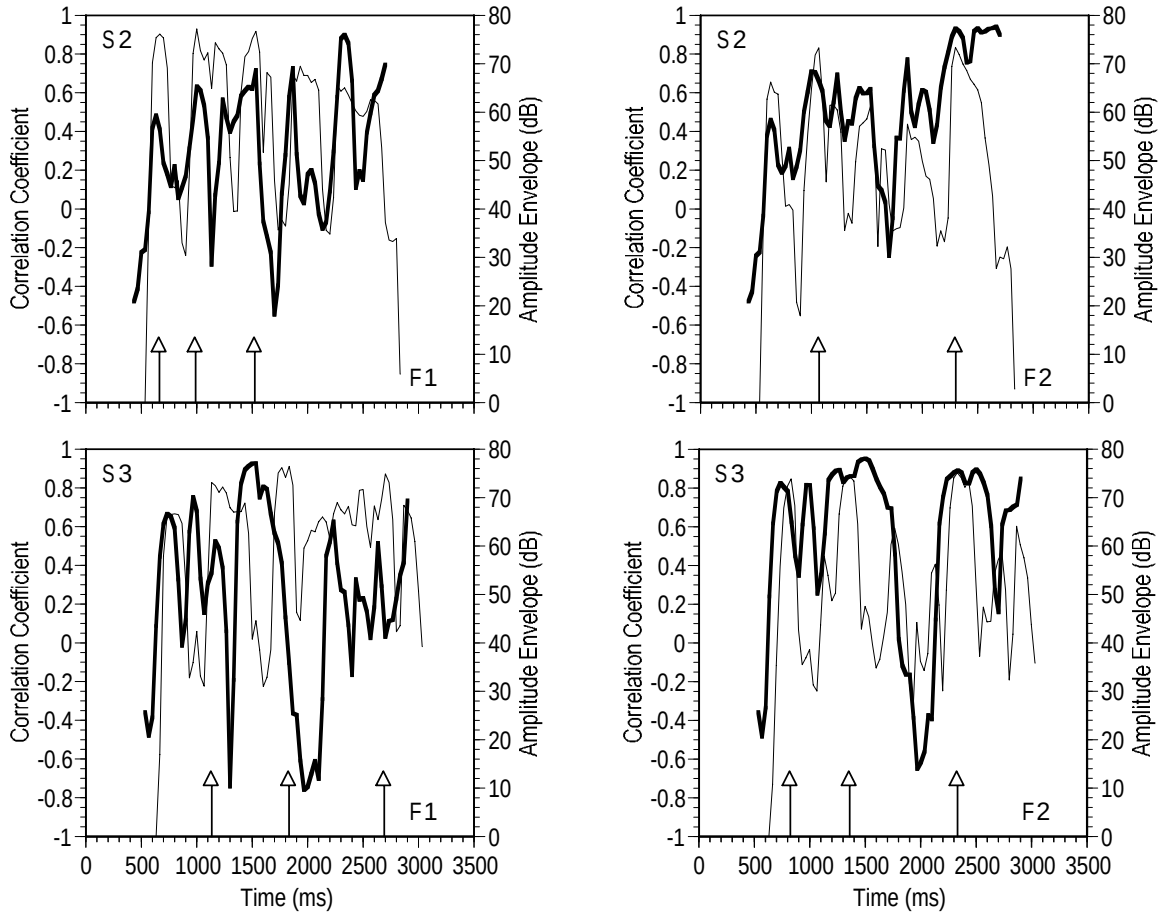

Figure 2. Local correlations (bold line) between area of mouth opening and speech amplitude envelope (see text for explanation). For each panel, the left axis shows the Pearson product-moment correlation whereas the right axis shows the rms energy of the speech target signal averaged over successive 33-ms rectangular windows (thin line). Arrows show the temporal locations of the most prominent amplitude peaks in each target sentence (presumably the temporal regions most likely responsible for the detection response). The top two panels are for sentence S2 "Both brothers wear the same size". The bottom two panels are for the sentence S3 "Watch the log float in the wide river". The left two panels are for F1-filtered targets. The right two panels are for F2-filtered targets.

IV. DISCUSSION AND CONCLUSIONS

The data presented in this study support previous findings that speechreading provides useful information that can be used to cue listeners to the temporal and spectral locations of high energy regions in speech signals. Essentially, watching the variations in the movement of the

mouth during speech production informs listeners both *when* in time and *where* in the spectrum to expect signal energy. By focusing the attention of the listener to specific spectro-temporal locations in the speech waveform, the ability to hear speech in noisy backgrounds is improved.

Correlation data (Grant and Seitz, 2000) comparing the acoustic envelope from various spectral regions of speech to variations in the area of lip opening during speech production show that amplitude envelopes from the F2 speech region appear to have greater coherence with visible oral kinematics than envelopes derived from other spectral regions (e.g., F1, F3, and wideband). It is hypothesized that this coherence is due primarily to the association between changes in the place of constriction in the front cavity of the vocal tract and rapid changes in F2 formant frequency (Stevens and House, 1955; Stevens, 1998). Furthermore, the fact that there is greater modulation in F2 frequency than in F1 frequency with changes in place of articulation suggests that the acoustic energy observed at the output of a filter with a fixed bandwidth centered in the F2 speech region (as used in the present study) might also be highly modulated. Thus, there appears to be a natural linkage between front cavity area and lip shape and the acoustic fluctuations in energy associated with frequencies in the F2 region. Moreover, this information appears to be at least partially available through speechreading and can be used to improve speech detection thresholds in noise.[2]

For automatic speech recognition (ASR) and noise-reduction algorithms for speech enhancement, these findings offer new and potentially interesting possibilities. For example, it may be possible to construct a temporal filter based on variations in inter-lip distance or area of mouth opening that estimates the amplitude envelope of mid-frequency speech bands. Such a

time-varying filter could then be used to process the acoustic environment so as to enhance probable speech signals and reduce extraneous acoustic signals that are unrelated to the visual speech dynamics. Recent experiments in ASR have demonstrated that combining mouth shape with acoustic data can improve recognition performance with noisy speech (Girin et al., 1998, in press).

The integration of visual and acoustic speech information allows for robust and reliable speech recognition that is greatly resistant to noise and reverberation. Bisensory integration appears to proceed rather automatically and with a fairly high degree of efficiency (Braida, 1991; Grant and Seitz, 1998; Massaro, 1998; Massaro and Cohen, 2000), especially for normal-hearing subjects. The question of just where in the speech perception process auditory-visual integration takes place has been a topic of much discussion. The present data, as with the previous study by Grant and Seitz (2000), demonstrate that physical correspondence between visible speech kinematics and acoustic modulations of speech output provides an opportunity for auditory and visual speech data to merge at the level of signal detection. Whether this represents an extremely early phase of auditory-visual integration is unclear, since it is possible that improvements in masked thresholds, as evidenced by significant BCMP, may occur for more central reasons (Watson and Kelly, 1981). However, it does bring into question whether it is possible to model auditory-visual speech processing strictly in terms of independent auditory and visual speech processes.

ACKNOWLEDGMENTS

This work was supported by Grant Number DC00792 from the NIDCD and the Department of Clinical Investigation, Walter Reed Army Medical Center under Work Unit #2590-99. I wish to thank Sid Bacon, Andrew Faulkner, and an anonymous reviewer for their helpful comments. A preliminary report of this work was presented at the 139th Meeting of the Acoustical Society of America, Atlanta, GA, 2000. All subjects participating in this research 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.

REFERENCES

American National Standards Institute (1989). "Specifications for Audiometers". ANSI S3.6-1989, American National Standards Institute, New York.

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

Girin, L., Feng, G. and Schwartz, J.-L. (1998). "Fusion of auditory and visual information for noisy speech enhancement: a preliminary study of vowel transitions", Proc. ICASSP'98, Seattle, WA, pp. 1005-1008.

Gordon, P.C. (1997a). "Coherence masking protection in brief noise complexes: Effects of temporal patterns," J. Acoust. Soc. Am. **102**, 2276-2283.

Gordon, P.C. (1997b). "Coherence masking protection in speech sounds: The role of formant synchrony," Percept. Psychophys. **59**, 232-242.

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

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.

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

Levitt, H. (1971). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467-477.

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

Massaro, D.W., and Cohen, M.M. (2000). "Tests of auditory-visual integration efficiency within the framework of the fuzzy logical model of perception," J. Acoust. Soc. Am. **108**, 784-789.

Stevens, K.N. (1998). *Acoustic Phonetics*. MIT Press, Cambridge, MA

Stevens, K.N., and House, A.S. (1955). "Development of a quantitative description of vowel articulation," J. Acoust. Soc. Am. **27**, 484-493.

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

Watson, C.S., and Kelly, W.J. (1981). "The role of stimulus uncertainty in the discrimination of auditory patterns," in *Auditory and Visual Pattern Recognition,* edited by D.J. Getty and J.H. Howard, Jr., Lawrence Erlbaum Associates, Hillsdale, NJ., pp. 37-59.

FOOTNOTES

1. One caveat in interpreting these results is that the F1 and F2 filters had very different bandwidths (700 Hz versus 1400 Hz, respectively). Thus, there are significant differences in spectrum level across the two filter conditions. However, for present purposes, these differences in spectrum level probably have little or no impact on the current results. First, the peak amplitude levels for F1 and F2 target sentences were quite similar (within 2 dB) after the filter outputs were scaled to have the same overall rms level. Second, the primary comparison across filter conditions is made using BCMP measures, which are themselves direct comparisons between auditory and auditory-visual threshold values within a given sentence and a given filter condition. Thus, any level differences due to filtering are probably unimportant because the BCMP measures are primarily a comparison across conditions (A versus AV) and not filter.

2. It has been argued here and elsewhere (Grant and Seitz, 2000) that BCMP is attributed primarily to the correlation between local peaks in the speech amplitude envelope and visible lip shape information. Another possibility is that there is an association between speechreading and rapid changes in F2 frequency often observed at boundaries between consonants and vowels. These sudden changes in F2 frequency, as opposed to local amplitude peaks, may be the cue underlying the observed BCMP. However, it is important to remember that at detection threshold, so little of the speech signal is available due to the very poor signal-to-noise ratio, that it seems more plausible to consider amplitude peaks, rather than F2 frequency, as the primary cue.