THE USE OF VISIBLE SPEECH CUES (SPEECHREADING) FOR DIRECTING AUDITORY ATTENTION: REDUCING TEMPORAL AND SPECTRAL UNCERTAINTY IN AUDITORY DETECTION OF SPOKEN SENTENCES

Ken W. Grant and Philip F. Seitz

Army Audiology and Speech Center, Walter Reed Army Medical Center, Washington D.C. 20307-5001

Presented at the 16th International Congress on Acoustics and the 135th Meeting of the Acoustical Society of America, Seattle, Washington 20-26 June 1998

ABSTRACT

Classic accounts of the benefits of speechreading to speech recognition treat auditory and visual channels as independent sources of information that are integrated early in the speech perception process, most likely at a pre-categorical stage [1,14,21-22]. The question addressed in this study was whether visible movements of the speech articulators could be used to improve the detection of speech in noise, thus demonstrating an influence of speechreading on the processing of low-level auditory cues. Nine normal-hearing subjects detected the presence of spoken sentences in noise under three conditions: auditory-only (A), auditory-visual with a visually matched sentence (AV_M), and auditory-visual with a visually unmatched sentence(AV_{UM}). When the video matched the target sentence, detection thresholds improved by about 1.5 dB relative to the auditory-only and auditory-visual unmatched conditions. The amount of threshold reduction varied significantly across target sentences (from 0.73 to 2.19 dB). Analysis of correlations within each sentence between area of mouth opening and four acoustic envelopes (wideband, F1 region, F2 region, and F3 region) suggested that AV_M threshold reduction is conditioned by the degree of auditory-visual spectral comodulation.

INTRODUCTION

Past studies have demonstrated the benefits of auditory-visual (AV) speech perception over either listening alone or speechreading alone. The addition of visual cues can be effectively equivalent to an improvement in the speech-to-noise ratio (S/N) of as much as 10 dB for spondaic words [21], and about 4-5 dB for more difficult connected speech materials such as the IEEE/Harvard sentence lists [6,11]. Since each 1-dB improvement in S/N corresponds roughly to a 10 percent increase in intelligibility [6,15], the addition of speechreading can mean the difference between failure to understand and near perfect comprehension, especially in noisy environments.

The relationship between the intelligibility benefit provided by speechreading and the type of speech information provided by independent auditory and visual inputs has been studied extensively [1,6,8-9,14,20-22]. In this study, we focus on the potential importance of cross-modality temporal comodulation between the acoustic speech signal and visible movements of the talker's lips for detecting speech in noise.

Repp et al. [16] were the first to examine the potential influence of speechreading on the detection of acoustic speech signals. Their results failed to show a change in detection sensitivity thresholds. In our opinion, this result can be traced to a limited accuracy in synchronizing A and V stimulus components and, more importantly, in the selection of a speech modulated masker which itself was comodulated with the visible speech signals. The failure to accurately synchronize A and V stimulus components weakened the correlation between the two signals, which is known to affect the amount of masking release [12]. Furthermore, by using a masking signal that was modulated by the speech envelope, the chances of showing a *release from masking* was greatly reduced since the speech target signal, the comodulated visual signal, and the masker all had the same temporal structure. In the present study these problems were eliminated by using a continuous noise masker having no temporal comodulation with the speech signals and equipment capable of precise auditory-visual alignments within ±2 ms. The primary question addressed in this study was whether the detectability of a masked speech signal is improved by the addition of temporally comodulated visual speech information[5,21-22].

EXPERIMENT 1: AUDITORY AND AUDITORY-VISUAL SPEECH DETECTION THRESHOLDS

METHODS

Masked thresholds for detecting speech were obtained from nine normal-hearing subjects under three conditions: auditory alone (A), auditory-visual with matching visual stimulus (AV_M), and auditory-visual with mismatched visual stimulus (AV_{UM}). In the AV_M condition, target audio sentences were presented along with simultaneous congruent visual lipread information. In the AV_{UM} condition, target audio sentences were presented along with simultaneous incongruent visual lipread information. Six sentences from the IEEE/Harvard sentence lists served as stimuli [11]. Three sentences were used as auditory targets. The matching video from these sentences were used in the AV_M conditions whereas the video from three different sentences was selected randomly and used in the AV_{UM} conditions. For both AV conditions, video lipread information was available equally in both observation intervals. Subjects were tested binaurally under headphones using a two-alternative forced-choice tracking procedure with nine interleaved tracks (3 conditions x 3 target sentences). For each track, a target sentence plus white noise was presented in one interval. In the other interval, only the noise was presented. When the interval contained both signal and noise, the noise was started 100-300 ms before the signal and continued for 100-300 ms after the signal ended. The duration of the noise lead was varied randomly from trial to trial to increase temporal uncertainty.

The subject's task was to identify the interval containing the sentence. The speech signal level was held constant at approximately 50 dB SPL. The intensity of the white noise masker was varied independently for each track according to a 3-down, 1-up adaptive tracking procedure using an initial step size of 3 dB and a final step size of 1-dB [13]. Threshold estimates for each track were computed as the mean of the noise levels between reversal points for each of the last six ascending runs.

RESULTS AND DISCUSSION

Figure 1 shows the masked thresholds for each condition and each sentence in terms of the S/N ratio averaged across subjects. For each sentence, there was a reduction in masked threshold for the AV_M condition relative to the other two conditions. No difference in S/N was observed between the A and AV_{UM} conditions. That is, the presentation of uncorrelated V information does not have a negative effect on auditory detection thresholds.

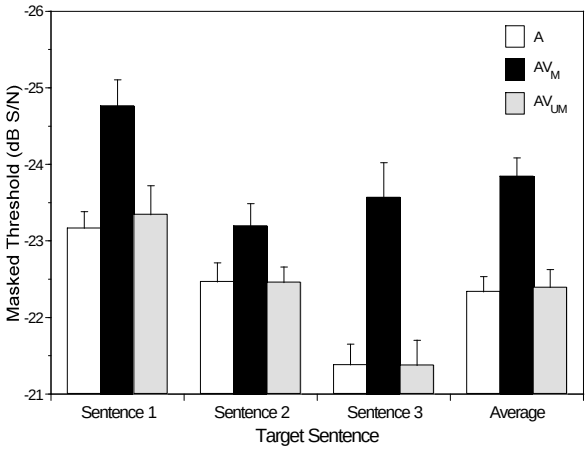

Figure 1. Signal-to-Noise (S/N) ratio for speech detection thresholds as a function of listening condition and target sentence. A = Audio Only; AV_M = Matching Video; AV_{UM} = Mismatched Video. Error bars show +1 standard error.

Using the detection thresholds obtained from the A condition as a reference, we computed the average masking difference for the AV_M and AV_{UM} conditions as well as the masking difference for each of the three target sentences (Figure 2). A masking level difference, or *bimodal coherence masking protection* (BCMP), was observed for all three sentences in the AV_M condition (black bars). In contrast, there was no BCMP for the AV_{UM} condition (striped bars). The average BCMP obtained in the AV_M condition was 1.5 dB. The range of BCMP for the three sentences was 0.73 to 2.19 dB.

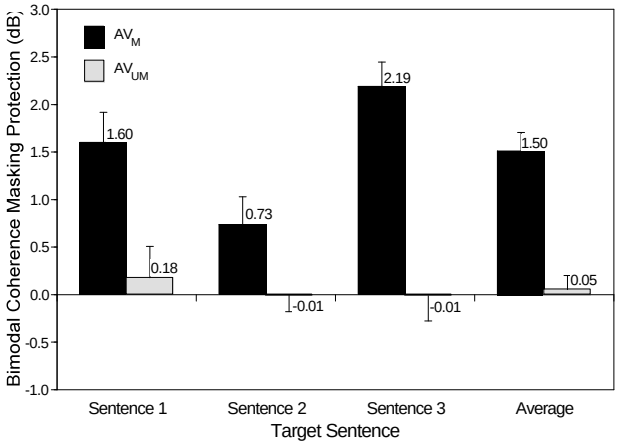

Figure 2. Bimodal coherence masking protection in dB relative to the A only condition. AV_M = Black Bars; AV_{UM} = Striped Bars. Error bars show +1 standard error.

To assess the statistical significance of these effects a repeated measures analysis of variance (ANOVA) was run with condition and sentence as within-subject trial factors. The main effect of condition [$F(2,16) = 40.94$, $p < 0.0001$] and sentence [$F(2,16) = 31.42$, $p < 0.0001$] was significant. The interaction of condition and sentence was also significant [$F(4,32) = 5.02$, $p = 0.003$]. Post hoc comparisons confirmed that the A and AV_{UM} conditions required a significantly greater speech-to-noise ratio than did the AV_M condition for all three sentences and that the

amount of BCMP observed in the AV_M condition was greater for sentence 3 than for sentence 2. The difference in BCMP observed between the A and AV_M condition for sentences 1 and 2 or for sentences 1 and 3 were not significant. These data show that cross-modality comodulation can offer protection from noise masking in much the same manner as shown by Gordon [5] for acoustic stimuli in what has been dubbed *coherence masking protection* (CMP). The magnitude of this protection (roughly 1.5 dB) is consistent with a reduction of temporal uncertainty observed in earlier signal detection experiments when a light or other visual signal is used to mark the onset of an acoustic signal masked by noise [3-4,23]. The strength of the effect, however, may depend on the degree of correlation between the A and V signals. In the following experiment, we asked whether the correlation between area of mouth opening and acoustic energy in various bands of speech can be used to determine the degree of BCMP.

EXPERIMENT 2: CORRELATION BETWEEN ACOUSTIC AND VIDEO MEASURES

BACKGROUND

The *bimodal coherence masking protection* (BCMP) observed in the Experiment 1 might be attributed to the comodulation that exists between visible movements of the speech articulators and acoustic variations in energy over time. The detection response depends on some part of each sentence, presumably the region with the best S/N ratio, rising above the noise masker. The sentences differed with respect to the temporal location of the region with the best S/N. For S1 ("To make pure ice you freeze water"), the point of detection was in the vicinity of the phrase "pure ice". For S2 ("Both brothers wear the same size"), it was near the phrase "both brothers". For S3 ("Watch the log float in the wide river") it was near the phrase "log float". In the A condition it is impossible to distinguish which sentence is being presented when near threshold, and the point of detection (in time) is uncertain. Under the AV_M condition, the visual information can inform the subject which sentence is being presented and may allow the subject to follow along until the point of detection is reached. Under these conditions, the subject can focus their auditory attention more efficiently thereby reducing stimulus uncertainty.

If this account is correct, we would expect that the degree of correlation between visible movement of the articulators and variations in acoustic energy over time should differ for different sentences, being greater for those sentences that give rise to more BCMP. In this experiment, we measured the area of lip opening and correlated this with the acoustic envelope fluctuations of each sentence. In addition, given that it is quite plausible that energy fluctuations in each sentence may be different for different bands of speech, we also calculated the envelopes for three filtered bands roughly corresponding to the first three formant regions. Based on past results from speechreading experiments showing that subjects extract primarily place-of-articulation cues (known to be associated primarily with F2 transitions), we predict that visible lip area functions will be correlated best with the F2 envelope band and correlated least with the F1 envelope band.

METHODS

Individual video frames from sentence 3 (“Watch the log float in the wide river”) and sentence 2 (“Both brothers wear the same size”) were digitized and analyzed for area of mouth opening, height of mouth opening, and width of mouth opening. These two sentences were chosen because they resulted in the greatest (S3) and least (S2) BCMP (see Figure 2). All measures were made using SigmaScan Software.

Sentences were digitized and filtered into 4 distinct spectral regions corresponding roughly to F1 (100-800 Hz), F2 (800-2200 Hz), and F3 (2200-6500 Hz) formant regions, as well as a wideband condition (100-6500 Hz). The rms output from the resulting filtered waveforms was computed for successive 33.33 ms time intervals, corresponding to the timing of each video frame. Figures 3 and 4 show the sentence waveform, and four envelope functions (wideband, F1, F2, and F3 envelope) for sentence S3 and S2 respectively.

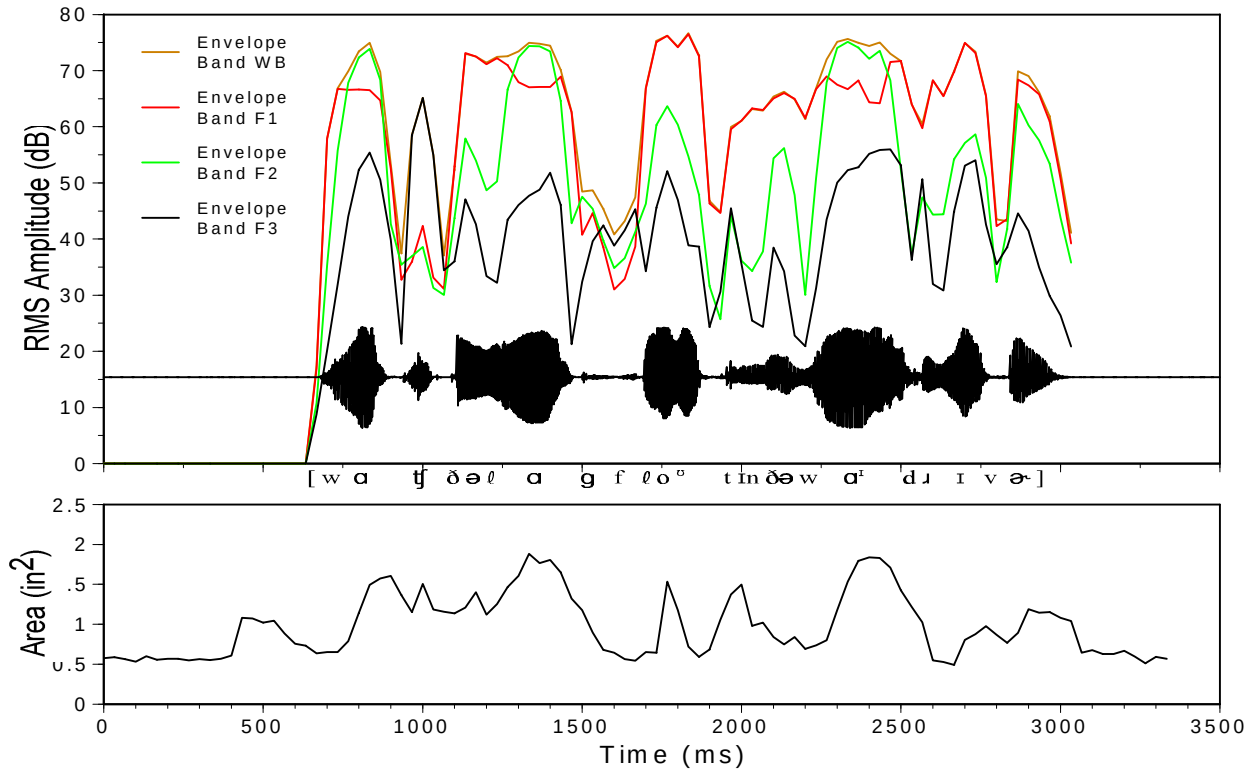

Figure 3. Amplitude-time waveform, RMS envelope functions, and area of mouth opening for sentence S3 “Watch the log float in the wide river”. WB Envelope Band (100-6500 Hz); F1 (envelope Band (100-800 Hz); F2 Envelope Band (800-2200 Hz); F3 Envelope Band (2200-6500 Hz.).

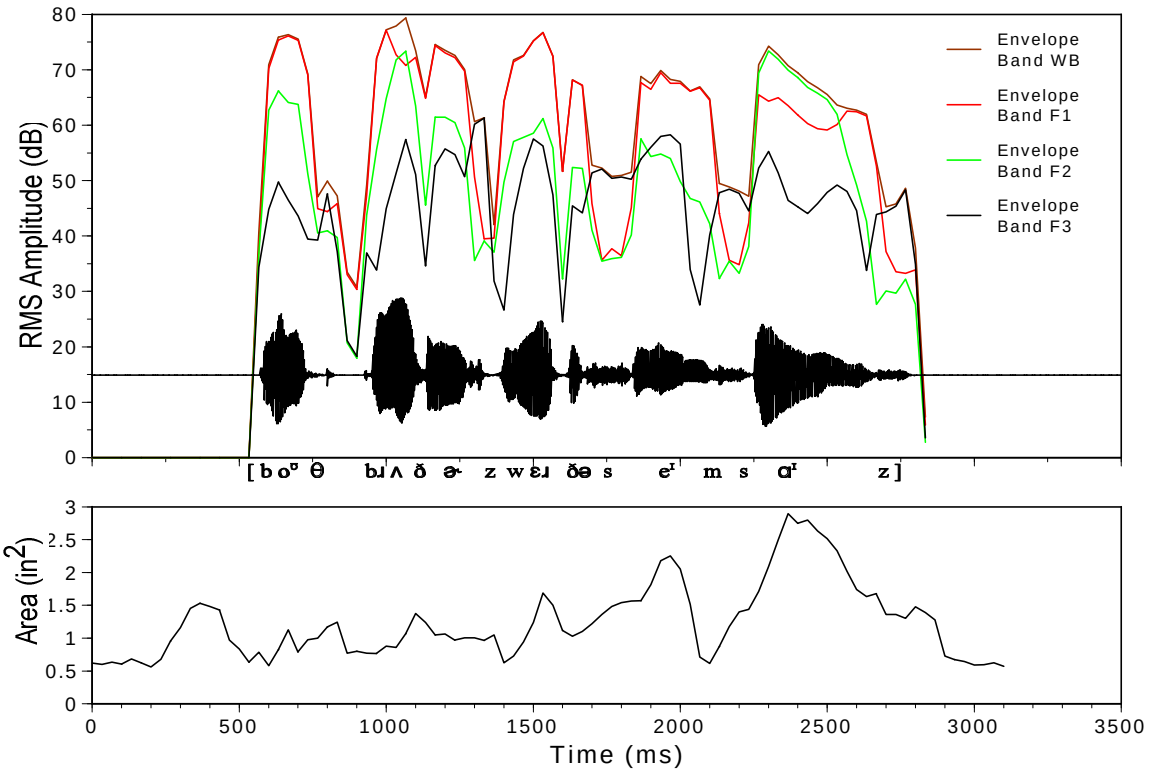

Figure 4. Same as Figure 3 except for sentence S2 “Both brothers wear the same size”.

RESULTS AND DISCUSSION

Given that visual movement of the speech articulators generally precedes acoustic output, we used cross-correlation plots between the area function and envelope functions for each sentence to help identify the lag (in frames) that resulted in the maximum correlation. Lags of up to ± 5 frames (≈165 ms) were deemed appropriate based on previous data on the effect of AV asynchrony on speech intelligibility[7] which show that, on average, subjects are able to tolerate about 150-200 ms of audio delay relative to the video signal without a detrimental effect on intelligibility. Inspection of the cross-correlation plots revealed that the maximum correlation was obtained with a lag of -1 frame, or about 33 ms.

Figure 5 and Table 1 show the correlations (with a -1 frame lag between V and A signals) within each sentence between area of mouth opening and acoustic envelope. As predicted, sentence S3 had higher correlations than S2. Furthermore, the band showing the highest correlation for both sentences was the F2 envelope band whereas the band showing the lowest correlation was the F1 envelope band. Across the two sentences, the correlations for S3 were significantly higher than those for S2 for each of the four envelope conditions.

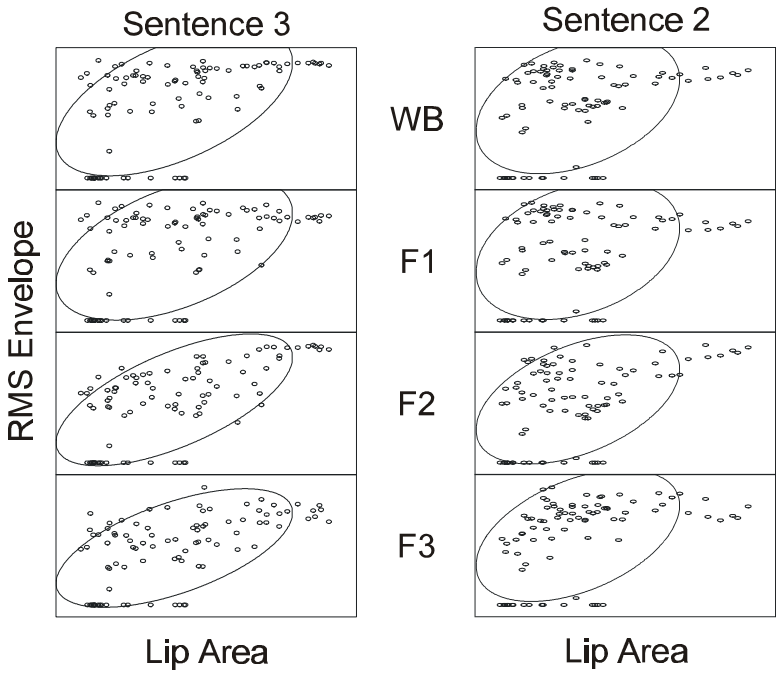

Figure 5. Correlation between area of mouth opening and RMS envelope functions for sentence S3 and S2. A 75% ellipse assuming a gaussian bivariate distribution is also shown.

Table 1. Correlation between area of mouth opening and envelope function. All correlations are significant ($p < 0.01$). Within each sentence, correlation with F2 envelope was best, whereas correlation with F1 envelope was worst. Differences across sentences for each envelope function are significant ($p < 0.05$).

	Sentence 3	Sentence 2
WB Envelope	0.52	0.35
F1 Envelope	0.49	0.32
F2 Envelope	0.65	0.47
F3 Envelope	0.60	0.45

To estimate the degree of correlation for the AV_{UM} condition, we correlated the area function for S3 with the envelope functions for S2 and *vice versa*. The correlations ranged from 0.32 to 0.41 which were significantly lower than the correlations for the AV_M condition with S3 for all four envelope functions and for S2 for the F2 envelope function.

As mentioned earlier, the detection experiment required that only a brief segment of the sentence be audible. Thus, high correlations across the entire sentence may not be crucial for detecting

these brief moments. Perhaps what is more important is the degree of local correlation at or near the specific points of detection for each sentence. In other words, if the correlation between area of mouth opening and envelope is high in the region of envelope maxima, we would predict a relatively large amount of BCMP. On the other hand, if the correlation between area of mouth opening is low, we would predict a relatively small amount of BCMP.

Local correlations between area of mouth opening and envelope were computed by selecting the first 10 frames from each function (333 ms) and then sliding the 10-point window by 1 frame. Plots of this moving 10-point correlation are shown in Figures 6 and 7 for sentence S3 and S2, respectively. Also shown in the figures are the wideband and F2 envelope function so that the points of correlation maxima can be compared to points of envelope maxima. For convenience, only the correlation with the F2 envelope function is displayed which was previously shown to have the highest correlation among the four envelope bands.

In Figure 6 (sentence S3), the local correlations are quite high ($r > 0.8$) in the region of the presumed point of detection ("log float"), as well as at several other envelope peaks. In comparison, the local correlations shown in Figure 7 (sentence S2) are substantially lower near the presumed point of detection ("Both brothers"). Note also, that even though the local correlation in Figure 7 reaches a high value near the end of the sentence, on the word "size", this high degree of comodulation between the area of mouth opening and envelope may be of little use because the overall energy at the end of the sentence is less than that at the beginning. Subjects may not have been able to take advantage of this correspondence because the S/N at the end of the sentence is too low to make this the point of detection.

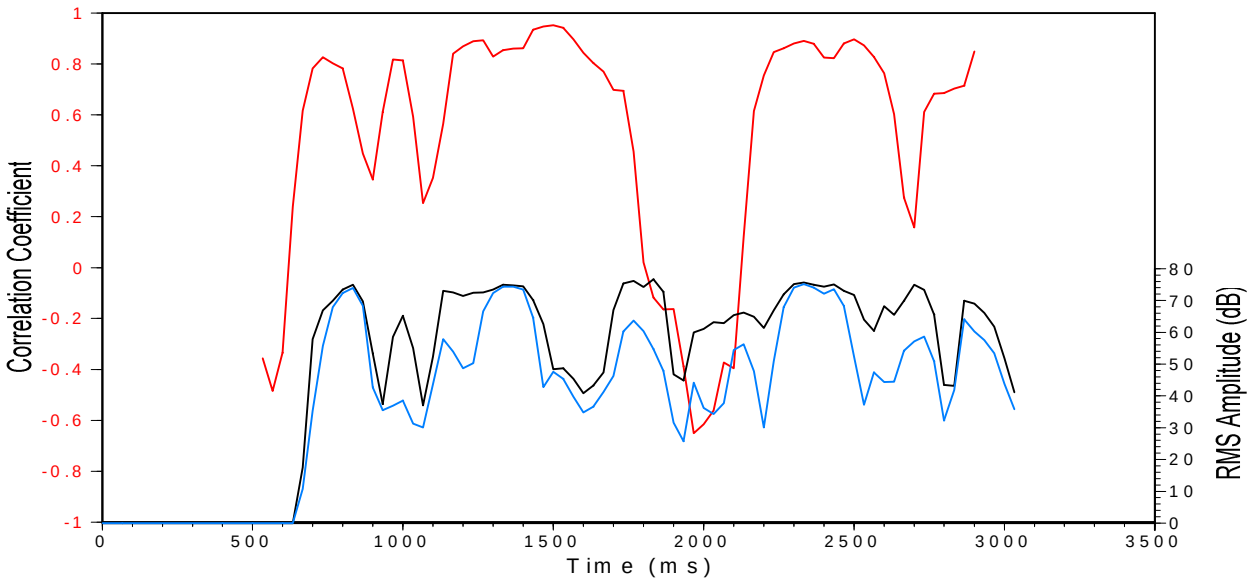

Figure 6. 10-point moving correlation between area of mouth opening and F2 envelope function for sentence S3. Red = correlation coefficient; Black = WB envelope; Blue = F2 envelope. Note the high correlation in the vicinity of the envelope peaks.

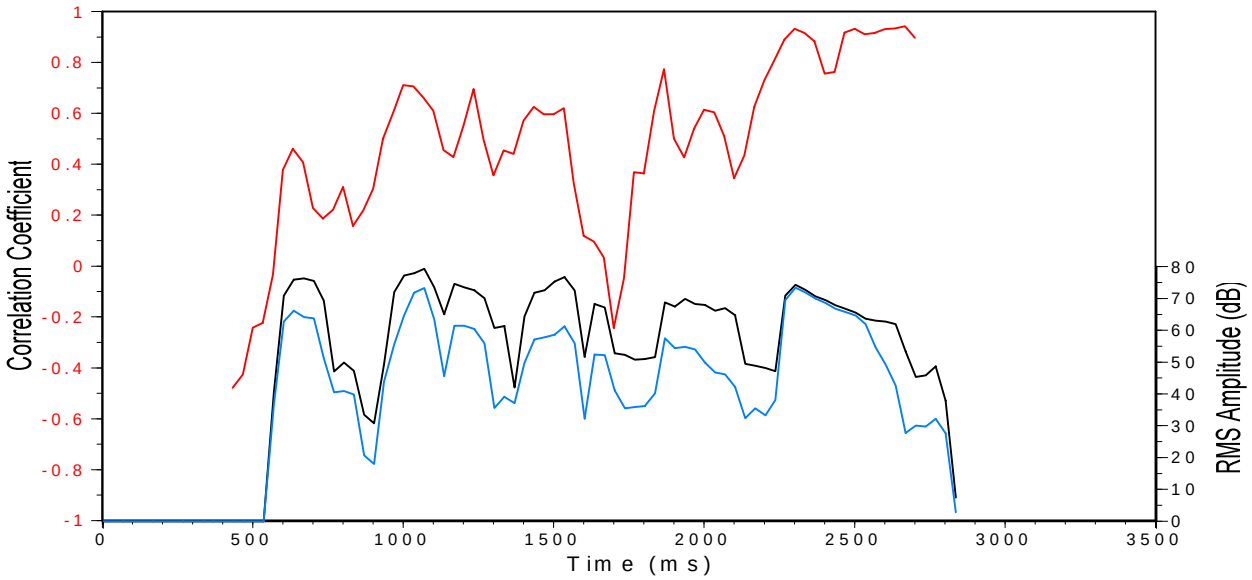

Figure 7. Same as in Figure 6 except for sentence S2. Note the relatively low correlation in the vicinity of the envelope peaks, especially at the beginning of the sentence.

CONCLUSIONS

In speech, the visible modulation of the area of mouth opening is correlated with the acoustic modulation of overall amplitude. As Summerfield stated [22], the "opening of the mouth is ... generally loosely, and occasionally tightly, related to the overall amplitude contour" (pg. 11). Our measures extend this notion by highlighting the relation between the area of mouth opening and the acoustic energy modulations in the F2 and F3 regions. This correspondence is exactly what one would predict given speechreaders' abilities to extract primarily place-of-articulation information, known to be carried by F2 and F3 transitions. Thus, we could say that F2 transitions, and to a lesser extent F3, are visible on the talker's lips.

This cross-modality comodulation is useful for informing the observer that the visual and acoustic information belong to the same articulatory event and should be processed together. By this account, we could say that visual information informs auditory processing of sound by directing attention and reducing temporal and spectral uncertainty.

However, our data also support the idea that visual signals that are comodulated with acoustic signals can have the effect of modulating the sensitivity of the auditory system. According to this view, listening under AV conditions is quite different than listening under A conditions. One can easily imagine that the operating state of the auditory system changes under visual influence. This is consistent with neurophysiological results showing that auditory cortex is greatly influenced by speechreading [2, 17], and with the existence of multisensory cells in the superior colliculus. [19] These cells tend to be maximally excited under conditions of weak unimodal input, a principle referred to as inverse effectiveness. Thus, under synchronized AV conditions with a weak A target signal as in Experiment 1, we can assume that the activity in these multisensory cells is relatively high leading to a different pattern of sensory activation than would be

experienced in the A alone condition. One result of this AV neurophysiological state might be that an observer's ability to extract an acoustic signal from a noise background is enhanced. This suggests that models of AV speech recognition that treat A and V input as essentially independent [1, 14] may not be correct. A model that combines the integration of unimodal information with bimodal sensory activation may better explain the great advantages readily observed in AV speech recognition with regard to intelligibility, speed of processing [18], and enhanced speech detection in noise.

ACKNOWLEDGMENTS

This work was supported by grant number DC00792 from the NIDCD and Walter Reed Army Medical Center. 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

1. Braida, L.D., *Quarterly J. Exp. Psych.* **43**, 647-677 (1991).
2. Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C., McGuire, P.K., Woodruff, P.W., Iversen, S.D., & David, A.S., *Science*, 276, 593-596 (1997).
3. Egan, J.P., Greenberg, G.Z., & Schulman, A.I., *J. Acoust. Soc. Am*. **33**, 771-778 (1961).
4. Egan, J.P., Schulman, A.I., & Greenberg, G.Z., *J. Acoust. Soc. Am*. **33**, 779-781 (1961).
5. Gordon, P.C. *J. Acoust. Soc. Am*. **102**, 2276-2283 (1997).
6. Grant, K.W., & Braida, L.D., *J. Acoust. Soc. Am.* **89**, 2952-2960 (1991).
7. Grant, K.W., & Seitz, P.F., *J. Acoust. Soc. Am.* (in press).
8. Grant, K.W., & Walden, B.E., *J. Acoust. Soc. Am.* **100**, 2415-2424 (1996a).
9. Grant, K.W., & Walden, B.E., *J. Speech Hear. Res.* **39**, 228-238 (1996b).
10. Grant, K.W., Walden, B.E., & Seitz, P.F., *J. Acoust. Soc. Am.* **103**, (1998 in press).
11. IEEE, *Institute of Electrical and Electronic Engineers*, New York (1969).
12. Hall, J.W. III, Haggard, M.P., & Fernandes, M.A., *J. Acoust. Soc. Am*. **76**, 50-56.
13. Levitt, H., *J. Acoust. Soc. Am.* **49**, 467-477 (1971).
14. Massaro, D.W., *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Hillsdale, NJ: Lawrence Earlbaum Assoc., 1987.
15. Miller, G.A., Heise, G.A., & Lichten, W., *J. Exp. Psych.* **41**, 329-335 (1951).
16. Repp, B.H., Frost, R., & Zsiga, E., *Quarterly J. Exp. Psych.* **45**, 1-20 (1992).
17. Sams, M., Aulanko, R., Hamalainen, M., Hari, R., Lounasmaa, O.V., Lu, S.T., & Simola, J., *Neuroscience Letters*, 127, 141-145 (1991).
18. Seitz, P.F., *J. Acoust. Soc. Am.* **101**, 3155.
19. Stein, B.E., & Meredith, M.A. *The Merging of the Senses*. Cambridge, MA: MIT Press, (1993).
20. Sumby, W.H. & Pollack, I., *J. Acoust. Soc. Am.* **26**, 212-215 (1954).
21. Summerfield, Q., *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **335**, 71-78 (1992).
22. Summerfield, Q., in B. Dodd and R. Campbell (Eds.) *Hearing by Eye: The Psychology of Lip-Reading.* Hillsdale NJ: Lawrence Erlbaum Associates, 1987, pp. 3-52.
23. Watson, C.S., & Nichols, T.L., *J. Acoust. Soc. Am*. **59**, 655-668 (1976).