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, DC 20307-5001

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. 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. Normal-hearing subjects were required to detect the presence of spoken sentences in noise under three conditions: auditory-only (A), auditory-visual with a visually matched sentence (AVM), and auditory-visual with a visually unmatched sentence (AVU).

When the video matched the target sentence, detection thresholds improved by about 1.6 dB relative to the auditory-only and auditory-visual unmatched conditions. The amount of threshold reduction varied significantly across target sentences, possibly reflecting the degree of visual and audio temporal and spectral comodulation.

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) by as much as 10 dB for spondaic words (14), and about 4-5 dB for more difficult connected speech materials such as the IEEE/Harvard sentence lists (5,9). Since each 1-dB improvement in S/N corresponds roughly to a 10 percent increase in intelligibility (5,12), it is fair to say that 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 (1,5,6,7,8). In this study, we focus on the potential importance of cross-modality temporal comodulation between variations in the acoustic (A) speech signal and the visible movements of the talker's lips (V) and how the temporal coherence between the two modalities may help protect the target speech signal from the effects of masking (4,15).

Repp et al. (13) 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 ability to accurately synchronize A and V stimulus components and in the selection of a speech modulated masker which itself was comodulated with the visible speech signals. 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.

METHODS

Masked thresholds for detecting speech were obtained from normal-hearing subjects under three conditions: auditory alone (A), audiovisual with matching visual stimulus (AVM), and audiovisual with mismatched visual stimulus (AVU). In the AVM condition, target audio sentences were presented along with simultaneous congruent visual lipread information. In the AVU condition, target audio sentences were presented along with simultaneous incongruent visual lipread information. Six sentences from the IEEE/Harvard sentence lists served as stimuli (9). Three sentences were used as auditory targets. The matching video from these sentences were used in the AVM conditions whereas the video from three different sentences were used in the AVU conditions. 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 noise was presented in one interval. In the other interval, only the noise was presented. Under AV conditions, video lipread information was available equally in both observation intervals. The subject's task was to identify the interval containing the sentence. The intensity of the white noise masker varied independently for each track according to a 3-down, 1-up adaptive tracking procedure using a 1-dB step size (10). The speech signal level was held constant at approximately 50 dB SPL. 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.

2335
RESULTS AND DISCUSSION

Using the detection thresholds obtained from the A condition as a reference, we computed the average masking difference for the AV\textsubscript{m} and AV\textsubscript{UM} conditions as well as the masking difference for each of the three target sentences (Figure 1). A significant masking level difference, or bimodal coherence masking protection (BCMP), was observed for all three sentences in the AV\textsubscript{m} condition (black bars). In contrast, there was no difference between the A and AV\textsubscript{UM} conditions (striped bars). The average BCMP obtained in the AV\textsubscript{m} condition was 1.6 dB. The range of BCMP for the three sentences was 0.9 to 2.2 dB. To assess the statistical significance of these effects a repeated measures analysis of variance (ANOVA) was conducted with condition and sentence as within-subject trial factors. The main effect of condition was highly significant [F(2, 10) = 34.9, p < 0.0001]. The main effect for sentence was also significant [F(2, 10) = 25.6, p = 0.0001] as was the interaction of condition and sentence [F(4, 20) = 3.9, p = 0.017]. Post hoc analyses confirmed that the A and AV\textsubscript{UM} conditions required a significantly greater speech-to-noise ratio than did the AV\textsubscript{m} condition and that the amount of MLD observed in the AV\textsubscript{m} condition was greater for sentence 3 than for sentence 2 (the difference in BCMP observed in the AV\textsubscript{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 (2) in what has been dubbed coherence masking protection (CMP). The magnitude of this protection (roughly 2 dB) is consistent with a reduction of for 6 subjects relative to the A alone condition. AV\textsubscript{m} (black bars) and AV\textsubscript{UM} (striped bars). Error bars show ±1 standard error.

FIGURE 1. Bimodal coherence masking protection in dB for 6 subjects relative to the A alone condition. AV\textsubscript{m} (black bars) and AV\textsubscript{UM} (striped bars). Error bars show ±1 standard error.

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


2336