PREDICTING AUDITORY-VISUAL SPEECH RECOGNITION IN HEARING-IMPAIRED LISTENERS

Ken W. Grant and Brian E. Walden (Army Audiology and Speech Center, Walter Reed Army Medical Center, Washington, DC 20307-5001)

ABSTRACT

Individuals typically derive substantial benefit to speech recognition from combining auditory (A) and visual (V) cues. However, there is considerable variability in AV speech recognition, even when individual differences in A and V performance are taken into account. In this paper, several possible sources of subject variability are examined, including segment perception, AV integration skill, and context usage. When these sources of variability are accounted for, predictions of AV speech recognition of nonsense syllables for normally-hearing and hearing-impaired listeners are excellent (R^2=0.96). Predictions for AV sentence recognition, however, are much poorer (R^2=0.44). These data will be discussed as part of a generalized model of AV speech recognition which includes the use of A and V unimodal cues, the integration of A and V cues, and the use of phonemic and semantic context.

[Work supported by NIH Grant DC 00792 and the Department of Clinical Investigation, Walter Reed Army Medical Center].

INTRODUCTION

In most communication settings, speech perception involves the integration of both auditory (A) and visual (V) information [1-4]. Further, auditory-visual (AV) speech perception is almost always better than either hearing or speechreading alone. This is especially true when the auditory signal has been degraded due to hearing loss or environmental noise.

Figure 1 shows fairly typical results obtained from intelligibility tests using low-context sentences [5] presented in a background of speech-shaped noise (S/N=0 dB) to hearing-impaired subjects. The hearing-impaired subjects tested had a variety of hearing-loss configurations ranging from mild to severe. For convenience, subjects are arranged along the abscissa in order of ascending A scores. As shown in the figure, all subjects demonstrated benefit from the addition of visual cues, but some subjects derived substantially more benefit than others, independent of their A score. For comparison, normally-hearing subjects tested under the same conditions achieved scores of 90%, 10%, and 98% for A, V, and AV conditions, respectively.

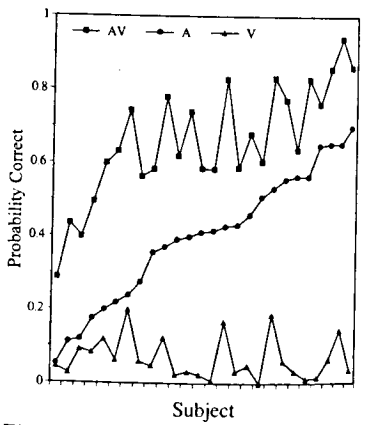

Figure 1. *A, V, and AV sentence recognition by individual hearing-impaired subjects.*

The exact amount of observed benefit depends on a number of variables. Among these are the individual's ability to recognize phonetic (e.g., consonants and vowels) and prosodic (e.g., intonation, duration, and stress) cues, the ability to integrate A and V cues, the difficulty of the speech materials, the physical conditions under which the speech is presented (e.g., noise, reverberation, lighting, viewing angle, etc.), and the individual's knowledge of the language and ability to make use of contextual constraints. Although much is known about the benefits of combining speechreading with audition, the relative effects of each of these factors on AV benefit is largely unknown. In this paper, we discuss some of the primary factors that are important to understanding how hearing and vision are combined in speech recognition.

SEGMENT RECOGNITION

Models of auditory-visual speech perception typically include auditory analyses, visual analyses, and more central processes common to both A and V modalities [6]. Since the classic study by Miller and Nicely [7], the recognition of speech segments (i.e., consonants and vowels) has typically been analyzed in terms of acoustic, phonetic, and articulatory features. Application of these analyses to AV recognition has shown that vision and hearing are often *complementary* in speech recognition under conditions of auditory signal degradation.

Figure 2 shows mean data for consonant feature recognition by normally-hearing subjects as a function of S/N for A, V, and AV conditions [8]. The top panel shows the data for voicing, whereas the middle and bottom panels show the data for manner of articulation and place of articulation, respectively. A and AV feature scores are shown by the dashed and solid lines. V feature scores are shown by the AV values displayed at -15 dB S/N. Notice that voicing cues obtained under AV conditions are determined by audition; that is, there is virtually no difference between AV and A conditions. In contrast, place-of-articulation cues are determined by speechreading. For this feature, AV scores remain virtually constant despite changes in A performance.

Although it may appear that both modalities contribute significantly to the recognition of manner-of-articulation cues, our analyses suggest that this feature is also determined by audition. It is important to remember that voicing, manner and place cues are not independent. Performance on one feature alters the expected chance performance on another. For example, if we assume that place-of-articulation cues are transmitted visually and that responses within place categories are distributed uniformly, then we would predict a visual manner score of 66% by chance alone. This is very close to the score for speechreading alone shown in the middle panel of the figure (depicted by the AV score at S/N = -15 dB). Thus, for these conditions, manner-of-articulation cues were derived primarily from the A condition.

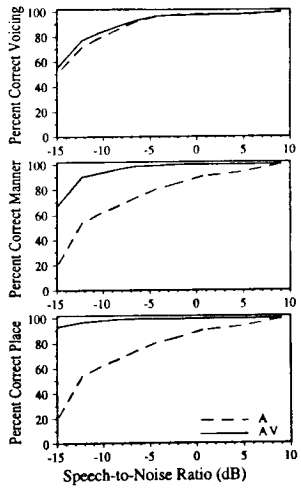

Figure 2. *A, V, and AV feature scores as a function of S/N.*

The complementary relation between auditory manner and voicing cues with visual place cues in speech recognition

has often been cited as a primary reason for the large advantages observed in AV consonant recognition relative to either A or V alone. In previous work, Walden, Prosek, & Worthington [9] developed a measure of AV redundancy that was able to account for a substantial amount of variability observed in AV consonant recognition by hearing-impaired listeners.

In a recent study, Grant and Walden [8] evaluated A, V, and AV consonant recognition by normally-hearing listeners under 12 different filtered-speech conditions. The filters were designed to create a range of A intelligibility scores with different patterns of perceptual confusions across A conditions. Confusion data obtained from each of the A filter conditions were subjected to a Sequential Information Feature Analysis [10], and the proportion of manner-plus-voicing information relative to the total amount of information received was calculated. This proportion represents, to a first approximation, the degree to which A and V information are complementary and shows the proportion of auditory information not obtainable by speechreading alone.

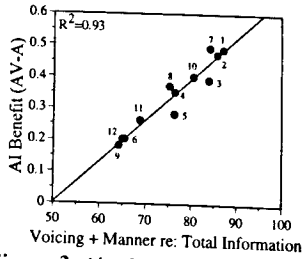

Figure 3. Absolute AI benefit as a function of AV complementation.

Figures 3 and 4 show the results of this analysis after converting percent correct scores to articulation index (AI) units. Figure 3 shows the results for *absolute* benefit (AV-A), while Figure 4 shows the results for *relative* benefit [(AV-A)/(1-A)]. In both figures, the abscissa shows the amount of information received in the A condition for the combined voicing-plus-manner feature expressed as a percent of the total amount of information received. There is a strong relation evident for both benefit measures, indicating that the degree to which A and V conditions complement each other is highly predictive of AV benefit. It should be noted that relating AV benefit to overall A intelligibility resulted in substantially weaker correlations.

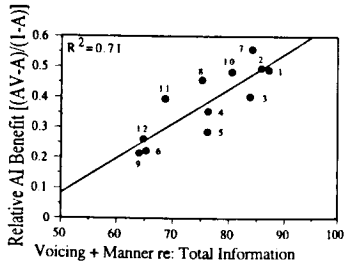

Figure 4. Relative AI benefit as a function of AV complementation.

Models of AV integration which make use of the entire A-alone and V-alone confusion matrices, such as the fuzzy logical model of perception (FLMP) proposed by Massaro [1] or the pre- and post-labelling models (PRE, POS) proposed by Braida [2], have been shown to predict AV consonant recognition more accurately than feature-based models when applied to data averaged across subjects.

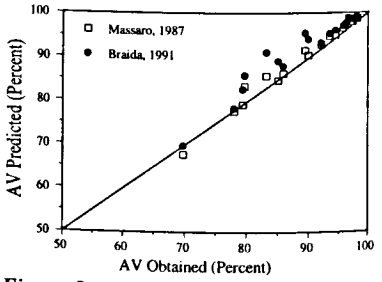

Figure 5. FLMP and PRE-model predictions for normal-hearing subjects.

Figure 5, for example, shows FLMP and PRE model predictions for the 8 S/N and 12 filtered-speech conditions shown in the previous figures. As can be seen, both models account for a large percentage of the variability in the obtained data (R^2 = 0.98 and 0.93 for the FLMP and PRE, respectively).

AUDIOVISUAL INTEGRATION

Whereas the ability to recognize segmental information from A and V speech signals is undoubtedly fundamental to predicting AV recognition, the ability to integrate A and V speech cues is another essential determinant of AV performance [11,1]. With the development of recent quantitative models of multisensory integration [1-2], it is now possible to estimate a listener's integration ability, independent from their ability to recognize A and V speech cues. These models predict AV recognition based on the pattern of segmental confusions obtained for each separate modality. It should be noted, however, that some AV cues, such as the relative timing of lip movements to voicing onset, are multimodal, in that they exist only as inter-modality timing cues. McGrath and Summerfield [12] have suggested that better lipreaders may be sensitive to these cues. Given the accuracy of the FLMP and PRE models, intermodal-timing cues may play only a small role in AV speech perception.

Unlike feature-based models, the FLMP and PRE integration models attempt to make use of all available data obtained in separate A and V identification tasks, and are potentially optimum-processor models. Ideally, a subject's AV performance should never exceed predicted performance. Subjects who perform as predicted are able to make use of all of the available information derived from the unimodal conditions. On the other hand, subjects who perform more poorly than predicted fail to make optimal use of A and V cues.

Our initial efforts to apply the FLMP and PRE models as a gauge of subject integration ability suggests that the PRE model may be more suitable. With the FLMP, stimuli identified correctly in one modality but incorrectly in the other, are predicted to be incorrect in the combined AV condition. As Braida [2] noted, the FLMP does not properly account for structured errors and relies too heavily on unimodal accuracy. In contrast, the PRE model focuses more on the consistency of unimodal responses and not necessarily on accuracy.

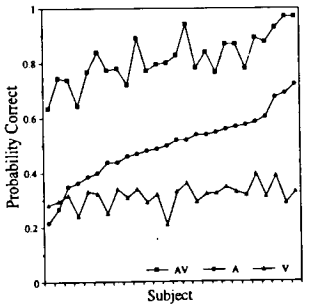

Figure 6. A, V, and AV consonant recognition by individual hearing-impaired subjects.

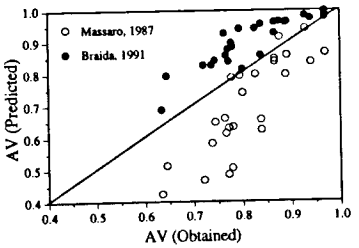

Figure 7. FLMP and PRE model predictions for hearing-impaired subjects.

Figures 6 and 7 show the results of a recent experiment examining consonant and sentence recognition in noise by 26 hearing-impaired subjects. Figure 6

shows A, V, and AV scores for each subject. As with Figure 1, subjects have been ordered along the abscissa according to A performance. Note first the large variability in AV recognition scores across subjects and the moderate correspondence between A and AV scores. This is additional evidence that the overall A score (or V score) does not allow accurate predictions of AV performance.

Predictions made by the FLMP and PRE models for the data shown in Figure 6 are displayed in Figure 7. The modeling results for individual subjects show that predictions of overall accuracy for are far less than perfect. FLMP predictions (R^2=0.71) consistently underpredicted performance when the input unimodal scores were low. Predictions by the PRE model on the other hand (R^2=0.77), were either equal to or greater than obtained performance. Thus, the PRE model, unlike the FLMP, behaves more like an optimal integrator. Some subjects are able to achieve this level of performance, whereas others fall short.

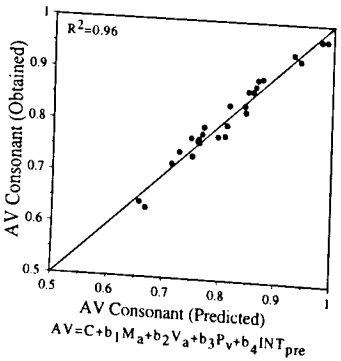

Figure 8. AV consonant predictions using a 4-factor feature-based model.

We have used this feature of the PRE model to estimate the extent to which individual subjects are able to integrate A and V cues optimally. Specifically, the difference between obtained and predicted AV performance was used as an index of AV integration skill. The combination of auditory voicing (V_A), auditory manner (M_A), visual place (P_V), and derived AV integration estimates (INT_{PRE} or INT_{FLMP}) was used in a 4-factor model to predict AV consonant recognition for hearing-impaired subjects. A and V feature scores were obtained from a SINFA analyses of the unimodal conditions and expressed as the percent of conditional information received for that feature relative to the total amount of all information received. For comparison, integration estimates were derived from AV predictions made by both PRE and FLMP models.

Figure 8 shows the results using the PRE model integration estimates. The excellent correlation obtained is impressive considering the simplicity of the model; that is, a linear addition of three unimodal feature scores and a measure of subject integration. Similar attempts to predict AV recognition without integration estimates, or with estimates derived from the FLMP model, led to significantly smaller correlations (R^2=0.822 and R^2=0.823, respectively).

WORD AND SENTENCE PROSODY

Even if AV segment recognition for individual subjects could be predicted perfectly, there would still be the problem of relating segment scores to word and connected-speech scores. One obvious difference between segmental recognition tasks and word and sentence recognition tasks is that the latter two contain important prosodic information related to stress, intonation, and rhythmic structure. The basic function of prosody in speech is to provide information about lexical, grammatical, and emotional aspects of the spoken message [13-14]. Further, individual differences in the ability to extract prosodic information appears to be an important factor in determining AV performance for words and sentences.

Acoustic analyses of prosody have shown that the cues for syllabification, stress, intonation, and phrasing include variations in fundamental frequency, segment and syllable duration, and amplitude envelope [15-20]. In general, speechreaders are not very good at extracting these cues. F0 variations are largely undetectable, and acoustic durational cues signifying segment lengths and intervocalic closure durations are often visually blurred or incompletely specified due to articulator movements that are either too rapid to follow visually or that occur behind the teeth [21-23]. Thus, as with voicing and manner-of-articulation cues, prosodic contrasts detected through audition are highly complementary to speechreading.

An important question related to the use of prosody in speech perception is whether subjects demonstrate variability in their judgements of prosodic contrasts. In a recent study, Grant and Walden [24] measured the ability of normally-hearing listeners to identify syllable number, syllabic stress, intonation, and rhythmic phrase structure in filtered words, phrases, and sentences. The filters used were approximately equal in intelligibility (AI=0.1), and spanned the frequency range from 300-5000 Hz. For some subjects, prosodic features were reliably extracted throughout the frequency spectrum. Other subjects, however, had considerable difficulty identifying sentence intonation and phrase structure from high-frequency speech regions.

Although Grant and Walden did not measure AV performance, the variability which they observed in subjects' abilities to extract suprasegmental cues from various parts of the frequency spectrum may be related to the variability observed in AV speech recognition. It is well known, for example, that there are substantial differences in AV performance among hearing-impaired observers who have the same average auditory recognition scores [25-26]. Given the highly complementary nature of acoustic prosody and speechreading cues, it is possible that some of the variability observed in AV word and sentence recognition tasks may be related to individual differences in the auditory recognition of prosodic and rhythmic cues.

WORD AND SENTENCE CONTEXT

Words and sentences provide listener's with many additional cues besides the usual segmental and suprasegmental cues. For example, identifying *nonsense* syllables requires that each separate consonant and vowel segment be received accurately. However, with *meaningful* words, lexical constraints make it possible to identify words correctly without having to resolve all of the individual segments. Similarly, words presented in isolation typically require more information than if the words were presented in sentences. In order to achieve the desired relationship between segment scores and word and sentence scores, these contextual variables need to be taken into account.

Following Boothroyd and Nittrouer [27], phonemic and semantic constraints can be represented quantitatively by using simple power-law equations. In Equation 1, the recognition of a CVC word is assumed to be equal to the recognition of its component parts. If each of these parts is statistically independent then,

$$P_w = P_p^n, \quad (1)$$

where P_w is the probability of recognition of the whole word, P_p is the probability of recognition of each independent segment, and n is the number of segments in the word. However, in real words, the segments are not independent and it is not required that all segments be received for the word to be recognized. Therefore, for real words,

$$P_w = P_p^j, \quad (2)$$

where $1 \leq j \leq n$. For monosyllabic words, j is approximately 2.5 [4,27].

Equation 3 relates words in isolation to words in sentences,

$$P_s = 1 - (1 - P_w)^k, \quad (3)$$

where P_s is the probability correct for words in sentences, P_w is the probability correct for words in isolation, and k is a free parameter (greater than one) reflecting the degree of predictability or context of the sentence materials. For low-context sentence materials such as the IEEE/Harvard set [5], k is approximately 1.14. For sentence sets with a higher degree of predictability (e.g., CUNY sentences), k is approximately 4.5 [4].

Combining equations 2 and 3 with appropriate estimates of j (≈ 2.5) and k (≈ 1.14), and substituting P_c for P_p gives

$$P_s \approx 1 - (1 - P_c^{2.5})^{1.14}, \quad (4)$$

where P_c is the proportion correct for consonants and P_s is the proportion of words correct for IEEE/Harvard sentences.

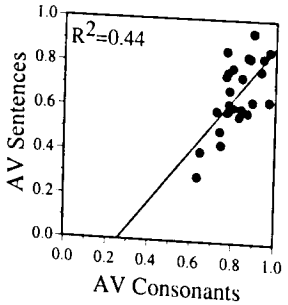

Figure 9. Relation between AV consonant recognition and sentence recognition.

Figure 9 shows the relation between AV consonant and AV sentence recognition for hearing-impaired subjects. Application of k- and j-factors appropriate for IEEE sentences (as described above) would adjust the range of consonant recognition scores to better match the range of observed sentence scores, but does nothing to reduce the variability across subjects. To accomplish this, individual differences in context-usage must be taken into account. Studies to estimate k- and j-factors for individual subjects, as opposed to sets of speech materials, are currently underway. Additionally, other measures of word and sentence context effects are being explored.

SUMMARY

Predictions of AV speech recognition ultimately depend on an understanding of how lexical access is affected by information provided by auditory and visual sources, the processes by which information is integrated, and the impact of top-down contextual constraints. Our efforts thus far to evaluate these factors in individual subjects have been limited mainly to consonant recognition, the recognition of certain prosodic contrasts, and segmental integration skills. Ongoing efforts to expand this work to include vowel recognition, sentence integration, and semantic context usage will no doubt improve our overall understanding of AV speech perception.

REFERENCES

[1] Massaro, D.M. (1987). *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Hillsdale, NJ: Lawrence Earlbaum Assoc.

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

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

[4] Rabinowitz, W.M., Eddington, D.K., Delhorne, L.A., & Cuneo, P.A. (1992). "Relations among different measures of speech reception in subjects using a cochlear implant," J. Acoust. Soc. Am. 92, 1869-1881.

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

[6] MacLeod, A. and Summerfield, Q. (1987). "Quantifying the contribution of vision to speech perception in noise," British J. Audiol. 21, 131-141.

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

[8] Grant, K.W., & Walden, B.E. (1993). "Evaluating the articulation index for auditory-visual consonant recognition," J. Acoust. Soc. Am. 94, 1887.

[9] Walden. B.E., Prosek, R.A., & Worthington, D.W. (1974). "Predicting audiovisual consonant recognition performance of hearing-impaired adults," J. Speech Hear. Res. 17, 270-278.

[10] Wang, M.D., Reed, C., and Bilger, R. (1978). "A comparison of the effects of filtering and sensorineural hearing loss on patterns of consonant confusions," J. Speech Hear. Res. 21, 5-36.

[11] Erber, N.P. (1975). "Auditory-visual perception of speech," J. Speech Hear. Res. 40, 481-492.

[12] McGrath, M. & Summerfield, Q. (1985). "Intermodal timing relations and audio-visual speech recognition by normal-hearing adults," J. Acoust. Soc. Am. 77, 678-685.

[13] Crystal, D. (1979). "Prosodic Development," In P. Fletcher & M. Garman (Eds.), *Language Acquisition* (pp. 33-48). Cambridge: Cambridge University Press.

[14] Kent, R., & Read, C. (1992). *The Acoustic Analysis of Speech*. San Diego: Singular Publishing Group.

[15] Lieberman, P. (1965). "On the acoustic basic of the perception of intonation by linguists," Word, 21, 40-54.

[16] Lehiste, I. (1976). Suprasegmental features of speech. In N.J. Lass (Ed.), *Contemporary Issues in Experimental Phonetics* (pp. 225-239). New York: Academic Press.

[17] Christie, W.M. (1974). "Some cues for syllable structure perception in English," J. Acoust. Soc. Am. 55, 819-821.

[18] Streeter, L. (1978). "Acoustic determinants of phrase boundary perception," J. Acoust. Soc. Am. 64, 1582-1592.

[19] Scott, D.R. (1982). "Duration as a cue to the perception of a phrase boundary," J. Acoust. Soc. Am. 71, 996-1007.

[20] Smith, M.R., Cutler, A., Butterfield, S., & Nimmo-Smith, I. (1989). "The perception of rhythm and word boundaries in noise-masked speech," J. Speech Hear. Res. 32, 912-920.

[21] Risberg, A. (1974). The importance of prosodic speech elements for the lipreader. In H.B. Nielson & B. Klamp (Eds.), *Visual and Audiovisual Perception of Speech VI. Danavox Symposium* (pp. 153-164). Scand. Audiol. (Suppl. 4).

[22] Risberg, A., & Lubker, J.L. (1978). "Prosody and speechreading," Speech Transmission Lab - Quarterly Progress Status Report, 4, 1-16.

[23] Boothroyd, A. (1988). "Perception of speech pattern contrasts from auditory presentation of voice fundamental frequency," Ear and Hearing, 9, 313-321.

[24] Grant, K.W. and Walden, B.E. (1992). "The transmission of prosodic information via selected spectral regions of speech," J. Acoust. Soc. Am. 92, 2300-2301.

[25] Montgomery, A.A., Walden, B.E., Schwartz, D.M., & Prosek, R.A. (1984). "Training auditory-visual speech reception in adults with moderate sensorineural hearing loss," Ear and Hearing, 5, 30-36.

[26] Walden, B.E., Busacco, D.A., & Montgomery, A.A. (1993). "Benefit from visual cues in auditory-visual speech recognition by middle-aged and elderly persons," J. Speech Hear. Res. 36, 431-436.

[27] Boothroyd, A. and Nittrouer, S. (1988). "Mathematical treatment of context effects in phoneme and word recognition," J. Acoust. Soc. Am. 84, 101-114.