

Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration

Ken W. Grant, Brian E. Walden, and Philip F. Seitz

Walter Reed Army Medical Center, Army Audiology and Speech Center, Washington, DC 20307-5001

(Received 19 March 1997; revised 10 December 1997; accepted 24 January 1998)

Factors leading to variability in auditory-visual (AV) speech recognition include the subject's ability to extract auditory (A) and visual (V) signal-related cues, the integration of A and V cues, and the use of phonological, syntactic, and semantic context. In this study, measures of A, V, and AV recognition of medial consonants in isolated nonsense syllables and of words in sentences were obtained in a group of 29 hearing-impaired subjects. The test materials were presented in a background of speech-shaped noise at 0-dB signal-to-noise ratio. Most subjects achieved substantial AV benefit for both sets of materials relative to A-alone recognition performance. However, there was considerable variability in AV speech recognition both in terms of the overall recognition score achieved and in the amount of audiovisual gain. To account for this variability, consonant confusions were analyzed in terms of phonetic features to determine the degree of redundancy between A and V sources of information. In addition, a measure of integration ability was derived for each subject using recently developed models of AV integration. The results indicated that (1) AV feature reception was determined primarily by visual place cues and auditory voicing+manner cues, (2) the ability to integrate A and V consonant cues varied significantly across subjects, with better integrators achieving more AV benefit, and (3) significant intra-modality correlations were found between consonant measures and sentence measures, with AV consonant scores accounting for approximately 54% of the variability observed for AV sentence recognition. Integration modeling results suggested that speechreading and AV integration training could be useful for some individuals, potentially providing as much as 26% improvement in AV consonant recognition. [S0001-4966(98)02305-4]

PACS numbers: 43.71.An, 43.71.Es, 43.71.Kg, 43.71.Ma [WS]

INTRODUCTION

For all but the most profoundly hearing-impaired individuals, auditory-visual (AV) speech recognition has consistently been shown to be more accurate than auditory-only (A) or visual-only (V) speech recognition. Although this is especially true when the auditory signal is distorted (e.g., due to hearing loss, environmental noise, or reverberation), the influence of visual cues on speech recognition is not limited to conditions of auditory distortion. Even with fully intact speech signals, visual cues can have an impact on recognition. McGurk and MacDonald (1976) demonstrated that when the auditory production of one consonant is synchronized with the visual production of another consonant, most observers will perceive a third consonant that is not represented by either auditory or visual modality. For example, when auditory /ba/ is presented with visual /ga/, the perceived result is often /da/. This illusion, known as the "McGurk Effect," occurs even when the auditory signal is perfectly intelligible.

Another example of the influence of visual speech cues on intact auditory signals occurs when listeners are asked to repeat unfamiliar phrases, as when learning a second language or when presented with grammatically complex passages. For instance, when asked to shadow the speech of a native French speaker, students with four years of French training performed significantly better when they were able

to see the speaker's face as compared to when they could only hear his voice (Reisberg *et al.*, 1987). Similarly, when native speakers of English were asked to shadow passages from Kant's *Critique of Pure Reason* spoken by another native English speaker, performance was significantly better when visual cues were available (Reisberg *et al.*, 1987).

These brief examples suggest that the perception of speech is inherently multimodal. Cues extracted from both auditory and visual sources are integrated early in the perceptual analysis of speech, and this information is used without much regard to the sensory modality of origin (Massaro, 1987; Summerfield, 1987). Yet, in spite of the superior status of auditory-visual speech recognition relative to auditory-only speech recognition, recent explications of speech perception often omit, or only mention briefly, the role of visual speech information (Diehl and Kluender, 1989; Klatt, 1989; Stevens, 1989; Halle and Stevens, 1991; Greenberg, 1995; Lindblom, 1996; Ohala, 1996). In contrast, recent studies of automatic speech recognition have begun to emphasize the importance of incorporating visual speech information along with acoustic information to improve recognition performance, especially in noisy environments (Stork and Hennecke, 1996).

From an applied perspective, a theory of AV speech perception would be an extremely valuable tool for addressing the communication problems encountered by hearing-impaired individuals or by normally hearing individuals in

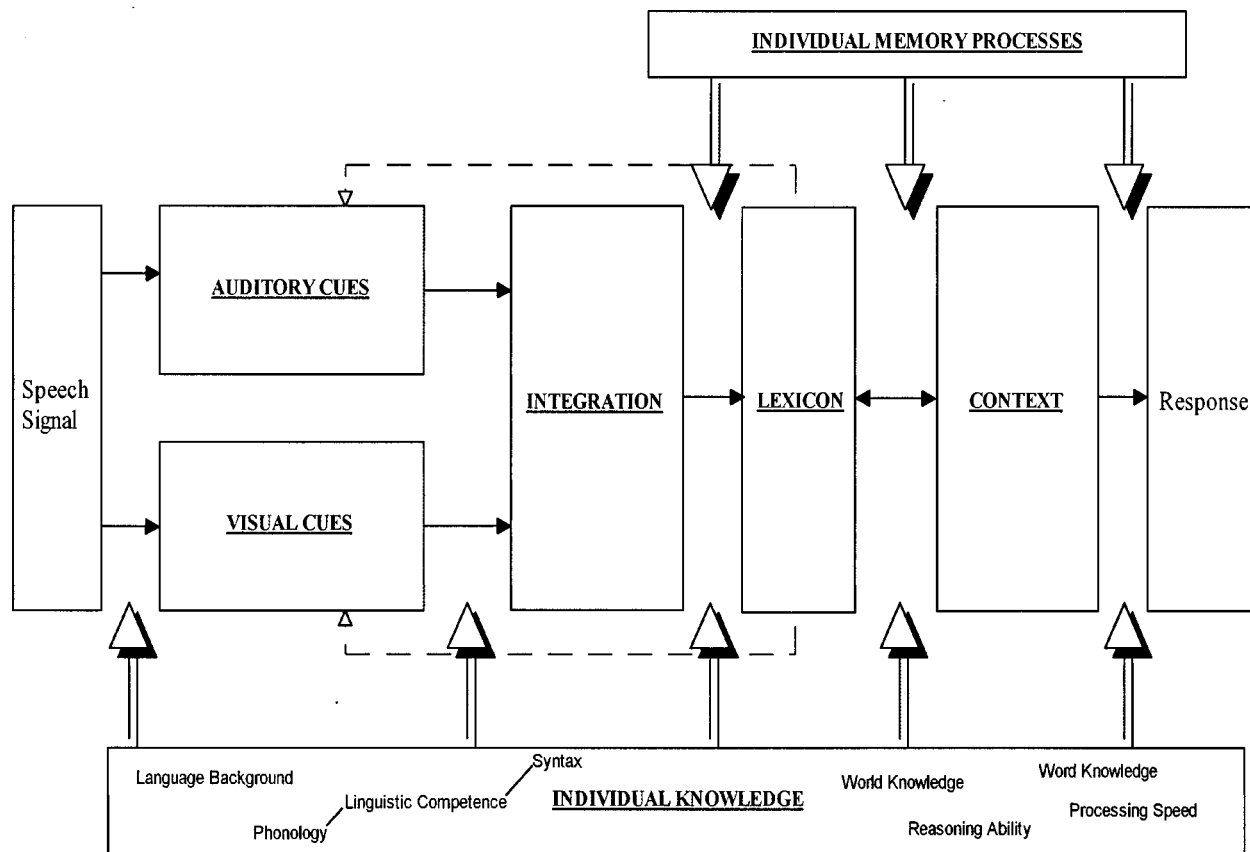


FIG. 1. Schematized framework of auditory-visual speech recognition.

noisy or reverberant environments. Such a theory would provide a conceptual framework that could serve as a guide for developing signal-processing strategies and/or rehabilitation programs when AV speech perception is less than perfect. Consider the simple conceptual framework shown in Fig. 1. A speech signal composed of both optical and acoustic information is presented. The listener-observer extracts signal-related segmental and suprasegmental cues from each modality, integrates these cues, and applies top-down semantic and syntactic constraints in an effort to interpret the message before making a response. The basic components, bottom-up signal-related cue extraction, integration, and top-down linguistic processes, are common to most speech perception theories (e.g., Liberman *et al.*, 1967; Stevens and House, 1972; Studdert-Kennedy, 1974). The major distinction drawn here from auditory-only theories of speech perception is that in an audiovisual communication environment, cues from the visual modality must be considered, and the integration of A and V cues, both within and across modalities, must be described (Massaro, 1987).

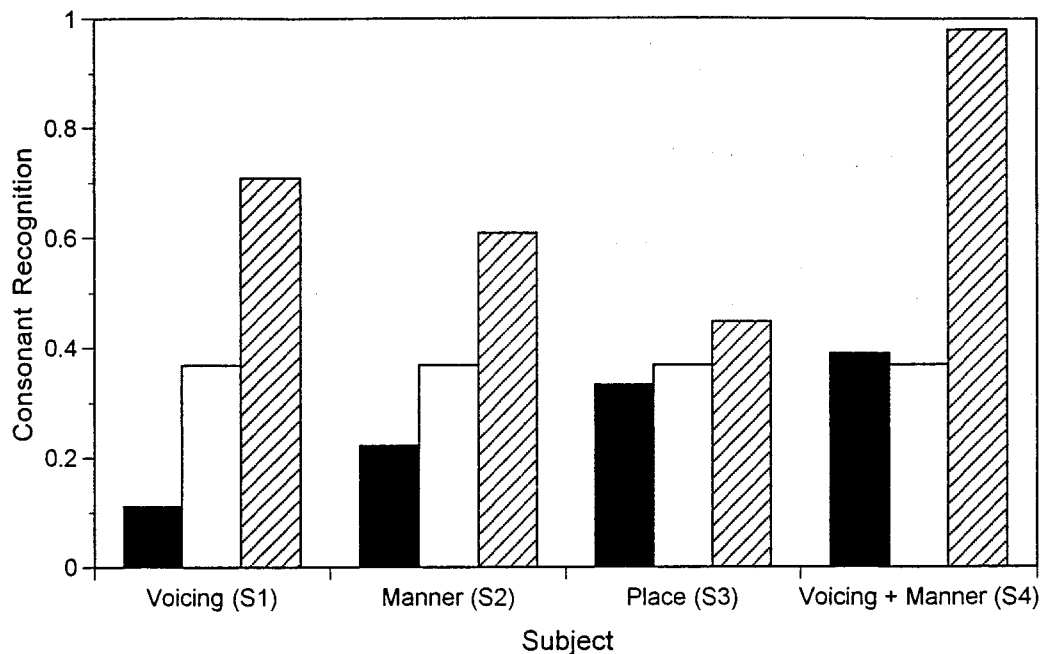
From this perspective, consider a hearing-impaired individual whose AV recognition of words and sentences is less than perfect. In order to evaluate the exact nature of the communication problem, it is necessary to determine whether the communication failure is due to poor reception of auditory and/or visual cues, difficulty in integrating A and V cues, difficulty in applying linguistic and contextual constraints, reduced working memory capacity, or a combination of these factors. If the problem is determined to be primarily

difficulty in receiving A or V cues, signal-processing strategies to enhance the relevant cues may be used. If, on the other hand, the problem is determined to be difficulty in integrating A and V cues or difficulty in applying top-down language processing rules, then training and practice techniques may be applied. Simply knowing the individual's AV sentence or word recognition performance is not sufficient for determining a plan for rehabilitation.

In order to use the framework displayed in Fig. 1 as a guide for rehabilitation, three questions must be addressed: (1) What are the most important cues for AV speech recognition that can be extracted from acoustic and visual speech signals? (2) Is it possible to measure an individual's ability to integrate auditory and visual cues separate and apart from their ability to recognize syllables, words, and sentences? (3) What are the most important nonsignal-related "top-down" processes that contribute to individual variability in AV speech recognition? A brief discussion relating to each of these questions is presented below.

1. What are the most important cues for AV speech recognition that can be extracted from acoustic and visual speech signals?

Clearly, the answer to this question cannot come from studies of auditory speech perception alone. For example, auditory place-of-articulation cues, which are extremely important for A-only speech recognition, are not as important in AV speech recognition because this information is readily



- A Auditory consonant recognition based on perfect transmission of indicated feature. Responses within each feature category were uniformly distributed (see Table 1).
- V Visual Consonant recognition. Data from 11 normally-hearing subjects (Grant and Walden, 1996a).
- ▨ PRE Predicted AV consonant recognition based on PRE model of integration (Braida, 1991).

FIG. 2. Scores for A, V, and (predicted) AV conditions for four hypothetical subjects with average speechreading ability. Auditory-visual predictions were made with the PRE model using an average normal-hearing V consonant confusion matrix and an appropriate A confusion matrix with perfect recognition of the specified speech feature. Each set of bars represents a subject with perfect auditory feature recognition for voicing (S1), manner-of-articulation (S2), place-of-articulation (S3), and voicing-plus-manner (S4), respectively.

available through the visual channel. On the other hand, auditory voicing and manner-of-articulation cues are extremely important in AV speech recognition because this information is difficult to obtain visually. A careful accounting of the cues extracted from the separate unimodal conditions can reveal the degree of redundancy or complementarity between A and V cues. If A and V cues are redundant, there will be less benefit resulting from their combination than if the cues are complementary (Walden *et al.*, 1977; Grant and Walden, 1996a).

The relevancy of cue redundancy between A and V conditions as a predictor of AV performance is important for the present discussion because it forces us to recognize that when comparing two different listening conditions (or two listeners with different degrees of hearing loss, or two different hearing aids), the condition resulting in the higher rate of auditory intelligibility need not result in the highest AV score or the greater improvement when combined with speechreading. This point is illustrated graphically in Fig. 2.

Figure 2 shows predicted AV consonant recognition results for four different hypothetical subjects. For each hypothetical subject, three bars are shown. The black bar represents the auditory recognition score on a consonant identification task that would result from perfect recognition

of either voicing (S1), manner-of-articulation (S2), place-of-articulation (S3), or the combination feature, voicing-plus-manner (S4). Table I shows the feature classification for the set of consonants used in this example. The auditory recognition scores (black bars) were calculated assuming perfect recognition of each specified feature and a uniform distribution of responses within each feature category. For example, for the hypothetical subject who receives perfect voicing information (S1), the confusion matrix shows a uniform distribution of responses among all voiced consonants and a uniform distribution of responses among unvoiced consonants. Since the subject is perfect in recognizing the voicing feature, there are no cases of voiced-voiceless confusions. Once constructed in this manner, the confusion matrix can then be scored for overall accuracy by simply observing the number of correct responses (lying along the main diagonal of the confusion matrix) divided by the total number of trials.

The unfilled bars are the same for each subject and represent consonant recognition performance of a typical speechreader (Grant and Walden, 1996a).¹ The shaded bars represent AV predictions made by Braida's Pre-Labeling Integration Model (1991).² Note that a subject who has perfect voicing recognition (S1) derives substantially more AV benefit (e.g., difference between AV and A recognition) accord-

TABLE I. Feature classification for voicing, manner, and place categories.

VOICING					
Voiced b,d,g,m,n,v,ð,z,ʒ,dʒ			Unvoiced p,t,k,f,θ,s,ʃ,tʃ		
MANNER OF ARTICULATION					
Stop b,p,g,k,d,t		Nasal m,n	Fricative v,f,ð,θ,z,s,ʒ,ʃ		Affricate dʒ,tʃ
PLACE OF ARTICULATION					
Bilabial b,p,m	Lingua-Velar g,k	Lingua-Alveolar d,t,n,s,z	Lingua-Dental ð,θ	Lingua-Palatal ʒ,ʃ,dʒ,tʃ	Labio-Dental v,f

ing to Braida's model than subjects who obtain either perfect manner (S2) or-perfect place recognition (S3), even though these latter subjects have higher overall A recognition scores. Finally, a subject who obtains perfect voicing and manner recognition (S4) is predicted to have a nearly perfect AV score. This implies that hearing-impaired listeners with average speechreading skills who can extract voicing and manner information from the A condition may be expected to have high AV consonant scores, regardless of their overall A recognition score. Figure 2 also demonstrates that improvements in absolute unimodal recognition scores that may result from various forms of signal processing (e.g., hearing aids) may not produce the best bimodal recognition score.

2. Is it possible to measure an individual's ability to integrate auditory and visual cues separate and apart from their ability to recognize syllables, words, and sentences?

The recent development of quantitative models of multisensory integration (Massaro, 1987; Blamey *et al.*, 1989; Braida, 1991) has made it possible to derive measures of AV integration that are independent of the individual's ability to extract cues from the A and V modalities. These models make predictions of AV performance based on speech feature recognition or from the detailed patterns of confusions obtained for each separate modality. If these predictions are based on optimum processing models, then the extent to which the actual obtained AV performance approximates the predicted performance may be used as an index of AV integration skill. That is, subjects whose AV performance is well-predicted by the models are better at integrating A and V cues than subjects whose actual AV performance is over-predicted by the models.

Some of the integration models considered in this study are not optimum processor models, and observed AV scores can be higher than predicted scores. Nevertheless, deviations between predicted and obtained AV performance may still be used as an index of integration skill. For example, suppose that predictions from a particular model were based only on a subject's overall score from each separate modality and did not take into account the pattern of confusions generated in the A and V conditions. Even though such a model would likely underpredict real performance, the extent to which subjects' AV recognition scores exceed model predictions might indicate better integration abilities.

3. What are the most important nonsignal-related "top-down" processes that contribute to individual variability in AV speech recognition?

Individuals interpret speech signals in conjunction with stored linguistic knowledge (Lindblom, 1996). This knowledge base includes the individual's vocabulary and other properties of the lexicon, their knowledge of semantics and syntax, their use of word-level and sentence-level context to compensate for misperceptions and impoverished acoustic and visual information, and memory processes and strategies for lexical access based on partial information in the signal. Measures of variability for each of these sources of linguistic knowledge are likely to reveal interesting subject differences which may impact on AV recognition of speech. Unfortunately, this potentially important source of variability in AV performance continues to receive little attention in the literature.

The experiments reported here represent the first in a series of studies to evaluate the importance of these different factors (cue extraction, integration, top-down processing) in determining AV speech recognition. In the present study, we focus mainly on the first two factors and describe individual variability in AV consonant and sentence recognition by hearing-impaired persons in terms of two factors: the ability to recognize consonants auditorily and visually, and the ability to integrate A and V cues. Admittedly, consonant recognition represents only one class of several potentially important "bottom-up signal-related" factors (as compared to the recognition of consonant clusters, vowels, word and sentence stress, syllabification, etc.). One reason for choosing to investigate consonant recognition is that in English, consonants account for a large proportion of the total segmental information making up lexical items (Kucera and Francis, 1967). Furthermore, many of the acoustic cues that help distinguish one consonant from another are often of low intensity and involve rapid, short-duration spectral transitions making them difficult to hear under conditions of hearing-impairment or environmental distortion (e.g., noise or reverberation). These two attributes of consonants (frequency of occurrence in the language and acoustic characteristics) lead us to hypothesize that variability in consonant recognition will contribute strongly to variability in word and sentence recognition.

This study is divided into three main parts: obtaining measures of auditory (A), visual (V), and auditory-visual

(AV) consonant recognition, deriving a measure of auditory-visual integration ability from A and V consonant recognition data, and obtaining measures of A, V, and AV sentence recognition. All measures were obtained on individual hearing-impaired patients. The overall objective was to relate the variability in consonant recognition and AV integration ability to variability in AV sentence recognition.

I. METHODS

A. Subjects

Twenty-nine subjects between the ages of 41 and 88 years (mean age=65.0) were recruited from the patient population of the Army Audiology and Speech Center, Walter Reed Army Medical Center. All had acquired sensorineural hearing losses due primarily to noise exposure. The average three-frequency (500, 1000, and 2000 Hz) pure-tone threshold for the better ear was 33 dB HL (range: 0–63.3 dB HL re; ANSI, 1989). The average two-frequency (2000 and 4000 Hz) pure-tone threshold was 53.5 dB HL (range: 20–77.5 dB HL). No additional audiometric criteria were imposed in subject selection. Because variability in the patterns of auditory consonant confusions was important to the planned analyses, subjects with a variety of configurations and severity of hearing loss were included. In order to include a wide range of speech recognition abilities, potential subjects were initially screened on their ability to recognize IEEE Harvard sentences (IEEE, 1969) in noise [signal-to-noise ratio (S/N) of 0 dB]. Screening of subjects continued until 4–5 subjects with scores at each performance level of approximately 10%, 20%, 30%, 40%, 50%, and 60% correct were identified (i.e., 5 subjects at 10%, 5 subjects at 20%, etc.). All subjects were native speakers of American English and had normal or corrected-to-normal vision. None had prior experience as subjects of experiments involving visual or bisensory speech recognition. Although many of the subjects were experienced hearing-aid users, all testing was conducted binaurally under headphones with speech levels approximating 85 dB SPL. When eligible subjects were paid for their participation in the screening process, and if selected, during the main experiment.

B. Procedure

1. Consonant recognition

Consonant recognition was measured separately for A, V, and AV presentations. Speech materials consisted of eighteen medial consonants (/p,b,t,d,k,g,m,n,s,z,f,v,θ,ð,ʃ,ʒ,tʃ,dʒ/) surrounded by the vowel /a/. Ten productions of each /a/-consonant-/a/ (aCa) stimulus were spoken by a female talker of American English and recorded on optical disk (Panasonic TQ-3031F). The audio portion of each production was digitized (16-bit A/D, 20-kHz sampling rate), normalized in level, and stored on computer (Vanguard 486). For auditory-visual presentations, the digitized computer audio and optical disk video portions of each production were realigned using custom auditory-visual control software. Alignments were checked using a dual trace oscilloscope to compare the original and digitized productions of each utterance and were found to be within ± 2 ms.

Prior to testing, subjects were familiarized with the consonant set and the use of a touch screen in making responses. Care was taken to ensure that each subject understood the task and could make use of the eighteen consonant category labels. Auditory-visual practice blocks of 36 trials, each with trial-by-trial feedback, were presented binaurally through headphones (Beyer DT-770) at a comfortable listening level appropriate for each subject. Speech materials were presented in quiet with subjects seated in a sound-treated room facing a 19-in. color video monitor (SONY PVM 2030) situated approximately 5 ft from the subject. After one or two initial practice blocks, subjects were required to achieve performance levels on three additional consecutive practice blocks of 92% correct consonant recognition or better. Only one subject (out of 30 tested) failed to meet this requirement and was eliminated from further testing.

Following familiarization training, data were obtained on consonant recognition in noise. On each test trial, a single aCa production was presented. The overall level of the speech signal was approximately 85 dB SPL. The aCa utterances were mixed with a speech-shaped noise matched to the average long-term spectrum of the stimulus set and presented at a S/N ratio of 0 dB. This level of noise was chosen to insure a sufficient number of recognition errors for the planned analyses. A different sample of noise was used for each of the 180 tokens (18 consonants \times 10 productions). Each speech token began 50 ms after the noise started and ended 50 ms before the noise finished. Subjects were required to identify the consonant presented, selecting their response from a touch screen terminal displaying all 18 possible consonant labels. Subjects were tested in blocks of 72 trials. Ten blocks each were presented in the A, V, and AV conditions, yielding a total of 40 trials per consonant per condition. The order of A, V, and AV conditions was randomized for each subject. No feedback was provided.

2. Auditory-visual integration

Three recent models of auditory-visual speech recognition (Massaro, 1987; Blamey *et al.*, 1989; Braida, 1991) were used for estimating the efficiency with which individual subjects integrate A and V cues. Each of these models predicts AV recognition scores from A and V scores alone. Differences between predicted and obtained AV scores were used as a measure of each subjects' integration efficiency. Below, a brief description of each model is presented along with the methods used to derive individual integration measures.

a. Braida (1991)—PRE. In the Prelabeling Model of Integration, confusion matrices from A and V consonant recognition are subjected to a special form of multidimensional scaling (MDS) and interpreted within a Theory of Signal Detection (e.g., Green and Swets, 1966; Macmillan *et al.*, 1988). The model provides a spatial interpretation of the ability to distinguish between consonants analogous to that derived from traditional multidimensional scaling (Borg and Lingoes, 1987). However, unlike traditional MDS, the scaled distances between consonants in the separate A and V spaces are converted to a common metric, d' , explicitly reflecting the correctness of responses. The decision process assumes a

comparison between stimulus attributes (modeled as a multidimensional vector of cues, \vec{X}) and prototypes or response centers (\vec{R}) in memory. Subjects are assumed to respond R_k if and only if the distance from the observed vector of cues \vec{X} to \vec{R}_k is smaller than the distance to any other prototype. A subject's sensitivity $d'(i, j)$ in distinguishing stimulus S_i from stimulus S_j is given by

$$d'(i, j) = \|\vec{S}_i - \vec{S}_j\| = \sqrt{\sum_{k=1}^D (S_{ik} - S_{jk})^2}, \quad (1)$$

where $\|\vec{S}_i - \vec{S}_j\|$ is the distance between the D -dimensional vector of cues generated by stimuli S_i and S_j .

In the present study, estimates of stimulus and response centers that best fit a given confusion matrix were obtained iteratively using a KYST procedure (Kruskal and Wish, 1978). For the first iteration, \vec{S} and \vec{R} are assumed to be aligned. Subsequent iterations attempted to improve the match between predicted and obtained matrices (using a x^2 -like measure) by displacing slightly both stimulus and response centers. Each iteration assumed 5120 presentations per consonant token yielding a total of 92 160 trials per matrix. This number was selected to reduce the stimulus variability in each simulation to approximately 1/10th of the variability in the data. The MDS fits were further optimized by choosing either two- or three-dimensional solutions depending on which gave the best fit to the unimodal matrix.

Prelabeling model predictions for AV performance are made solely on the basis of unimodal performance. Assuming that A and V cues are combined optimally, the decision space for the AV condition is the Cartesian product of the space for the A condition and the space for the V condition. Thus the relation between a subject's sensitivity in the AV condition and the corresponding unimodal sensitivities, assuming no perceptual interference (e.g., masking or distraction) across modalities, is given by

$$d'_{AV}(i, j) = \sqrt{d'_A(i, j)^2 + d'_V(i, j)^2}. \quad (2)$$

Predictions of AV consonant recognition were made using the A and V consonant recognition data described above and compared to AV consonant recognition obtained by individual subjects. Since Braida's Prelabeling Model is an optimum processor model, predicted AV scores will always equal or exceed observed AV scores. A subject's integration efficiency, as predicted by the model, was given by the difference between predicted and observed AV recognition scores (with zero difference indicating perfect integration).

b. Massaro (1987)—FLMP. In the Fuzzy Logical Model of Perception (FLMP), auditory and visual channels are considered to be independent sources of information about the identity of the AV stimulus. Continuously valued features are evaluated, integrated, and matched against prototype descriptions in memory, and an identification decision is made on the basis of the relative goodness of match of the stimulus information with the relevant prototype descriptions. In the multimodal case, each speech segment possesses a set of feature values for the auditory channel and a set of feature values for the visual channel. For example, in the syllable /ba/, the lips coming together represent a visual fea-

ture, whereas second and third formant transitions appropriate for /ba/ represent acoustic features. To relate these sets of informational values to each other, a common metric based on fuzzy truth values (Zadeh, 1965) is used. Truth values ranging between 0 and 1 are used to represent the goodness-of-fit between each particular feature in the information set and the prototypical feature value in memory. Although feature values are considered the basis for the match between a given speech signal and its prototypical representation, these values are not observed directly. Instead, they are inferred from the response patterns obtained from separate auditory-only and visual-only recognition experiments. To predict bimodal performance, auditory-visual feature values are assumed to be proportional to the product of unimodal feature values. According to Massaro (1987) and Massaro and Friedman (1990) a multiplicative combination of feature values predicts an optimal level of performance given multiple sources of information. The multiplicative rule for AV integration used by Massaro is given by

$$P_{AV}(R_j|S_i) = \frac{P_A(R_j|S_i) \times P_V(R_j|S_i)}{\sum_{k=1}^N P_A(R_k|S_i) P_V(R_k|S_i)}, \quad (3)$$

where $P_{AV}(R_j|S_i)$ is the conditional probability that response R_j is given when the AV stimulus S_i is presented, and $P_A(R_j|S_i)$ and $P_V(R_j|S_i)$ are the conditional probabilities for the A and V unimodal cases, respectively. The denominator is the sum of the conditional probabilities in the i th row of the confusion matrix and normalizes the feature values for the N responses to conditional probabilities.

In our implementation of the FLMP, a fixed form of the model was used where the parameter values representing the goodness-of-fit of any particular token (presented in either the A or V modality) to its prototypical memory representation are assumed to be equal to the unimodal performance level for that token.³ In other words, the truth values were derived directly from the original confusion matrices for the unimodal conditions. For example, if a subject correctly identified auditory /ba/ 85% of the time, then $P_A(R_{ba}|S_{ba}) = 0.85$. Similarly, if the subject responded "va" 10% of the time given this same auditory stimulus, then $P_A(R_{va}|S_{ba}) = 0.10$, and so on. Thus the set of A and V conditional probabilities derived from the A and V confusion matrices were used directly with the multiplicative rule to predict AV responses.

Braida (1991) has shown that the fixed FLMP model makes reasonably good predictions of multimodal scores without systematically under- or over-predicting observed accuracy. However, the model has a tendency to underestimate AV performance when A performance is particularly poor and overestimate AV performance when A performance is relatively good. As with the PRE model, integration efficiency was defined as the difference between predicted and observed AV consonant recognition scores.

c. Blamey et al. (1989)—PROB. According to the model proposed by Blamey et al. (1989), AV speech recognition errors occur only when there are simultaneous errors in both A and V recognition. Model predictions are stated in

terms of percentage of information transmitted (%IT). Thus given the probability of an auditory error $[1 - \%IT(A)]$ and visual error $[1 - \%IT(V)]$, the AV error rate is given by

$$AV_{ERR} = (1 - \%IT_A)(1 - \%IT_V), \quad (4)$$

and the AV score is given by

$$1 - (AV_{ERR}). \quad (5)$$

The advantage of this simple probabilistic model (PROB) is that it can be used to predict a wide range of speech recognition scores, including information transmission scores for speech features (e.g., voicing, manner-of-articulation, etc.), words, and sentence recognition scores. The disadvantage is that human observers often do better than the model predicts (Blamey *et al.*, 1989), suggesting that the model is not optimal. The Blamey *et al.* model represents a probabilistic combination of the A recognition response and the V recognition response made independently of each other. However, Braida (1991) showed that predicted AV recognition scores are generally higher when information from the two modalities are integrated before response labels are assigned. Thus actual observed AV scores that exceed those predicted by Blamey *et al.* (1989) are likely the result of the observers' use of prelabeling integration processes, and, the greater the deviation from predicted score, the greater the integration efficiency.

3. Sentence recognition

As with consonant recognition, sentence recognition was measured separately for A, V, and AV presentations. Speech materials consisted of the IEEE/Harvard (1969) sentences. These are 720 phonetically balanced low-context sentences each containing five key words (e.g., "The *birch canoe slid* on the *smooth planks*"). The sentences are organized into 72 lists with 10 sentences in each list. These stimulus materials were filmed at the Massachusetts Institute of Technology using the same female talker used for the VCV materials and dubbed onto an optical disc (Panasonic TQ-3031F). The audio portion of the sentences were treated in the same manner as the consonants.

A and AV sentence recognition was measured at a S/N ratio of 0 dB. For AV presentations, the subjects viewed a 19-in. color video monitor (SONY PVM 2030) from a distance of approximately 5 ft.

The subjects were asked to write down what they perceived in A, V, and AV presentation modes. Subjects were encouraged to respond with as much of the sentence as they could, and to guess whenever they were uncertain. All sentences were scored for the percentage of key words correctly identified. A strict scoring criterion was used in which all tense (e.g., "-ed") and number (e.g., "-s") affixes were required to be correct for the key word to be correct. On the other hand, homophones were counted as correct (e.g., "read" for "red").

For each test condition (A, V, and AV), five lists of sentences (ten sentences/list) were presented. Thus each score was based on 250 words (5 words per sentence, 10 sentences per list, 5 lists per condition). The order of the test

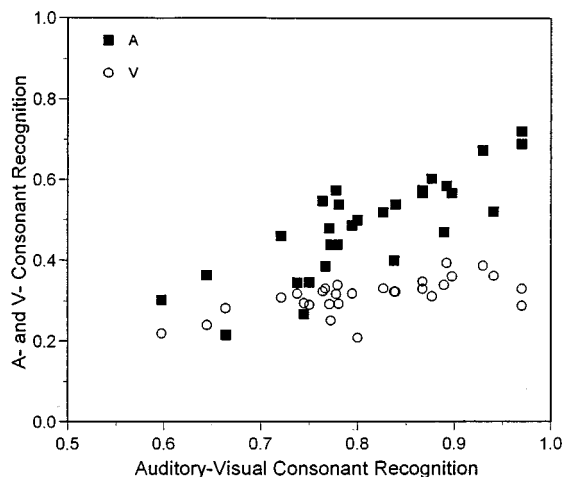


FIG. 3. A and V consonant recognition as a function of AV consonant recognition.

conditions and lists were randomized for each subject. No feedback regarding the correctness of responses was provided.

4. Summary

The specific question addressed by the various speech measures obtained in this study was how much of the individual variability in AV sentence recognition could be accounted for by variability in AV consonant recognition? Further, it was assumed that AV consonant recognition is determined primarily by auditory and visual cue extraction abilities and the efficiency at which the unimodal cues could be integrated. The ability to extract auditory and visual cues were determined by analyses of error patterns made in separate A and V consonant recognition tests. To measure integration ability, three models of integration (PRE, FLMP, and PROB) were used to predict AV consonant recognition scores from independent measures of A and V consonant recognition. Integration efficiency was defined as the difference between predicted and observed AV consonant recognition scores. This is the first report, as far as we are aware, to describe individual differences in AV integration efficiency, and how these differences may relate to speech recognition performance.

II. RESULTS

A. Consonant recognition

Variability in AV consonant recognition was examined in terms of A and V consonant recognition. In addition, consonant confusions made in each of the unimodal conditions were examined to determine the relation between AV recognition and unimodal feature recognition.

Focusing initially on overall recognition accuracy, Fig. 3 shows A, V, and AV consonant recognition scores for each of the 29 subjects. Auditory-visual recognition scores (displayed along the abscissa) ranged from 60% to 98% (mean = 81%, s.d. = 9.5%), whereas A recognition scores ranged from 20% to 74% (mean = 49%, s.d. = 12.3%). Speechreading (i.e., V-only) scores were less variable across subjects and

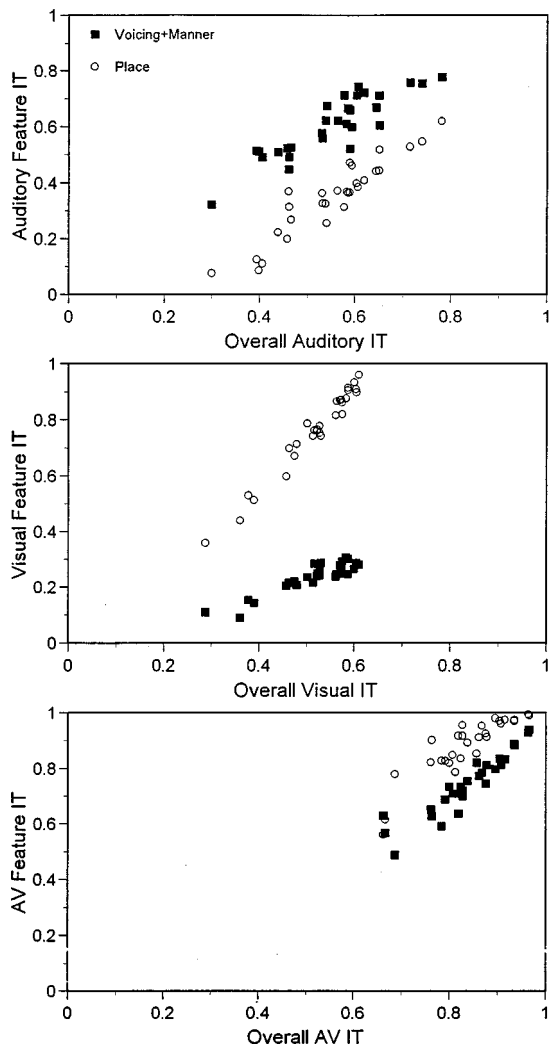


FIG. 4. Voicing-plus-manner and place information transmission for A, V, and AV modalities as a function of overall information transmitted (IT).

ranged from 21% to 40% (mean=31%, s.d.=4.3%). All subjects were able to benefit substantially from the combination of audition and speechreading with the mean performance difference between AV and A test conditions equal to 32% (s.d.=7%). For these materials, this amount of gain is roughly equivalent to a 5–6 dB improvement in signal-to-noise ratio (Grant and Walden, 1996a). Also apparent from the figure is the strong relation between AV and A scores ($r=0.82$) and the weaker relation between AV and V scores ($r=0.63$). Both of these correlations were highly significant ($p<0.001$).

The correspondence between unimodal and bimodal performance shown in Fig. 3 is based solely on overall recognition accuracy in each of the three receiving conditions. It is likely that even higher correlations between unimodal and bimodal performance would result if the patterns of A and V confusions were considered. Previous data obtained by Grant and Walden (1996a), as well as the predicted AV scores shown in Fig. 2, suggest that AV consonant recognition performance is determined in large part by the accuracy with which auditory voicing plus manner cues and visual place cues are extracted from the unimodal conditions. To investigate this, a feature-based information analysis (Miller and

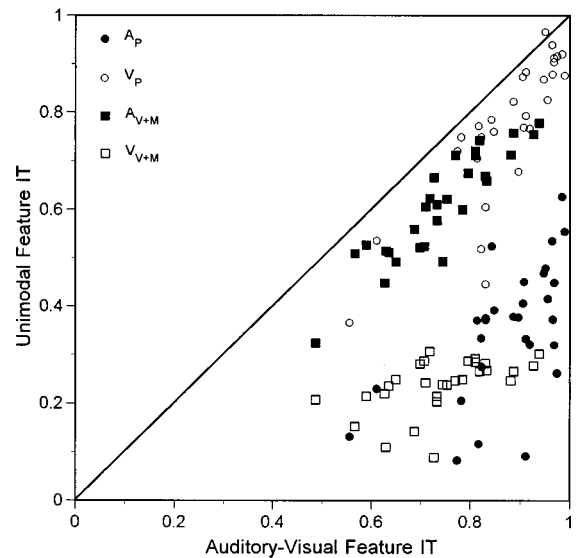


FIG. 5. A and V feature transmission scores as a function of AV feature transmission scores.

Nicely, 1995) was performed on the unimodal and bimodal confusion matrices of the 29 subjects. The accuracy of information transmission (IT) for a variety of speech features was determined, including a 7-category voicing+manner (V+M) feature and a 6-category place (P) feature as described in Table I. The three panels displayed in Fig. 4 show the results for the V+M and P features as a function of the overall IT for A (top), V (middle), and AV (bottom) conditions, respectively. As expected for auditory presentations (top panel) given the subjects' hearing losses and the relatively poor S/N, AV_{V+M} cues (mean=0.61) were received more accurately than A_p cues (mean=0.35). In contrast (middle panel), V_p cues (mean=76%) were received more accurately than V_{V+M} (mean=0.24). In the AV condition (bottom panel), AV_p cues (mean=0.88) were received with greater accuracy than AV_{V+M} cues (mean=0.74). This is noteworthy in that place cues, as opposed to voicing or manner cues, have been shown consistently to be more susceptible to the deleterious effects of noise and hearing loss (Miller and Nicely, 1955; Rabinowitz *et al.*, 1996). The implications of this observation for hearing-aid development and rehabilitation will be discussed later.

An attempt was made to relate unimodal feature recognition to AV feature recognition to explore further the connection between A and V segmental cues and AV recognition. Figure 5 shows scatter plots of four unimodal-to-bimodal relations: AV_p vs A_p (filled circles), AV_p vs V_p (open circles), AV_{V+M} vs A_{V+M} (filled squares), and AV_{V+M} vs V_{V+M} (open squares). All four correlations were significant ($r\geq 0.53$, $p\leq 0.003$). However the strongest correlations were between AV_p recognition and V_p ($r=0.83$) and between AV_{V+M} recognition and A_{V+M} ($r=0.91$). This is not to imply that A_p cues did not contribute to AV_p recognition or that V_{V+M} cues did not contribute to AV_{V+M} recognition. In fact, by combining V_p and A_p scores into a single P score (e.g. by using the PROB model), the correlation with AV_p ($r=0.90$) is significantly higher than the correlation between V_p and AV_p . A similar analysis regarding

AV_{V+M} recognition showed that combining the two unimodal V+M scores hardly improved the correlation previously obtained with the A_{V+M} score alone ($r=0.93$). Therefore, in describing the AV feature IT for the 29 hearing-impaired subjects, AV_{V+M} recognition was determined primarily by A_{V+M} whereas AV_P was determined by the combined influences of V_P and A_P.

B. AV consonant integration

Auditory-visual consonant recognition is assumed to be determined primarily by the amount and type of A and V cues that can be extracted from the speech signal (unimodal cue extraction), and the efficiency with which these cues can be combined across the two modalities (bimodal integration). In order to separate these two factors, a measure of AV integration must be devised. In this section, three models capable of predicting AV consonant recognition from A and V consonant recognition are used to derive such measures. Each model is used to predict AV consonant identification accuracy for individual subjects. The differences between obtained and predicted scores are then used as measures of individual integration skill. It should be noted that subjects with relatively good integration skills need not achieve better overall AV recognition scores than subjects with relatively poor integration skills. As we are using the term, integration only refers to the ability to combine the cues derived from the separate modalities. Therefore, subjects with very poor A and/or V cue resolution might still be excellent integrators if they use these cues to their fullest potential in bimodal perception. The net result, depending on the amount of A and V cues recognized, might still be a relatively poor overall AV score. Conversely, subjects with excellent cue resolution but relatively poor integration skills may end up with a relatively high AV score by virtue of a high A or V score alone. These examples suggest that integration skill might be better related to AV benefit than to AV score. For the purpose of this study, AV benefit was defined as $(AV-A)/(100-A)$, or the recognition improvement relative to the total possible improvement given an individual's A-alone score (Sumbly and Pollack, 1954).

Predicted versus observed AV consonant recognition scores are shown in Fig. 6. AV predictions made by the PRE, FLMP, and PROB models of integration are shown separately in the top, middle, and bottom panels of the figure. The Pearson correlations between observed and predicted AV recognition scores were fairly similar for the three models: 0.89 for the PRE, 0.83 for the FLMP, and 0.89 for the PROB. As indicated by the position of the data points relative to the line $AV_{OBSERVED}=AV_{PREDICTED}$ shown in each panel, predictions by the PRE model were either equal to or greater than obtained performance, whereas predictions by the FLMP and PROB models generally underpredicted AV performance.⁴ This was especially true for the FLMP when the input unimodal scores were low. In the fixed FLMP model, stimuli identified correctly in one modality but incorrectly in the other are predicted to be incorrect in the combined AV condition [see Eq. (3)]. As Braida (1991) noted, the fixed FLMP model does not properly account for struc-

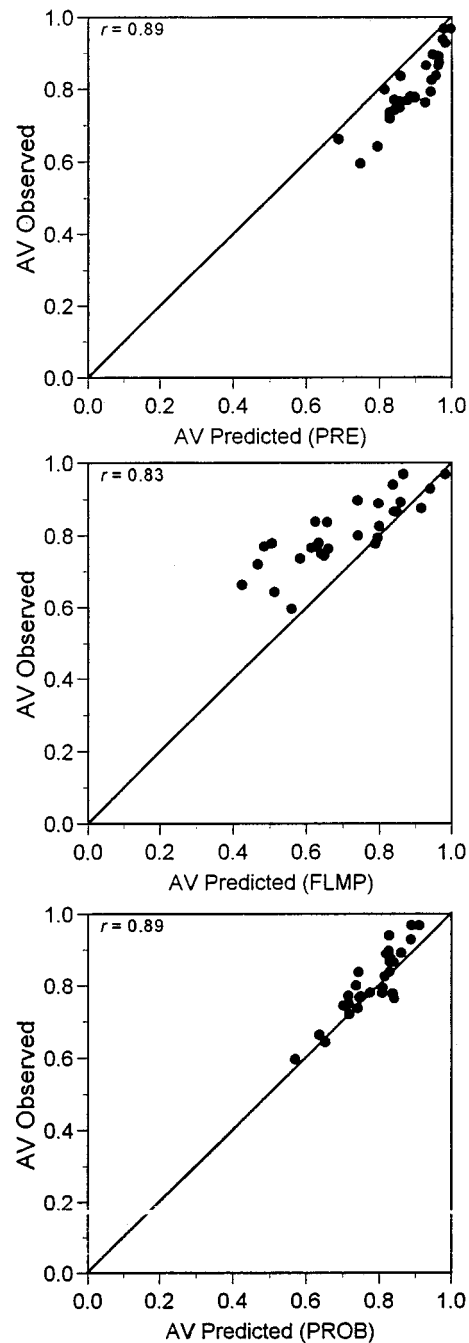


FIG. 6. PRE, FLMP, and PROB model predictions of AV consonant recognition.

tured errors and relies too heavily on unimodal accuracy. In contrast, the PRE model focuses more on the consistency of unimodal responses (as determined by MDS) and not necessarily on accuracy. Thus the PRE model makes a prediction of optimal performance (not necessarily optimal fit) and would therefore be expected to over predict observed scores.

In Fig. 7, the derived measure of integration skill (i.e., the difference between predicted and observed scores for individual subjects) is shown with respect to the amount of relative AV benefit obtained. When compared in terms of the ability to predict the relative benefit provided by speechreading, the derived measure based on PRE model predictions accounted for approximately 56% of the variance in relative

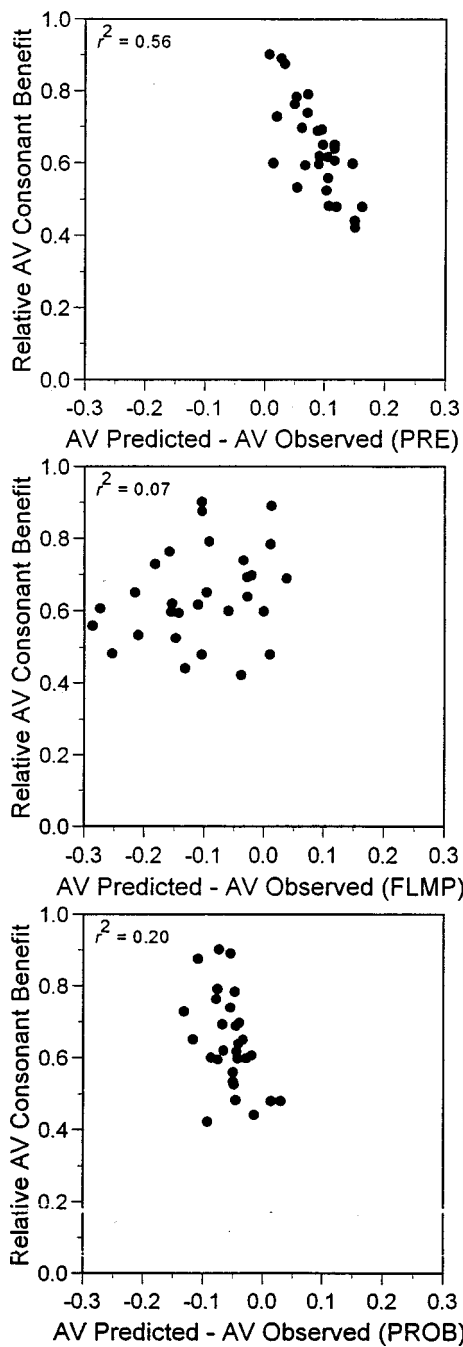


FIG. 7. Relation between derived integration measure ($AV_{\text{PREDICTED}} - AV_{\text{OBSERVED}}$) and relative AV benefit for consonants.

AV benefit. In comparison, the PROB model accounted for roughly 20% of the variance, whereas the FLMP accounted for about 7% of the variance. The correlations between derived measures of integration efficiency and relative AV benefit were significant for the PRE and PROB models, but not for the FLMP. Empirically, the PRE model appears to provide a better estimate of integration efficiency than either the FLMP or PROB model.

Because the PRE model predicts optimum AV recognition performance, anything less than perfect AV consonant recognition for subjects who are well predicted by the PRE model is probably due to poor cue extraction and not to poor integration. On the other hand, subjects whose AV consonant

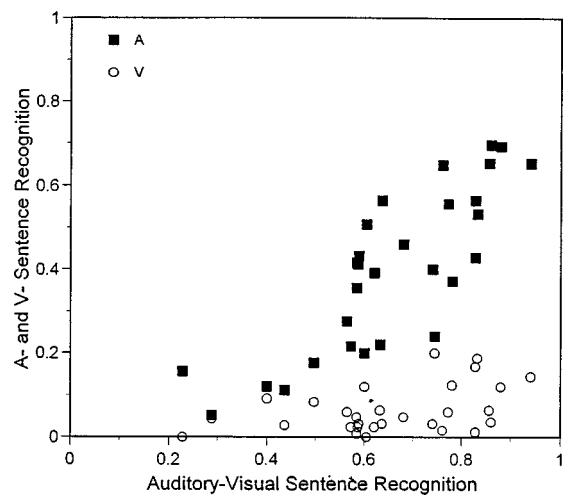


FIG. 8. A and V sentence recognition as a function of AV sentence recognition.

recognition scores fall below PRE model predictions may not be integrating all the unimodal cues available to them. For these subjects, improved AV speech recognition scores may be achieved by focusing on integration training in addition to cue extraction. The amount of AV improvement in consonant recognition that could be expected with integration training is indicated by the performance difference between predicted and observed AV scores shown along the abscissa in the top panel of Fig. 7. The average difference was 8.5% and the maximum difference was 16.1%.

C. Sentence recognition

Recognition results for IEEE/Harvard sentences are shown in Fig. 8. As in Fig. 3, A and V recognition results are shown in relation to AV recognition scores. Auditory-visual sentence recognition of key words (shown along the abscissa) ranged from 23% to 94% (mean=65.5%; s.d.=17.5%). Auditory sentence recognition ranged from 5% to 70% (mean=39.7%; s.d.=19.1%) and visual sentence recognition scores ranged from 0% to 20% (mean=6.5%; s.d.=5.6%). Although there was no appreciable correlation between the subject's A and V scores, and in spite of the very low speechreading scores overall, all subjects benefited from the addition of visual cues. Every subject's AV score was better than his/her A score, even when the V score was 0% correct. A strong correlation was evident between A and AV sentence recognition ($r=0.82$; $p<0.001$). The correlation between V and AV sentence recognition was also significant, but weaker ($r=0.44$; $p<0.02$).

The benefit obtained from combining A and V cues in sentence recognition varied substantially across individuals. To address this issue, relative AV benefit scores, $(AV-A)/(100-A)$, were computed for each subject. The average relative benefit was 44% (s.d.=17.8%) with a maximum of 83% and a minimum of 8.5%. The large individual differences in the amount of relative benefit received were only weakly related to the subject's A performance ($r=0.34$). A much better accounting of individual differences in relative AV benefit for sentences was provided by subjects' speechreading scores, in that better speechreaders obtained more AV

benefit for sentences ($r=0.72$).⁵ The relatively strong correlation between V sentence recognition and relative AV sentence benefit, in conjunction with the weaker relation observed between visual consonant recognition and relative AV consonant benefit ($r=0.55$), suggests that better speechreaders of sentence materials are able to extract more visual information than the poorer speechreaders with regard to certain cues such as word segmentation and stress, both of which would prove useful in the AV sentence condition (Risberg, 1974; Risberg and Lubker, 1978; Grant and Walden, 1996b). Obviously, these cues would play a much smaller role in consonant recognition than in sentence recognition which might help explain the apparent greater significance of speechreading for AV sentences than for AV consonants.

A second possibility may be that the better speechreaders engage useful higher-level cognitive skills in addition to the bottom-up information extracted from the visual speech signal. With meaningful sentence materials, speechreading requires that subjects not only extract signal cues from visual speech movements, but form linguistic wholes from perceived fragments as well. This ability to perform *perceptual closure* (Wertheimer, 1938; O'Neill and Oyer, 1961; Watson *et al.*, 1996) is useful regardless of modality, especially when the input signals are ambiguous. Therefore, it is possible that better speechreaders of sentence materials have better perceptual closure skills than poorer speechreaders, which would facilitate AV sentence benefit.

D. Relations among consonant and sentence recognition

According to Fig. 1, auditory-visual consonant recognition is primarily a function of A and V cue extraction and cue integration. AV sentence recognition, on the other hand, is a function of A and V signal-cue extraction, cue integration, lexical processes, memory processes, and the use of linguistic and world knowledge in conjunction with semantic and syntactic context. If this conceptual framework is correct, then the proportion of variance in AV sentence recognition data not accounted for by AV consonant recognition is most likely due to individual differences in *top-down* speech recognition processes.

Figure 9 shows the relation between consonant recognition and sentence recognition for A, V, and AV conditions. In all three receiving modalities, the correlations were highly significant ($p<0.007$). For the A and AV conditions, respectively, 52% and 54% of the variance observed in sentence recognition was accounted for by consonant recognition. For the V condition, 25% of the variance in visual sentence recognition could be accounted for by visual consonant recognition. This reduction in the amount of variance explained for the V modality (compared to either A or AV modalities) is probably due to the relatively narrow range of scores observed for visual consonant recognition (mean = 31.2%, s.d.=4.3%). In general, however, the data show that variability in A and AV sentence recognition is determined to a great extent by *bottom-up* processes related to cue extraction and cue integration.

With regard to individual differences in the benefit received from combining A and V cues in sentence recogni-

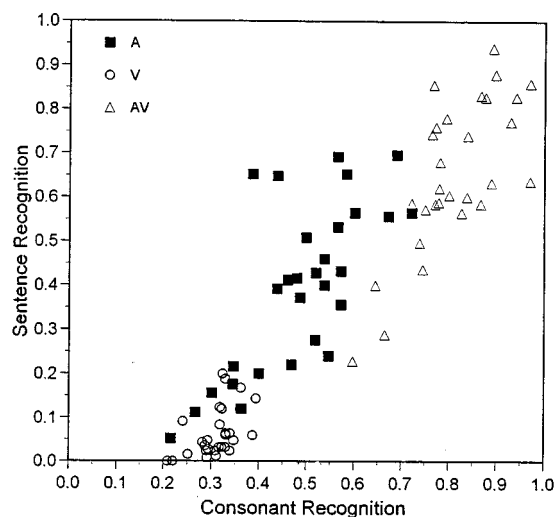


FIG. 9. Relation between consonant recognition and sentence recognition for A, V, and AV modalities.

tion, the results suggested that speechreading ability, measured either with nonsense syllables or with meaningful sentences, was the most consistent unimodal predictor of AV sentence benefit. For example, measures of relative AV sentence benefit correlated significantly with visual consonant recognition ($r=0.66$, $p<0.001$) and with visual sentence recognition ($r=0.72$, $p<0.001$). In contrast, relative AV benefit was not significantly correlated with auditory consonant or auditory sentence recognition. Finally, although the correlations between measures of relative AV benefit for consonants and relative AV benefit for sentences was significant ($p=0.02$), relative AV consonant benefit accounted for only 18% of the variance in relative AV sentence benefit.

III. DISCUSSION

The conceptual framework schematized in Fig. 1 highlighted a variety of *bottom-up* and *top-down* sources of information that are likely to be important in AV speech recognition. The general purpose of this work was to further delineate the various factors and perceptual processes that determine individual variability in AV sentence recognition. In this study, we focused on consonant recognition and AV integration at the segmental level and the relation between consonant recognition and the recognition of words in sentences.

Auditory-visual consonant recognition can be reasonably well described as a simple combination of visual place-of articulation cues and auditory manner-plus-voicing cues. This rather simple conception of the problem of predicting AV consonant recognition in individual listeners is supported by the results of Massaro's FLMP and Braida's Prelabeling Model of integration, and by earlier work by Grant and Walden (1996a). As shown in Fig. 2, model predictions of AV recognition given average speechreading skill, depend primarily on which speech features are resolved in the auditory condition, and not on the overall accuracy of auditory speech recognition. For example, individuals who receive only voicing or manner cues through audition are predicted to have higher AV consonant recognition scores

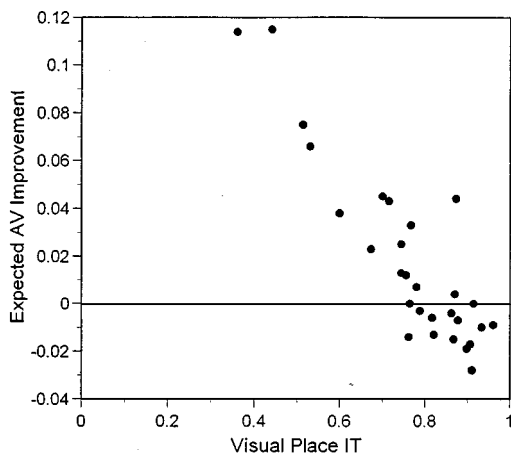


FIG. 10. Predicted AV consonant improvement that would likely result from speechreading training.

than individuals who receive only auditory place-of-articulation cues. These model predictions support the idea that integration of complementary A and V speech features results in higher AV recognition scores than integration of redundant features, as is the case when the A modality contributes primarily place-of-articulation cues. A similar conclusion can be drawn from the study by Grant and Walden (1996a). In that study, normal-hearing listeners were presented with filtered bands of speech in combination with speechreading. Low-frequency bands of speech, which conveyed primarily voicing and manner information, resulted in higher AV recognition of consonants than high-frequency bands which conveyed mostly place information. This was the case, even though the high-frequency bands were often more intelligible than the low bands when presented without speechreading.

In the present study, the ability to extract V+M and P cues from the A and V conditions was quite variable across subjects. For example, when compared to previous speechreading results obtained with normal-hearing subjects (Grant and Walden, 1996a), the HI subjects displayed a significantly wider range of performance for the recognition of V_p cues (36%–96% for the HI subjects as compared to 73%–94% for the NH subjects). For HI subjects with relatively low V_p recognition (below 70%) speechreading training would seem appropriate. Using Braida's model we can estimate the amount of AV improvement expected to result from such training. Model fits were computed for each HI subject using the average visual confusion matrix from Grant and Walden (1996a) in combination with individual auditory confusion matrices and compared to original model fits obtained with the subject's own visual confusion matrix. This average or *generic* visual matrix had a V_p score of 0.87 and represents conservatively what we feel can be achieved through speechreading training (Walden *et al.*, 1977). Figure 10 shows the predicted AV improvements obtained for each subject. According to this analysis, subjects with V_p recognition less than 0.7 would show improvements between 3% and 11.5%.

With regard to auditory consonant features, it is clear that strategies for improving the information transmission of

the V+M feature need to be developed, either through auditory training or signal processing. For our subjects, the A_{V+M} feature was received more accurately than the A_p feature, whereas the AV_{V+M} feature was received less accurately than the AV_p feature. Note that the finding for AV feature recognition is precisely the opposite result of classic feature studies of auditory-only listening conditions (e.g., Miller and Nicely, 1955). Manner and voicing information can be represented by time-intensity envelope cues which tend to be more resistant to the effects of hearing loss or background noise than are the rapid spectral transitions and spectral shapes associated with A_p cues (Van Tasell *et al.*, 1987; Rosen, 1989). It is reasonable, therefore, for researchers and clinicians concerned with improving auditory speech recognition to advocate hearing aid fitting and other rehabilitative strategies to improve A_p reception (Rabinowitz *et al.*, 1992). However, to improve AV speech recognition, improving A_{V+M} reception (as opposed to A_p reception) needs to be the focus of rehabilitative efforts, assuming that the individual has at least average speechreading and integration abilities.

The integration processes by which A and V speech cues are combined are not well understood. The results of our efforts to measure this ability across HI listeners suggests that there is substantial intersubject variability in integration efficiency and that better integrators are likely to derive more AV benefit than poorer integrators, at least for consonant recognition.

The integration abilities of individual listeners currently are not subject to direct observation. Instead, they must be derived from speech recognition performance obtained from unimodal and bimodal receiving conditions, and models of the integration process. The approach used in this study was an attempt to partial out the contributions of information extraction and information processing as they relate to AV speech recognition. Braida's PRE model is a theoretically optimal integrator which produces the best possible recognition score for each subject, given their A and V consonant recognition data. The predictions of this model can be used, among other clinical applications, to estimate the performance that may be possible if the subject was able to integrate the available information from the two modalities perfectly. The difference between predicted and observed AV scores suggests that, on average, a gain of 8.5% (maximum gain=16.1%) could be expected for the subjects of this study with appropriate integration training. This model also allows us to test the effectiveness of certain rehabilitation strategies or signal processing algorithms with respect to predicted AV gain. For example, if we eliminated a particular auditory consonant confusion (e.g., /aba/ vs /ava/) through training or signal processing, it would be possible to calculate the predicted improvement in overall AV recognition score. Similarly, one could estimate the overall benefit of combining several rehabilitation strategies, such as speechreading training and integration training. For example, for the 14 subjects who either received less than 70% place information in the V condition or who were over-predicted by the PRE model by more than 10%, we would expect a potential combined training benefit (speechreading training and integration training) of 9% to 26% depending on the individual subject.⁶

In comparing speech recognition results with nonsense aCa syllables and IEEE sentences, significant correlations were found within receiving condition (i.e., $A_{\text{CONS}} \propto A_{\text{SENT}}$, $V_{\text{CONS}} \propto V_{\text{SENT}}$, and $AV_{\text{CONS}} \propto AV_{\text{SENT}}$). Especially noteworthy was the finding that over 50% of the variability in AV (or A) sentence recognition could be accounted for by AV (or A) consonant scores. In addition, the relative AV benefit for consonants was significantly correlated with the relative benefit for sentences ($r=0.43$, $p=0.02$). Thus AV recognition of IEEE sentences and the amount of benefit provided by visual cues to audition is determined, in large measure, by the recognition of medial consonants, which can be fairly well predicted by separate measures of A and V consonant recognition and a measure of AV consonant integration.

When discussing the relationship between AV consonant recognition and AV sentence recognition, it is important to consider the many differences between these two sets of materials. It is well known that fluent speech productions like the IEEE sentences and carefully articulated nonsense syllable productions like the vCv consonant set differ with regard to phonetic (speaking rate, segment duration, etc.), phonological (vowel neutralization, flapping, etc.), lexical, morphosyntactic, and semantic factors. These differences are likely to weaken the association between segment articulation scores and sentence intelligibility scores. In addition, they suggest several possible processes important for fluent speech recognition, such as the ability to use lexical, semantic, and grammatical constraints (Fletcher, 1953; Boothroyd and Nittrouer, 1988), short-term memory, and the processing speed required to make lexical decisions (Seitz and Rakerd, 1996), that may vary significantly across individuals. Given this array of different factors separating the recognition of nonsense medial consonants in a single vowel context from that of meaningful sentence materials, the strength of the observed association across HI subjects found in this study is quite remarkable.

In summary, individual performance on AV speech recognition tasks involving words or sentences ultimately depends on 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 and memory processes. Our efforts thus far to evaluate these factors in individual subjects have focused mainly on nonsense syllable recognition (consonant recognition and the recognition of selected speech features), the recognition of certain prosodic contrasts (Grant and Walden, 1996b), and segmental integration skills. These studies have shown that the benefit of visual cues in AV speech recognition can be readily interpreted in terms of the degree of redundancy between A and V cues, with greater redundancy leading to smaller benefits. Furthermore, there appear to be substantial individual differences regarding the efficiency with which A and V segmental cues are integrated. These differences in integration abilities can also lead to fairly large differences in the amount of AV benefit observed. Taken together, an accounting of the redundancy between A and V features and a measure of integration efficiency can account for the bulk of the variability in AV segment recognition. Additional work is required, however,

to more fully account for the variability observed in the recognition of words and sentences. Ongoing efforts to expand this work to include the recognition of additional speech segments in different phonetic environments, multi-syllabic consonant and vowel sequences (to mimic speech rates found in fluent productions), measures of AV integration in connected speech, and measures of lexical redundancy and semantic context usage across individual subjects, will no doubt improve our overall understanding of AV speech recognition.

ACKNOWLEDGMENTS

This research was supported by Grant Nos. DC 00792 and DC 01643 from the National Institute on Deafness and Other Communication Disorders to Walter Reed Army Medical Center, and by the Clinical Investigation Service, Walter Reed Army Medical Center, under Work Unit # 2528. The authors would like to thank Dr. Louis Braida for fitting our consonant data with the Prelabeling Model of Integration. We would also like to thank Drs. Dominic Massaro and Michael Cohen for help in implementing the fixed form of the FLMP, and for numerous helpful discussions regarding theories of information extraction and information processing. Helpful comments on a earlier draft of the manuscript were provided by Dr. Winifred Strange and by two anonymous reviewers. 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.

¹The visual confusion matrix used for these model fits was obtained using the same consonant stimuli used in the present study. The results of 11 normal-hearing subjects were pooled to form a single confusion matrix. The overall percent information transmitted (IT) was 55%. The percent place IT was 87%.

²AV predictions based on Massaro's (1987) Fuzzy Logical Model of Perception (FLMP) produced nearly identical results as those shown in the figure. Descriptions of the PRE and FLMP model are presented in the methods section.

³In Massaro's evaluation of the FLMP, response probabilities obtained from the unimodal A and V confusion matrices are treated as approximations to the auditory and visual truth values. The program STEPIT (Chandler, 1969) is used to derive a set of optimal feature values which are then used in conjunction with Eq. (3) to predict the AV response. This iterative process is accomplished with prior knowledge of the AV response probabilities by minimizing the average root mean square deviation between the predicted and obtained AV responses, the original and adjusted A responses, and the original and adjusted V responses. Thus by manipulating the original A and V response patterns by small amounts, application of the multiplicative rule results in a better fit than if the multiplicative rule were applied to the original unimodal matrices. This version of the FLMP is known as the variable FLMP. The variable FLMP although providing excellent fits to AV data, can fail to demonstrate differences among subjects with similar A and V matrices but different AV matrices. Such subjects are likely to be fit equally well by the variable FLMP model which, according to Massaro, suggests that the integration processes are similar. However, for the purposes of measuring individual differences in AV integration, the model needs to distinguish among subjects according to the efficiency at which A and V cues are integrated, even when subjects use similar integration processes.

⁴The number of subjects whose observed AV consonant scores exceeded their predicted scores was 25 and 27 for the FLMP and PROB models, respectively.

⁵Although the average sentence speechreading score was extremely low, there was nevertheless a range of scores between 0% and 20% correct to

- support the correlations between speechreading performance and relative AV sentence benefit. It should be noted that with low-context IEEE materials, speechreading scores of 15%–20% reflect excellent speechreading skills.
- ⁶Estimates of training benefit assume that speechreading training would increase the number and recognition accuracy of place-of-articulation categories to that of the average normal-hearing subjects studied by Grant and Walden (1996a). Further, it was assumed that integration training would result in integration efficiency ratios (i.e., $AV_{OBSERVED}/AV_{PREDICTED}$) between 0.9 and 1.0.
- ANSI (1989). ANSI S3.6–1989, “Specifications for audiometers” (ANSI, New York).
- Blamey, P. J., Cowan, R. S. C., Alcantara, J. I., Whitford, L. A., and Clark, G. M. (1989). “Speech perception using combinations of auditory, visual, and tactile information,” *J. Rehab. Res. Dev.* **26**, 15–24.
- Boothroyd, A., and Nittrouer, S. (1988). “Mathematical treatment of context effects in phoneme and word recognition,” *J. Acoust. Soc. Am.* **84**, 101–114.
- Borg, I., and Lingoes, J. (1987). *Multidimensional Similarity Structure Analysis* (Springer-Verlag, New York).
- Braida, L. D. (1991). “Crossmodal integration in the identification of consonant segments,” *Q. J. Exp. Psych.* **43**, 647–677.
- Chandler, J. P. (1969). “Subroutine STEPIT—Finds local minima of a smooth function of several parameters,” *Behav. Sci.* **14**, 81–82.
- Diehl, R., and Kluender, K. (1989). “On the objects of speech perception,” *Ecological Psychol.* **1**, 121–144.
- Fletcher, H. (1953). *Speech and Hearing in Communication* (Krieger, Huntington, NY).
- Grant, K. W., and Walden, B. E. (1996a). “Evaluating the articulation index for auditory-visual consonant recognition,” *J. Acoust. Soc. Am.* **100**, 2415–2424.
- Grant, K. W., and Walden, B. E. (1996b). “The spectral distribution of prosodic information,” *J. Speech Hear. Res.* **39**, 228–238.
- Green, D. M., and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics* (Wiley, New York).
- Greenberg, S. (1995). “The ears have it: The auditory basis of speech perception,” in *Proceedings of the XIIIth International Congress of Phonetic Sciences*, edited by K. Elenius and P. Branderud, ICPhS 95, Stockholm, Sweden, Vol. 3, pp. 34–41.
- Halle, M., and Stevens, K. N. (1991). “Knowledge of language and the sounds of speech,” in *Music, Language, Speech, and Brain*, edited by J. Sundberg, L. Nord, and R. Carlson (Macmillan, Basingstoke, Hampshire), pp. 1–19.
- IEEE. (1969). *IEEE Recommended Practice for Speech Quality Measurements* (Institute of Electrical and Electronic Engineers, New York).
- Klatt, D. H. (1989). “Review of selected models of speech perception,” in *Lexical Representation and Process*, edited by W. Marslen-Wilson (MIT, Cambridge, MA), pp. 169–226.
- Kruskal, J. B., and Wish, M. (1978). *Multidimensional Scaling* (Sage, Beverly Hills, CA).
- Kučera, F., and Francis, W. (1967). *Computational Analysis of Present Day American English* (Brown U.P., Providence, RI).
- Lieberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). “Perception of the speech code,” *Psychol. Rev.* **74**, 431–461.
- Lindblom, B. (1996). “Role of articulation in speech perception: Clues from production,” *J. Acoust. Soc. Am.* **99**, 1683–1692.
- Macmillan, N. A., Goldberg, R. F., and Braida, L. D. (1988). “Vowel and consonant resolution: Basic sensitivity and context memory,” *J. Acoust. Soc. Am.* **84**, 1262–1280.
- Massaro, D. W. (1987). *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry* (Earlbaum, Hillsdale, NJ).
- Massaro, D. W., and Friedman, D. (1990). “Models of integration given multiple sources of information,” *Psychol. Rev.* **97**, 225–252.
- McGurk, H., and MacDonald, J. (1976). “Hearing lips and seeing voices,” *Nature (London)* **264**, 746–748.
- Miller, G. A., and Nicely, P. E. (1955). “An analysis of perceptual confusions among some English consonants,” *J. Acoust. Soc. Am.* **27**, 338–352.
- Ohala, J. J. (1996). “Speech perception is hearing sounds, not tongues,” *J. Acoust. Soc. Am.* **99**, 1718–1725.
- O’Neill, J. J., and Oyer, H. J. (1961). *Visual Communication for the Hard of Hearing* (Prentice-Hall, Englewood Cliffs, NJ).
- Rabinowitz, W. M., Eddington, D. K., Delhorne, L. A., and Cuneo, P. A. (1992). “Relations among different measures of speech reception in subjects using a cochlear implant,” *J. Acoust. Soc. Am.* **92**, 1869–1881.
- Reisberg, D., McLean, J., and Goldfield, A. (1987). “Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli,” in *Hearing by Eye: The Psychology of Lipreading*, edited by B. Dodd and R. Campbell (Earlbaum, Hillsdale, NJ), pp. 97–113.
- Risberg, A. (1974). “The importance of prosodic speech elements for the lipreader,” in *Visual and Audiovisual Perception of Speech VI. Danavox Symposium*, edited by H. B. Nielson and B. Klamp, *Scand. Audiol. Suppl.* **4**, 153–164.
- Risberg, A., and Lubker, J. L. (1978). “Prosody and speechreading,” *Speech Transmission Lab—Quarterly Progress Status Report*, **4**, 1–16.
- Rosen, S. (1989). “Temporal information in speech and its relevance for cochlear implants,” in *Cochlear Implant Acquisitions and Controversies*, edited by B. Fraysse and N. Cochard (Cochlear AG, Basel), pp. 3–26.
- Seitz, P. F., and Rakerd, B. (1996). “Hearing impairment and same-different reaction time,” *J. Acoust. Soc. Am.* **99**, 2602(A).
- Stork, D. G., and Hennecke, M. E. (Eds.). (1996). *Speechreading by Humans and Machines* (Springer-Verlag, New York), Models, Systems, and Applications Proceedings of the NATO Advanced Study Institute on Speechreading by Man and Machine, held in Castera-Verzudan, France, 28 August–8 September 1995, NATO ASI Series F: Computer and Systems Sciences, Vol. 150, 1996.
- Stevens, K. N. (1989). “On the quantal nature of speech,” *J. Phon.* **17**, 3–45.
- Stevens, K. N., and House, A. S. (1972). “Speech perception,” in *Foundations of Modern Auditory Theory* (Vol. II), edited by J. Tobias (Academic, New York), pp. 3–62.
- Studdert-Kennedy, M. (1974). “The perception of speech,” in *Current Trends in Linguistics*, edited by T. A. Sebeok (Mouton, The Hague), pp. 2349–2385.
- Sumbly, W. H., and Pollack, I. (1954). “Visual contribution to speech intelligibility in noise,” *J. Acoust. Soc. Am.* **26**, 212–215.
- Summerfield, Q. (1987). “Some preliminaries to a comprehensive account of audio-visual speech perception,” in *Hearing by Eye: The Psychology of Lip-Reading*, edited by B. Dodd and R. Campbell (Lawrence Erlbaum, Hillsdale, NJ), pp. 3–52.
- Van Tasell, D. J., Soli, S. D., Kirby, V. M., and Widin, G. P. (1987). “Speech waveform envelope cues for consonant recognition,” *J. Acoust. Soc. Am.* **82**, 1152–1161.
- Walden, B. E., Prosek, R. A., Montgomery, A. A., Scherr, C. K., and Jones, C. J. (1977). “Effect of training on the visual recognition of consonants,” *J. Speech Hear. Res.* **20**, 130–145.
- Watson, C. S., Qiu, W. W., Chamberlain, M. M., and Xiaofeng, L. (1996). “Auditory and visual speech perception: Confirmation of a modality-independent source of individual differences in speech recognition,” *J. Acoust. Soc. Am.* **100**, 1153–1162.
- Wertheimer, M. (1938). “Laws of organization in perceptual forms,” in *A Sourcebook of Gestalt Psychology*, edited by W. D. Ellis (Routledge & Kegan Paul, London), pp. 71–88.
- Zadeh, L. A. (1965). “Fuzzy sets,” *Information and Control* **8**, 338–353.