

Speech Perception in Perceivers With Hearing Loss: Synergy of Multiple Modalities

Dominic W. Massaro
Michael M. Cohen
Department of Psychology
University of California,
Santa Cruz

Although we would expect that the role of visible speech in multimodal speech perception would have the greatest relevance to individuals with hearing impairment, relatively few analytic studies have been done with these participants. Their adequate understanding of spoken language usually requires information from several modalities or other sources of information. The framework of the fuzzy logical model of perception (FLMP) is used to assess how individuals with hearing impairment evaluate and integrate multiple sources of information. Given this framework, a distinction can be made between *information* and *information processing*. Within this framework, we can ask what information differences and information-processing differences exist among individuals with normal hearing and those with hearing impairment. Four experimental studies from the literature are analyzed to address these questions. Test items are presented under both unimodal and bimodal conditions. Of central interest is the nature of the bimodal performance as a function of the unimodal performance. The findings show that, although information differences obviously exist across different populations, their information processing involved in pattern recognition appears to be the same and is well described by the FLMP.

KEY WORDS: speech perception, hearing loss, models of speech perception, multimodal perception, feature analysis

It is now generally accepted that persons are influenced by multiple sources of information in face-to-face communication. Two particularly powerful influences are the audible and visible consequences of speaking. We are influenced by both the sound and the sight of the speaker. Although this fact has probably been true since speech originated, only relatively recently did speech scientists become enthralled with bimodal speech perception. One convincing demonstration of the joint influence of the two modalities is to present auditory speech in noise (Sumby & Pollack, 1954). Perceptual accuracy improves when the perceiver also has sight of the speaker relative to the situation in which only the sound is available. An even more impressive demonstration is to experience conflicting audible and visible speech, such as an auditory /ba/ combined with a visual /da/ (McGurk & MacDonald, 1976). Perceivers in many instances perceive some other syllable, such as /va/ or /ða/, for the above combination (Massaro, 1998).

Although we would expect that the role of visible speech perception would have the greatest relevance to individuals with severe to profound hearing loss, relatively few analytic studies have been done with this population. Many individuals with severe and profound hearing

impairment may have some sound reception and measurable speech recognition ability with hearing aids or cochlear implants. In addition, it is well-known that visible speech alone is seldom sufficient for perceiving a message. In most situations, several different speech segments have equivalent facial and mouth movements. For example, the phonemes /b/, /p/, and /m/ are seen as equivalent and are said to belong to the same viseme class. Adequate understanding of spoken language in difficult situations for individuals with hearing loss thus requires information from several modalities. People with hearing loss accordingly offer a valuable population to study the integration of audible and visible information in speech perception.

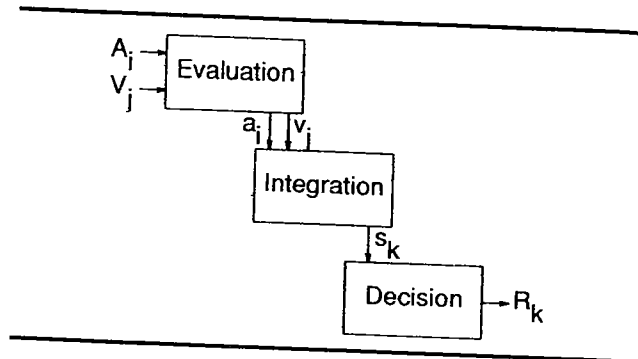
One goal of this paper is to advocate the study of bimodal speech perception in individuals with hearing loss. We will describe and analyze four comprehensive studies. These studies are chosen because their methodology is exemplary in terms of the stimulus items, the experimental conditions, and the data presentation. In both cases, a large selection of speech stimuli was tested, both unimodal and bimodal conditions were tested, and the complete confusion matrices either were published or made available to us. Our analysis goes beyond the previous ones, however, by utilizing the framework of the fuzzy logical model of perception (FLMP). We now provide a description of this framework.

Fuzzy Logical Model of Perception (FLMP)

A central assumption we make is that well-learned patterns are recognized in accordance with a general algorithm, regardless of the modality or particular nature of the patterns (Massaro, 1987). The FLMP consists of three operations: feature evaluation, feature integration, and decision. Continuously valued features are evaluated and integrated, and an identification decision is made on the basis of the relative goodness of match of the stimulus information with the relevant prototype descriptions. Figure 1 illustrates the three stages involved in pattern recognition.

Central to the FLMP are summary descriptions of the perceptual units of the language. These summary descriptions are called prototypes, and they include a conjunction of various properties called features. Soon after birth, infants quickly acquire knowledge about the meaningful segments of their language. This knowledge can be thought of as a set of features that characterize each segment. By features, we do not mean phonetic or phonological features but rather sensory primitives that inform the perceiver about a speech category. Speech perception is greatly determined by this knowledge of segments and their accompanying features. Iverson and

Figure 1. Schematic representation of the three processes involved in perceptual recognition. The three processes are shown to proceed left to right in time to illustrate their necessarily successive but overlapping processing. These processes make use of prototypes stored in long-term memory. The sources of information are represented by uppercase letters. Auditory information is represented by A_i and visual information by V_j . The evaluation process transforms these sources of information into psychological values (indicated by lowercase letters a_i and v_i). These sources are then integrated to give an overall degree of support, s_k , for each speech alternative k . The decision operation maps the outputs of integration into some response alternative, R_k . The response can take the form of a discrete decision or a rating of the degree to which the alternative is likely.



Kuhl (1995), for example, found that infants develop prototypical segments in the language but monkeys do not. Consistent with our cross-linguistic findings with adults, infants with different native languages have different prototypes (Massaro, 1998, Chapter 5). A prototype is a category, and the features of the prototype correspond to the prototypical values that an exemplar should have if it is a member of that category. The exact form of the representation of these properties is not known and may never be known. However, the memory representation must be compatible with the sensory representation resulting from the transduction of the audible and visible speech. Compatibility is necessary because the two representations must be related to one another. We assume that the functional prototypes in speech perception are open syllables (Massaro, 1972). To recognize the syllable /ba/, the perceiver must be able to relate the sensory information provided by the syllable itself to some memory of the category /ba/.

Prototypes become functional for the task at hand. In speech perception, for example, we might envision activation of all prototypes corresponding to the perceptual units of the language being spoken. For ease of exposition, consider a speech signal representing a single perceptual unit, such as the syllable /ba/. The sensory systems transduce the physical event and make available various features resulting from the sources of multimodal input. During the first operation in the model, the features are evaluated in terms of the prototypes in memory. For each feature and for each prototype,

featural evaluation provides information about the degree to which the feature in the speech signal matches the featural value of the prototype.

Given the necessarily large variety of features, it is necessary to have a common metric representing the degree of match of each feature (Summerfield, 1987; however, see Remez et al., 1994). The syllable /ba/, for example, might have visible featural information related to the closing of the lips and audible information corresponding to the rising second and third formant transitions. These two features must share a common metric if they eventually are going to be related to one another. That is, audible speech cannot be integrated with visible speech if both are not represented in common terms. To serve this purpose, fuzzy truth values (Zadeh, 1965) are used because they provide a natural representation of the degree of match. Fuzzy truth values lie between zero and one, corresponding to a proposition being completely false or completely true. The value .5 corresponds to something completely ambiguous, whereas .7 would be more true than false and so on. Fuzzy truth values, therefore, not only can represent continuous rather than just categorical information, they also can represent different kinds of information, such as audible and visible speech.

It should be noted that fuzzy truth values are not probabilities, although both lie between zero and one. To say that "A penguin is a bird to degree .85" is not the same as saying that "The probability that a penguin is a bird is .85." The former represents some measure of the degree to which the concept "penguin" matches the concept "bird," whereas the latter gives the probability that any given "penguin" exactly matches the concept "bird." Equivalent numerical values can correspond to different psychological representations. Although the FLMP makes the same predictions as Bayes theorem (Massaro, 1998, Chapter 4; Massaro & Friedman, 1990), the two formalizations are not equivalent psychological models.

Feature evaluation provides the degree to which each feature in the syllable matches the corresponding feature in each prototype in memory. The goal, of course, is to determine the overall goodness of match of each prototype with the syllable. All of the features contribute to this process, and the second operation of the model is called feature integration. That is, the features (actually the degrees of matches) corresponding to each prototype are combined (or conjoined in logical terms). The outcome of feature integration consists of the degree to which each prototype matches the syllable.

The third operation during recognition processing is decision. During this stage, the merit of each relevant prototype is evaluated relative to the sum of the merits of the other relevant prototypes. This decision operation is modeled after Luce's (1959) choice rule, called a relative goodness rule (RGR) by Massaro and Friedman

(1990). In Luce's choice axiom, the choice objects are represented by scale values (analogous to Case V of Thurstone, 1927). The choice axiom holds if and only if (a) the RGR holds, (b) the scale value representing an object does not change with changes in the response alternatives used in the choice task, and (c) the response alternatives defined as irrelevant do not enter into the RGR. The RGR stipulates that it is the relative goodness of match of the sensory input with a memory alternative that is critical, not its absolute goodness of match (see Massaro, 1998, pp. 264–268). This relative goodness of match gives the proportion of times the syllable is identified as an instance of the prototype. The relative goodness of match can also be mapped into a rating judgment indicating the degree to which the syllable matches the category. An important prediction of the model is that one feature has its greatest effect when a second feature is at its most ambiguous level. Thus, the most informative feature has the greatest impact on the judgment.

As described previously, prototypes are central to the description given by the FLMP. For illustrative purposes, consider the consonant-vowel (CV) syllables /ba/ and /da/. Although there are many different features representing each speech category, we formalize the model in terms of just two features, one auditory and one visual. We use the onsets of the second (F2) and third (F3) formants as the auditory features and the degree of opening of the lips at the onset of the syllable as the visual feature. If these were the only features, the prototype for /da/ would be

/da/: Slightly falling F2–F3 and Open lips

The prototype for /ba/ would be defined in an analogous fashion:

/ba/: Rising F2–F3 and Closed lips

Given a prototype's independent specifications for the auditory and visual sources, the value of one source cannot change the value of the other source. The integration of the features defining each prototype is evaluated according to the product of the feature values. If a_i represents the degree to which the auditory stimulus A_i supports the alternative /da/ (that is, has Slightly falling F2–F3) and v_j represents the degree to which the visual stimulus V_j supports the alternative /da/ (that is, has Open lips), then the outcome of prototype matching for /da/ would be given by the product of a_i and v_j .

$$s(/da/) = a_i v_j \quad (1)$$

where $s(/da/)$ is the overall degree of support, s_k , for $k = /da/$. With just two alternatives, /da/ and /ba/, we can make the simplifying assumption that the degree to which the audible speech supports the alternative /ba/ is $1 - a_i$. Thus, the support for the alternative /ba/ would be

$$s(/ba/) = (1 - a_i) (1 - v_j). \quad (2)$$

The third operation that is necessary before a behavioral judgment is made is decision, which follows the RGR. Implementing the RGR, it is predicted that the probability of a /da/ response given A_i and V_j is equal to the total support for /da/ divided by the sum of the support for all relevant alternatives, in this case $s(\text{da}/)$ and $s(\text{ba}/)$.

$$P(\text{da}/ | A_i, V_j) = \frac{s(\text{da}/)}{s(\text{da}/) + s(\text{ba}/)}$$

$$= \frac{a_i v_j}{a_i v_j + (1 - a_i)(1 - v_j)} \quad (3)$$

Given this framework, we are able to make a distinction between *information* and *information processing*. The sources of information from the auditory and visual channels make contact with the perceiver at the evaluation stage of processing. The reduction in uncertainty effected by each source is defined as information. In the fit of the FLMP, for example, the parameter values (a_i and v_j) indicating the degree of support from each modality correspond to information. These parameter values represent how informative each source of information is. Information processing refers to how the sources of information are processed. In the FLMP, this processing is described by the evaluation, integration, and decision stages. Thus we ask how well the FLMP describes performance.

Within this framework, we can ask what information differences exist between individuals with and without hearing impairment. Perceivers with hearing impairment obviously have less auditory information, but do they also differ in terms of visual information? In addition, do the two groups of perceivers process these information sources differently? More specifically, does the integration process differ for the two groups? It is possible that the groups differ with respect to the efficiency of integrating the audible and visible speech. Grant, Walden, and Seitz (1998) found that a measure of integration fell below the maximum that could be expected from an ideal combination of the audible and visible speech. We can ask similar questions within these groups. For example, does the integration process work the same way regardless of the degree of hearing impairment. By comparing individuals using hearing aids to those with cochlear implants, we can also address information and information-processing questions in terms of the nature of the assistive device. For example, it is conceivable that integration of the two modalities is more difficult with cochlear implants than with hearing aids. We now address these questions and begin with an analysis of an early study by Erber (1972).

Erber Study

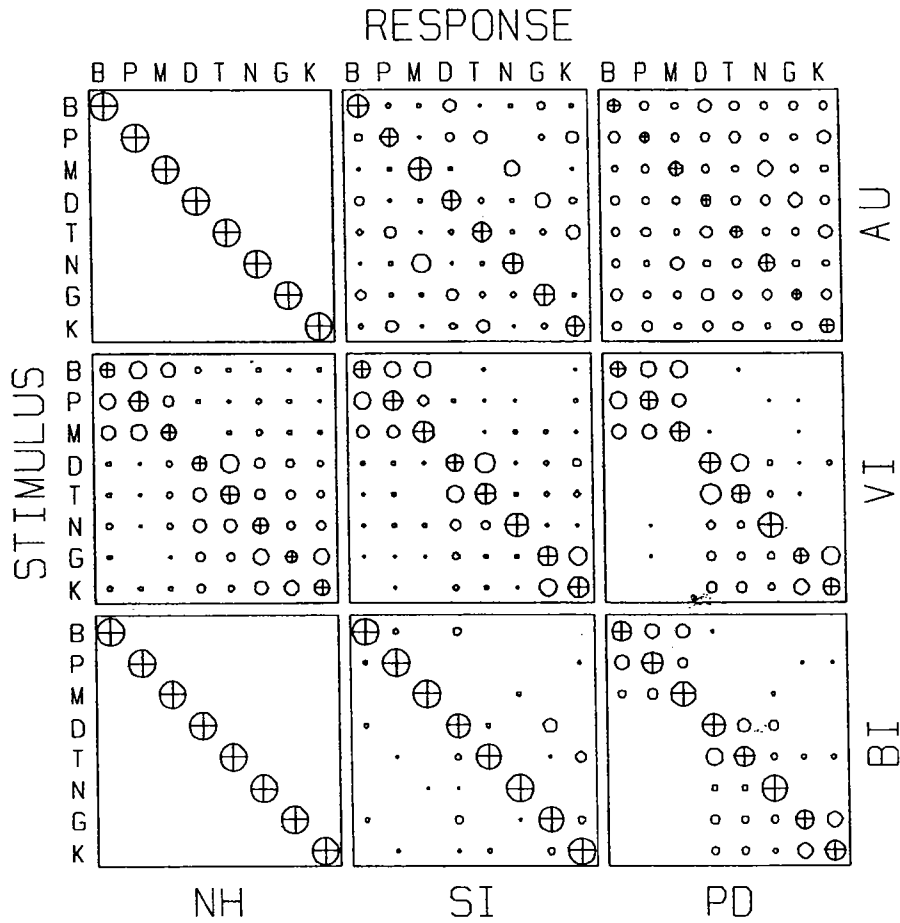
Erber (1972) tested three populations of children (adolescents and young teenagers between 9 and 15 years old): normal hearing (NH), severely impaired (SI),

and profoundly hearing-impaired (PI). The SI children had a hearing loss between 75 and 90 dB, and the PI had a hearing loss greater than or equal to 95 dB (for frequencies between 500 and 2000 Hz). All of the children with impaired hearing had sustained their loss before the acquisition of speech and language. They also had extensive experience with hearing aids and had at least 4 years of experience with the oral method of multimodal speech perception. The children used their hearing-assisted devices during the test. None of the children with normal hearing had any training in speechreading, whereas the children with hearing impairment had attended oral schools for the deaf with formal training in speechreading. There were 5 children in each of the three groups. The experimental test consisted of a videotape of the eight consonants /b, d, g, k, m, n, p, t/ spoken in a bisyllabic context /aCa/, where C refers to one of the eight consonants. The children had to respond with one of these eight alternatives. The test was presented under auditory, visual, and bimodal conditions. Each child was tested 45 times on each of the consonants under each of the three presentation conditions—an impressive number of observations. It is important to note that the talker's face was intensely illuminated so that the inside of the oral cavity was visible.

The overall performance differed greatly across the three groups. For the auditory condition, accuracy was .99, .50, and .21 for groups NH, SI, and PI respectively. For the visual condition, the scores were .32, .49, and .45. The bimodal scores were .99, .88, and .61. These scores show the overall differences among the groups and that group NH did not benefit from the visible speech given that they were already perfect in the auditory condition. The results from the other two groups, however, reflect the value that two ambiguous sources of information can have relative to either one alone.

Erber (1972) also presented the confusion matrices in tabular form. The results for the three groups under the three presentation conditions are shown in Figure 2 in the form of confusions matrices. The response probability is proportional to the area of the circles, and those on the main diagonal have a cross to indicate a correct response. Although the data appear to be somewhat complex, there are several obvious findings. First, the results for the NH group replicate what we already know: auditory speech is more informative than visible speech. Persons in the NH group were perfect in the auditory (AU) and bimodal (BI) conditions. Their errors in the visible (VI) speech condition reflect their inability to distinguish segments from within a viseme class. A viseme is a visible speech category, named analogously to the phoneme. A viseme usually contains several phonemes that tend to be indistinguishable from one another. Presented with the unimodal auditory speech, the SI group made many errors and the PI group made even more,

Figure 2. Confusion matrix given the auditory (AU), visual (VI), and bimodal (BI) conditions for three populations of children (adolescents and young teenagers): normal hearing (NH), severely impaired (SI), and profoundly deaf (PD). The response probability is proportional to the area of the circles, and those on the main diagonal have a cross to indicate a correct response. The results should be interpreted as both the observations and the predictions of the modality-analysis implementation of the FLMP because they were essentially equivalent to one another; the small differences are not noticeable in this type of plot.



performing near chance for some of the test items. Based on the performance on the unimodal visual test, however, these two groups appear to speechread somewhat better than the normal population—a result that is not always found.

An important outcome for our purposes is the performance gain the two hearing-impaired groups show in the bimodal condition relative to either of the unimodal conditions. This outcome reflects the synergy of multiple modalities in speech perception: two ambiguous sources of information can be combined to produce an unambiguous outcome. The overall bimodal performance of group SI is much better than that for group PI because the children in group SI supposedly have much more auditory information, as witnessed by their better unimodal auditory performance. We will now explore several implementations of the FLMP and a contrasting additive model in the analyses of these data.

Modality-Analysis Implementation

In this implementation, the FLMP is tested against results with multiple response alternatives in the same manner as with just two response alternatives. It is assumed that each modality supports each alternative to some degree, as described in the rationale for Equations 1 and 2. With more than two alternatives, it is necessary to estimate a unique parameter to represent the degree to which each source of information supports each alternative. We use aB_i to represent the degree to which the audible speech supports the alternative /ba/. The term vP_j would represent the degree to which the visible speech supports the alternative /pa/, and so on for the other response alternatives. Given both audible and visible speech, the total support for the alternative /ba/, $s(/ba/)$, would be

$$s(/ba/) = aB_i vB_j \quad (4)$$

and so on for the other test conditions and the other alternatives.

As in the case of just two alternatives, the probability of a particular categorization is assumed to be equal to the relative goodness-of-match

$$P(\text{ba}/|A_i, V_j) = \frac{s(\text{ba})}{\sum_r s(r)} \quad (5)$$

where $s(r)$ corresponds to the goodness of match for alternative r , and $\sum_r s(r)$ corresponds to the sum of the goodness of match values of all possible response alternatives.

With eight stimulus-response alternatives in the Erber study, each test stimulus provides different degrees of support for eight response alternatives. Given that we cannot determine these degrees of support before the test is actually carried out, they must be estimated from the actual judgments of the observers. A free parameter is needed to describe how much each source supports each of the test alternatives. In Erber's experiment, it is necessary to estimate eight free parameters for each of the eight test stimuli in each modality. Thus, 64 free parameters are required for the auditory modality and 64 for the visual modality. These constraints make it apparent that observations in the auditory, visual, and bimodal conditions are necessary in order to provide a test of the model. If only two of these three conditions are tested, then there would be as many free parameters as independent data points—an undesirable state of affairs. With all three conditions, on the other hand, we are able to test the model by predicting 3×64 data points with 2×64 free parameters. It has been claimed that the three conditions allow a parameter-free test of models (Braida, 1991; Grant & Walden, 1995). However, this type of analysis rests on the assumption that the unimodal conditions are noise-free estimates of the parameters used to predict the bimodal condition. A model is given its best chance by estimating the free parameters from all of the results rather than simply by predicting one of the three conditions based on the results observed in the other two conditions. (See Massaro, 1998, Chapters 10 and 11, for a discussion of the various factors in model testing, parameter estimation, and evaluating the goodness-of-fit of models.)

The goodness-of-fit of a model is given by the root mean squared deviation (RMSD) between the predicted and observed values. The best fit is that which gives the minimal RMSD. The RMSD is computed by (a) squaring the difference between each predicted and observed value, (b) summing across all conditions, (c) taking the mean, and (d) taking the square root of this mean. The RMSD can be thought of as a standard deviation of the differences between the 192 predicted and

observed values. The RMSD would increase as the differences increase. The smaller the RMSD value, the better the fit of the model.

The quantitative predictions of each model are determined by using the program STEPIT (Chandler, 1969). The model is represented to the program in terms of a set of prediction equations and a set of unknown parameters. By iteratively adjusting the parameters of the model, the program minimizes the RMSD. The outcome of the program STEPIT is a set of parameter values that, when put into the model, come closest to predicting the observed results. Thus, STEPIT maximizes the accuracy of the description of each model.

The fit of this model requires 8 a_i and 8 v_j parameters for each of the 8 response alternatives, for a total of 128 free parameters. A unique set of parameters is estimated for each of the three groups. The fit of the FLMP was .0009, .0121, and .0120 for the NH, SI, and PI groups, respectively. The predicted values are not plotted in Figure 2 because they would not be noticeably different from the observed values. Erber's results also reveal a strong complementarity between the audible and visible modalities in speech, which is discussed more fully in Massaro, 1998 (Chapter 14). Thus, the model is able to provide a very good fit when it is assumed that there are two modality-specific sources of information supporting the perceptual judgment.

It is possible that the FLMP assumption of multiplicative integration of the auditory and visible speech is not accurate. To test this, an additive model of perception (AMP) was fit to the observed data in the same manner as the FLMP. The difference between these two models is simply the conjunction rule: multiplicative versus additive.

$$s(\text{da}) = a_i + v_j \quad (5)$$

where $s(\text{da})$ is the overall degree of support, s_k , for $k = \text{da}$. All other aspects of the AMP are equivalent to the FLMP. The AMP is also mathematically equivalent to a single-channel model and a categorical model of perception (Massaro, 1998, Chapter 2). Thus a test of the additive model also provides a test of these models. The AMP gave RMSDs of .0019, .0598, and .0320 for groups NH, SI, and PI, respectively. Although the additive model provided an equally good description of the normal hearing children, it performed much poorer for the other two groups. Our interpretation of this result is that group NH with no noise in the auditory condition cannot be taken to provide a test of different models of integration. Their essentially perfect performance on the unimodal auditory condition makes the question of integration moot. For the two hearing-impaired groups, integration is critical because neither source provides adequate information. In these groups, the multiplicative integration of the FLMP proves superior. We can

conclude that the FLMP produced a better fit to the data than the AMP.

Feature Analysis Implementation

The model test we have presented in the previous section makes no assumptions about the psychophysical relationship among the different test items. A unique parameter is estimated for each possible pairing. For example, a unique parameter is estimated to represent the amount of support a visual /b/ provides for the response alternative /d/. To test the psychological reality of various linguistic features and to reduce the number of free parameters, we now articulate the FLMP in terms of audible and visible support for these features. This formulation has the potential to save a large number of features, because it is assumed that a given feature in a given modality has the same impact regardless of what segment it is in. Recall that the eight consonants were /b, d, g, k, m, n, p, t/. Following the tradition begun with Miller and Nicely (1955), we can define these eight segments by three features: voicing, nasality, and place. Table 1 lists the feature representation for 17 consonants that include the 8 used in this experiment. The other consonants in Table 1 are included because they were used in some of the other studies that we will analyze. The features duration and frication are not necessary to distinguish among these 8 auditory consonants. Three of the segments were voiced, two were nasalized, and three had a front, three a middle, and two a back place of articulation.

As noted in the development of the FLMP, we assume that features are simply sensory primitives that distinguish speech categories. Although the features used in the following tests are chosen to be equivalent to the linguistic features first used by Miller and Nicely (1955), they should be thought of as simply convenient labels for the underlying sensory features. Thus, for example, the auditory feature for place would not necessarily be equivalent to the parameter value for the visible feature for place. Thus the features at the evaluation stage are not linguistic, but perceptual.

It is important to stress that the feature values for one modality should be independent of the feature values for another modality. For example, we would expect that voicing and nasality would have informative feature values for auditory speech and relatively neutral feature values for visible speech. The place feature, on the other hand, would give relatively informative values for visible speech.

Thus, each of the eight syllables would be described by the conjunction of three features for unimodal speech and the conjunction of six features for bimodal speech. Even though each feature is defined as a specific value or its complement (e.g., voiced or voiceless), its influence in the perception of visible speech is represented by a value between 0 and 1. The parameter value for the feature indicates the amount of influence that feature has. Therefore, if the /ma/ and /na/ prototypes are each expected to have a nasal feature and the calculated parameter value for this feature is .90, then the nasal feature is highly functional in the expected direction. Alternatively, if the calculated parameter value for the nasal feature is .50, then the conclusion would be that the nasal feature is not functional at all. Because of the definition of negation as 1 minus the feature value, a feature value of .5 would give the same degree of support for a phoneme that has the feature as it should for a phoneme that doesn't have the feature. Finally, if the calculated parameter value is .20, then the nasal feature is functional but in the opposite of the expected direction. Finally, it should be noted that the features are not marked in this formulation; absence of nasality is as informative as presence of nasality. Thus if a nasal stimulus supports nasal response alternatives to degree .9, then a non-nasal stimulus also supports a non-nasal alternative to degree .9.

The overall match of a test stimulus to each syllable prototype was calculated by combining the feature matches according to the assumptions of the FLMP. These constraints dictate that (a) the features are the sources of information that are evaluated independently of one another, and (b) the features are integrated multiplicatively (conjoined) to give the overall degree of support for a syllable alternative. Thus, the overall degree of

Table 1. Feature set describing the 17 consonants used in the studies. The symbol "+" means voiced, nasal, fricative, or short duration. The symbols 1, 2, and 3 correspond to front, middle, and back articulations.

Feature	Phoneme																
	/b/	/p/	/m/	/d/	/t/	/n/	/g/	/k/	/ŋ/	/v/	/f/	/s/	/l/	/r/	/dʒ/	/ʃ/	/z/
Voicing	+	-	+	+	-	+	+	-	+	+	-	-	+	+	+	-	+
Nasal	-	-	+	-	-	+	-	-	+	-	-	-	-	-	-	-	-
Fricative	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	+	+
Place	1	1	1	2	2	2	3	3	3	1	1	2	2	3	1	2	2
Duration	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-

support for /ba/, s(/ba/), given the presentation of a /ba/ syllable, is

$$s(/ba/|/ba/) = a_v \times a_n \times a_p \times v_v \times v_n \times v_p, \quad (7)$$

where each feature value indexes a match between the feature in the stimulus and the corresponding feature in the /ba/ prototype. The features a_v correspond to auditory voicing, v_n to visual nasality, and so on. A mismatch between the feature in the stimulus and the corresponding feature in the prototype would be indexed by $(1 - f_i)$, where f_i corresponds to the modality's feature value. Thus, the support for the /ka/ prototype given presentation of a /ba/ syllable, is

$$s(/ka/|/ba/) = (1 - a_v) \times a_n \times (1 - a_p) \times (1 - v_v) \times v_n \times (1 - v_p), \quad (8)$$

where $(1 - f_i)$ indexes a mismatch between the feature in the stimulus and the corresponding feature in the /ka/ prototype.

After the overall degree of support for each syllable is calculated, the stimulus is categorized according to the RGR, which states that the relative probability of

choosing an alternative is the goodness of match of that alternative divided by the sum of the goodness of match of all alternatives. Thus, this model implementation parallels the previous one in all aspects except in terms of the featural description of the stimulus and response alternatives. The FLMP can thus be tested against the confusion matrix by estimating the amount of information in each feature and the featural correspondence between the stimulus and response prototypes. Thus, three parameters are necessary to describe the auditory information, and the same number are necessary to describe the visual. The feature-analysis FLMP was tested against the confusion matrices of the three groups of children. Figures 3, 4, and 5 give the fit of the feature-analysis model to the three groups, respectively. The fit of this model to three groups of observers gave an RMSD of .0328, .0409, and .0514 for the NH, SI, and PI groups, respectively. Although the RMSDs are much greater than those in the modality-analysis implementation, the number of free parameters has been reduced from 128 to just 6.

Figure 3. Observed (left panels) and predicted (right two panels) confusion matrices for the normal hearing (NH) children in the Erber study. The area of the circle is proportional to response probability. The predictions are for the modality-analysis (MA) and the feature-analysis (FA) implementation of the FLMP.

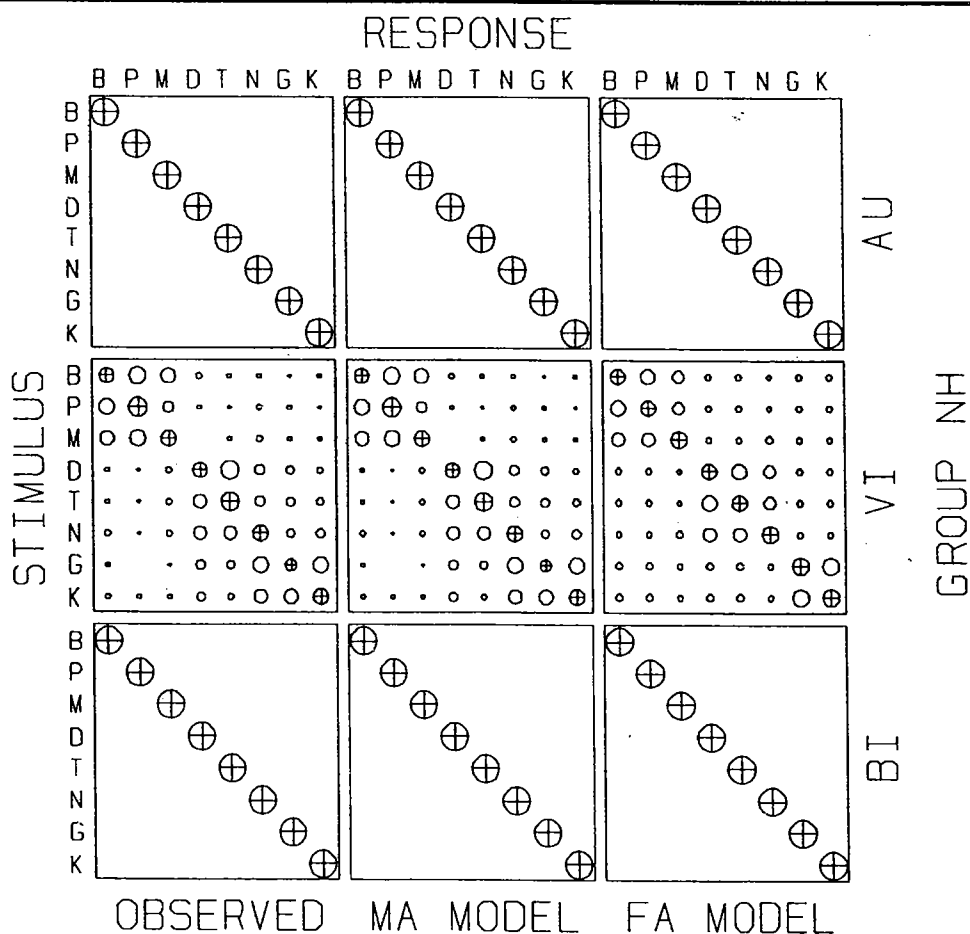


Figure 4. Observed (left panels) and predicted (right two panels) confusion matrices for the severely impaired (SI) children in the Erber study. The area of the circle is proportional to response probability. The predictions are for the modality-analysis (MA) and the feature-analysis (FA) implementation of the FLMP.

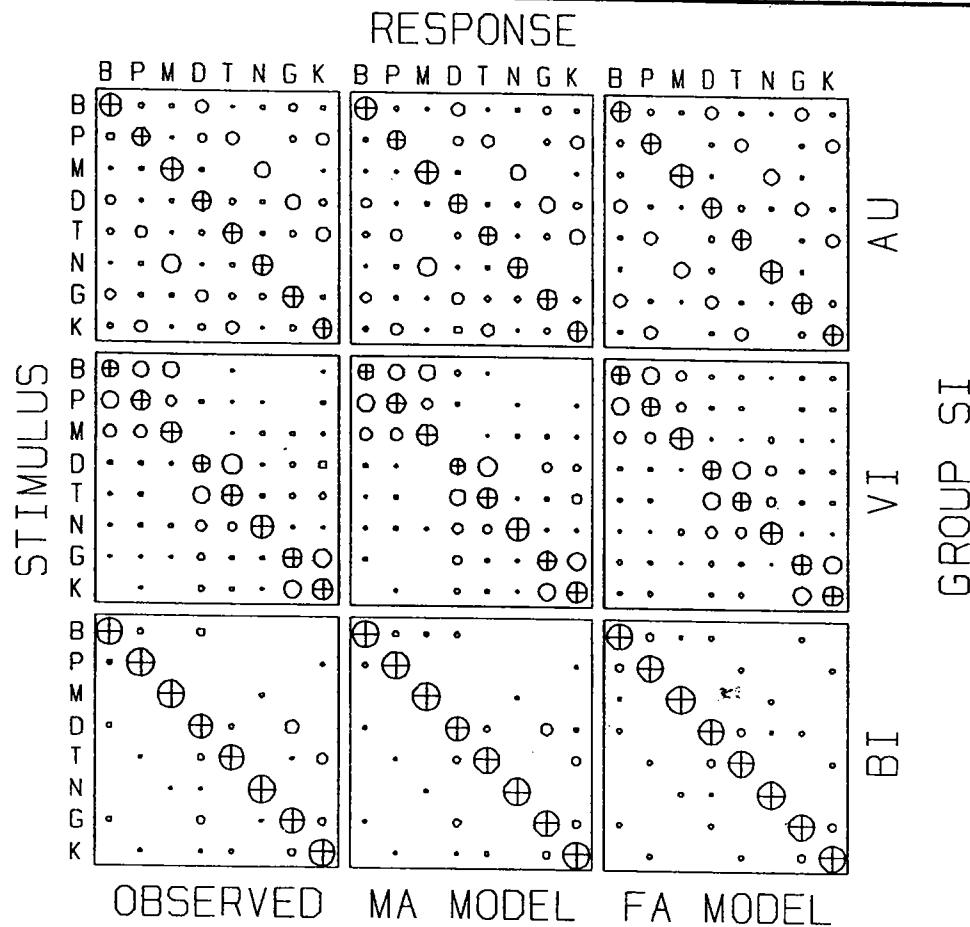


Table 2 gives the best fitting parameters of this model. A parameter value of near .5 means that very little information was transmitted about this feature. As expected, little or no information was transmitted by the face concerning voicing. Information about voicing was well conveyed by the auditory channel except for the children in group PI. The parameter values for place of articulation are reasonable. Visible speech transmits good information about place, whereas the information value for place from the auditory modality declines with hearing impairment. Finally, one might ask why visible speech conveys information about nasality,

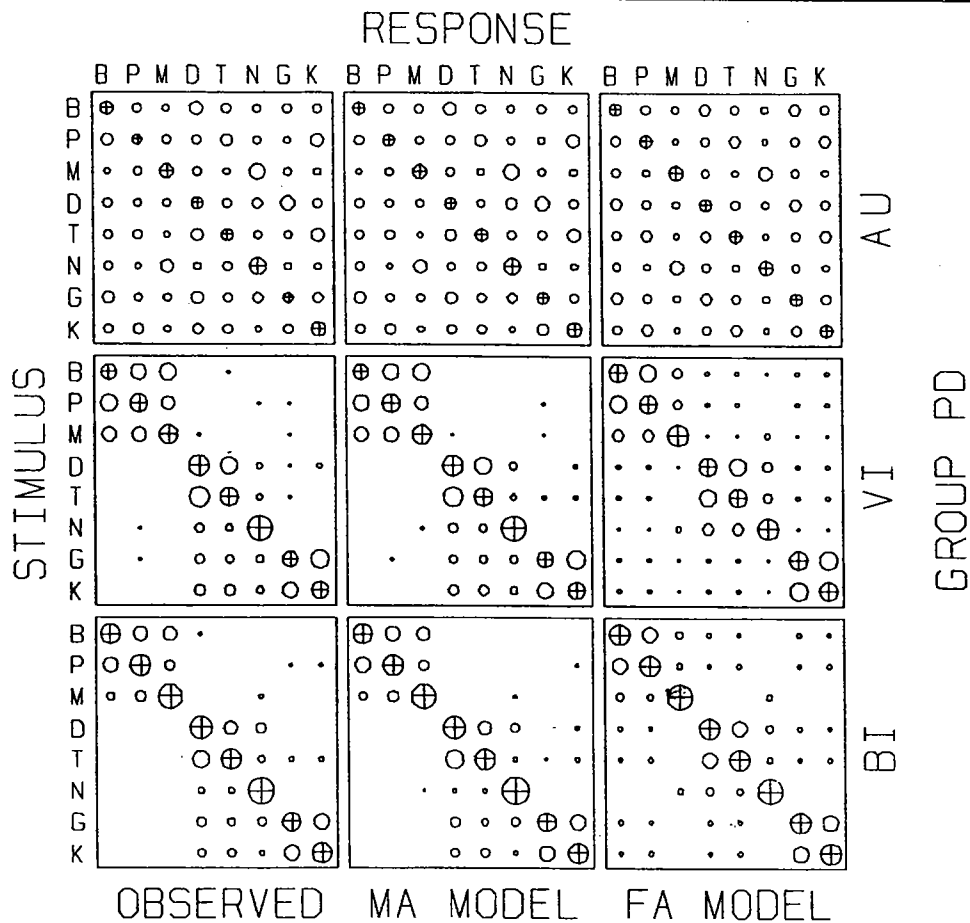
Table 2. Best fitting parameters for FA model of the Erber study.

Group	Visual			Auditory		
	VO	NA	PL	VO	NA	PL
NH	.5014	.6013	.8436	.9994	.9991	.9990
Si	.5565	.7803	.9376	.9022	.9455	.6869
PI	.5198	.7694	.9358	.6181	.7106	.5381

especially for the two groups of hearing-impaired children. One possibility is that, because Erber's participants knew that the velar nasal was not tested, information about place also provided some information about nasality. If the visible speech supported a velar place of articulation, then the perceiver would have known that it was also non-nasal.

It is possible that the FLMP assumption of multiplicative feature integration is not accurate. To test this, an additive model of perception (AMP) was fit to the observed data in the same manner as the FLMP. The difference between these two models is simply the conjunction rule: multiplicative versus additive. The fit of this model to three groups of observers gave an RMSD of .2420, .1765, and .1279 for the NH, SI, and PI groups, respectively. These fits are roughly 3 to 7 times worse than those given by the multiplicative combination of features in the FLMP. Given that only group data were available to test the model, no inferential statistics are possible. Given previous work with individual fits and statistical tests, however, the large differences between

Figure 5. Observed (left panels) and predicted (right two panels) confusion matrices for the profoundly deaf (PD) children in the Erber study. The area of the circle is proportional to response probability. The predictions are for the modality-analysis (MA) and the feature-analysis (FA) implementation of the FLMP.



the RMSD values support the conclusion that the FLMP produced a better fit to the data than the AMP.

Dowell et al. Study

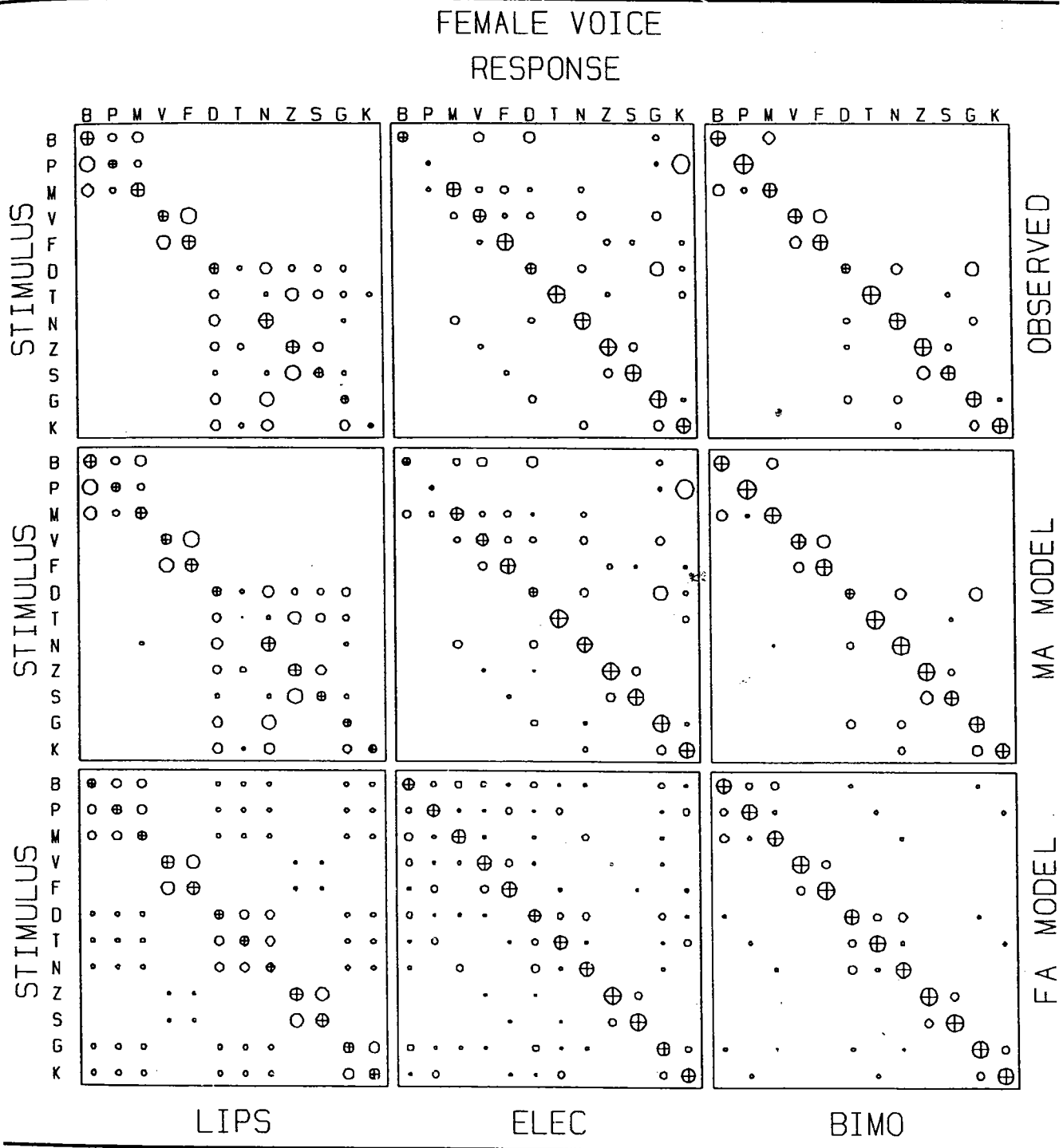
In this study, a patient with a multiple-channel cochlear implant was tested with just electrical stimulation, just lipreading, and both of these sources of information (Dowell et al., 1982). Twelve consonants were presented in an /aCa/ context. Twenty observations were made on each consonant spoken by a female speaker in one test and a male speaker in the other. The results are reported as a 12 × 12 confusion matrix under each of the three presentation conditions. The modality-analysis FLMP was applied to the results with the female and male speaker separately and gave RMSD values of .0263 and .0247, respectively (see Figures 6 and 7). The feature-analysis FLMP required five features and, therefore, reduced the number of free parameters from 288 to 10 in the prediction of the 432 data points. For this fit, the RMSD values increased to

.1061 and .0923, respectively. Table 3 gives the best fitting parameter values for the five linguistic features used in the model fit. These parameter values can be interpreted in the same manner as reported for the Erber fit. The model is capable of describing the integration of lipread information with electrical stimulation to the cochlea in the same manner as with normal hearing.

Agelfors Study

It is usually the case that neither hearing aids (HA) nor cochlear implants (CI) provide a sufficiently rich information source for perfect perception of speech in difficult situations. We also know too well that visible speech does not transmit the complete linguistic message. The synergy between two (degraded) channels, however, offers the potential of a robust communication environment for individuals with hearing loss who use one of these two assistive devices. Solid evidence for this conclusion comes from a study by Agelfors (1996). She

Figure 6. Observed (top panel) and predicted (bottom panels) confusion matrices for the female speaker in the Dowell et al. study. The predictions are for the modality-analysis (MA) and the feature-analysis (FA) implementation of the FLMP.

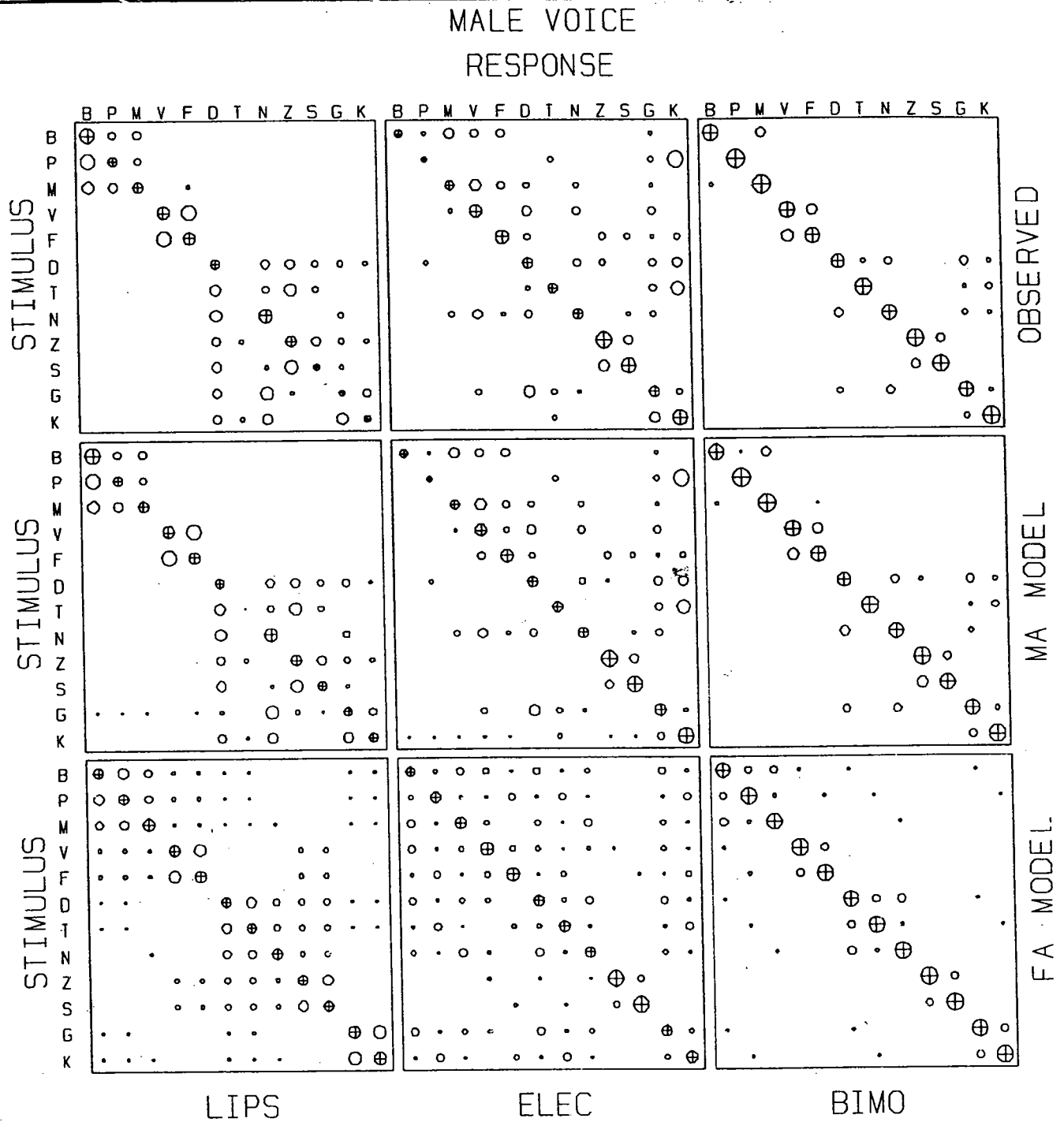


compared persons using HA and CI in several speech tests under auditory, visual, and bimodal presentations. One test involved the identification of 16 Swedish consonants presented in an /aCa/ context preceded by a carrier phase. The 16 consonants were /b, p, m, d, t, n, g, k, ʃ, v, f, s, l, r, d₃, f/. A videotape was made with four repetitions of each syllable presented in a random order. The

auditory level was adjusted by each participant to provide a comfortable listening level. The loudspeaker was turned off for the visual presentation.

According to the FLMP, there should be a super-additive effect of the binodal presentation relative to the unimodal conditions. The superadditivity results

Figure 7. Observed (top panel) and predicted (bottom panels) confusion matrices for the male speaker in the Dowell et al. study. The predictions are for the modality-analysis (MA) and the feature-analysis (FA) implementation of the FLMP.



from both complementarity and an optimal integration algorithm (Massaro, 1998, Chapter 14). The Agelfors study allows an answer to several additional questions beyond a test of the FLMP. First, it is possible to ask whether the gain with bimodal presentation is equivalent for HA and CI. Second, given that both groups were

split into subgroups with relatively good and relatively poor auditory sensitivity, we can ask whether the synergy of bimodal speech perception predicted by the FLMP holds for both of these subgroups. For the HA group, there were 12 participants with better hearing (HA+) and 3 with poorer hearing (HA-). For the CI group, there

Table 3. Best fitting parameters for FA model of the Dowell et al. study.

Speaker	Visual					Auditory				
	VO	NA	FR	PL	DU	VO	NA	FR	PL	DU
Female	.4921	.4915	.9844	.7990	.7852	.8166	.7509	.8160	.8036	.8905
Male	.4793	.6608	.8817	.9365	.2576	.8328	.6572	.7450	.6634	.9018

Table 4. Best fitting parameters for FA model of the Agelfors study.

Group	Visual					Auditory				
	VO	NA	FR	PL	DU	VO	NA	FR	PL	DU
CI+	.5712	.5552	.9842	.8828	.8039	.9337	.7584	.6672	.8265	.8522
CI-	.5824	.5791	.9999	.8906	.7816	.8231	.7580	.6141	.6276	.6329
HA+	.5839	.5111	.9913	.8985	.7720	.9148	.9026	.6759	.7923	.9528
HA-	.4960	.5490	.9925	.8820	.8297	.9475	.7439	.5470	.6305	.6158

were 8 participants with better auditory recognition (CI+) and 7 with poorer auditory recognition (CI-).

Parallel to the Erber study, we conducted two different model tests. The modality-analysis FLMP can be tested against the confusion matrix by estimating the amount of support that a modality-specific syllable presentation provides for each of the 16 consonants. Thus, (16 × 16) 256 parameters are necessary to describe the auditory information, and the same number is necessary to describe the visual information, for a total of 512. Given the three confusion matrices in each condition, there is a total of (3 × 256) 768 independent data points. As in the Erber and Dowell et al. studies, the ratio of data points to free parameters is thus 3 to 2.

Figures 8–11 give the observed and predicted results for the four subgroups of participants. As can be seen in the confusion matrices, superadditivity was obtained in all conditions. Furthermore, the FLMP gave a good description of each subgroup. The RMSDs were .0169, .0142, .0159, and .0207 for the CI+, CI-, HA+, and HI-, respectively.

As in the analysis of the Erber study, we tested the additive model. This model is identical in all respects to the FLMP except that an additive rather than a multiplicative integration is used. The additive model performed much more poorly than the FLMP, with RMSDs of .1201, .1079, .1276, and .1157, respectively, for groups CI+, CI-, HA+, and HA-. Thus, although the additive model had an additional free weight parameter, the goodness-of-fit was about six to eight times poorer than the fit of the FLMP.

We also carried out featural analyses, as was previously described for the Erber study. Given the larger set of syllables in the Agelfors study, the features of duration and frication were added to the feature set. Table 1 lists the feature representation for the 16 consonants.

The RMSDs were .0902, .0749, .0844, and .1084 for the CI+, CI-, HA+, and HI-, respectively. Although these values are about five times greater than the modality-analysis FLMP, the number of free parameters has been reduced from 512 to just 10. Table 4 gives the best fitting parameters of this model.

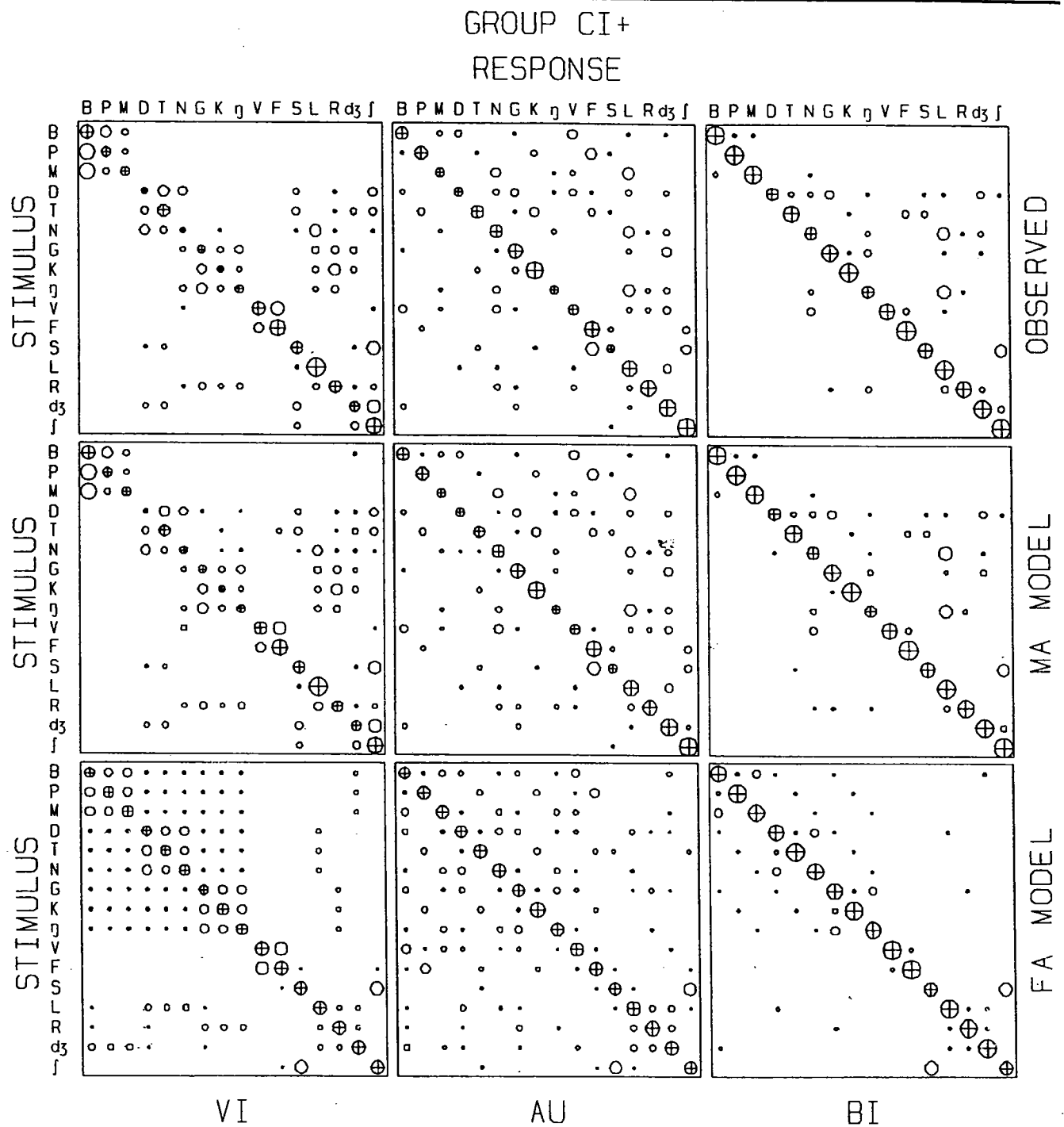
As in the other analyses, we tested an additive model. This model is identical in all respects to the feature-implementation FLMP except that an additive rather than a multiplicative integration is used. The additive model performed much more poorly than the FLMP, with RMSDs of .1533, .1275, .1579, and .1526, respectively, for groups CI+, CI-, HA+, and HA-. These values are much larger than those for the multiplicative integration, even though the same number of free parameters are used. Thus we conclude that the multiplicative integration given by the FLMP provides the best account of the confusion matrices.

Hearing Impairment in Older Adults

We also describe a study that employs a somewhat different methodology, but is equally analyzable within the context of the FLMP. Walden, Montgomery, Prosek, and Hawkins (1990) provide some comprehensive results on adults who became hearing-impaired with aging. These observers had a bilateral hearing loss predominantly in the high frequencies. They did not wear hearing aids during the experiment. The test items were synthetic auditory speech syllables along a 14-step /ba/-/da/-/ga/ continuum. The auditory syllables were presented either alone or paired with a visual /ba/ or a visual /ga/. The participants responded with /ba/, /da/, or /ga/.

Walden et al. (1990) presented the results in graphical form, which we scanned electronically to give numeric values. The average observed results are shown

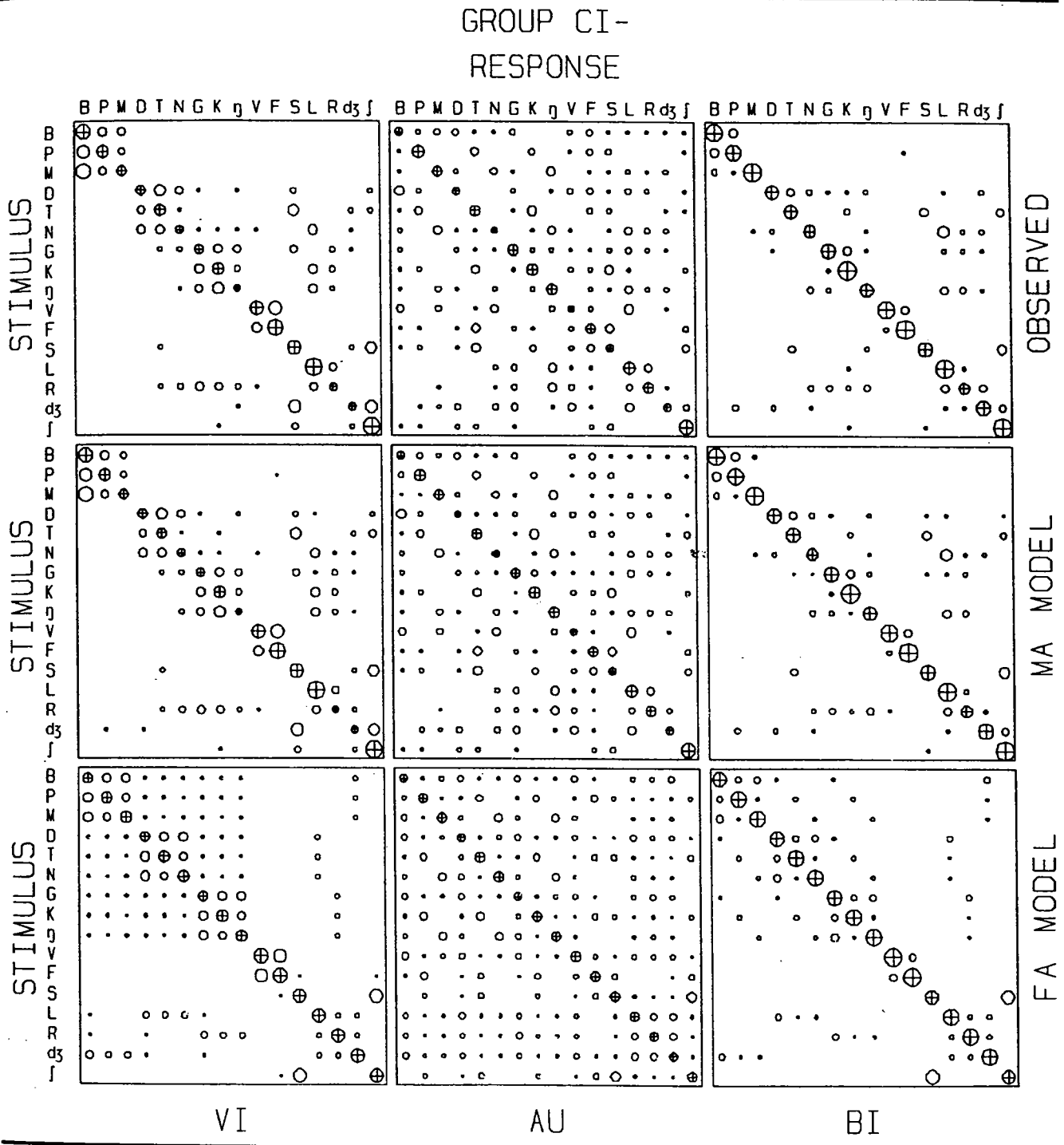
Figure 8. Observed (top panel) and predicted (bottom panels) confusion matrices for the eight observers with cochlear implants and relatively good auditory recognition in the Agelfors study. The predictions are for the modality-analysis (MA) and the feature-analysis (FA) implementation of the FLMP.



in Figure 12 together with the predictions of the FLMP. As can be seen in the figure, both sources of information had a large impact on performance. Although more complicated than our standard two response-task (Massaro, 1998), the results are easily understood. Although there is a very large impact of visible speech, the auditory speech

had an important and orderly influence. Changes along the auditory continuum had the expected effect. Stimuli at the /ba/ end of the auditory continuum were sometimes called /da/ and /ga/, indicating that auditory /ba/ was not as robust as auditory /da/ or /ga/. This difference seems responsible for the finding that the visual effect

Figure 9. Observed (top panel) and predicted (bottom panels) confusion matrices for the eight observers with cochlear implants and relatively poor auditory recognition in the Agelfors study. The predictions are for the modality-analysis (MA) and the feature-analysis (FA) implementation of the FLMP.

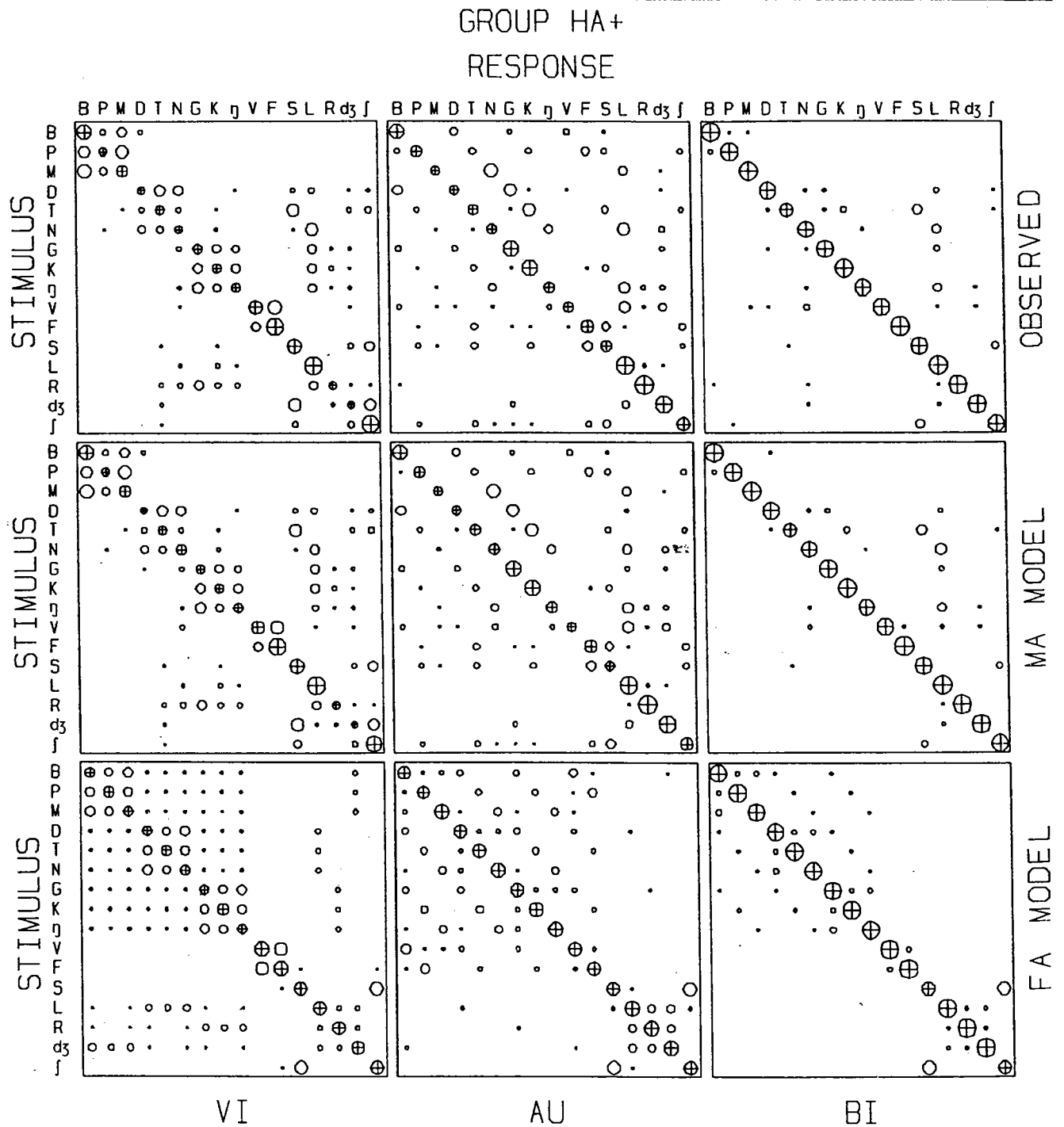


was larger at the /ba/ end of the auditory continuum. A visual /ba/ almost always produced /ba/ judgments for these auditory stimuli, and a visual /ga/ eliminated /ba/ judgments. A visual /ba/ gave a large number of /ba/ judgments, which tapered off across the continuum from /ba/ to /da/ to /ga/. Pairing a visual /ga/ with the auditory syllables

produced both /da/ and /ga/ judgments and very few /ba/ judgments.

Only the modality-analysis FLMP is appropriate for the fit of the Walden et al. results. Their study differs from the previous two by creating a continuum of ambiguous stimuli between just a few alternatives. The

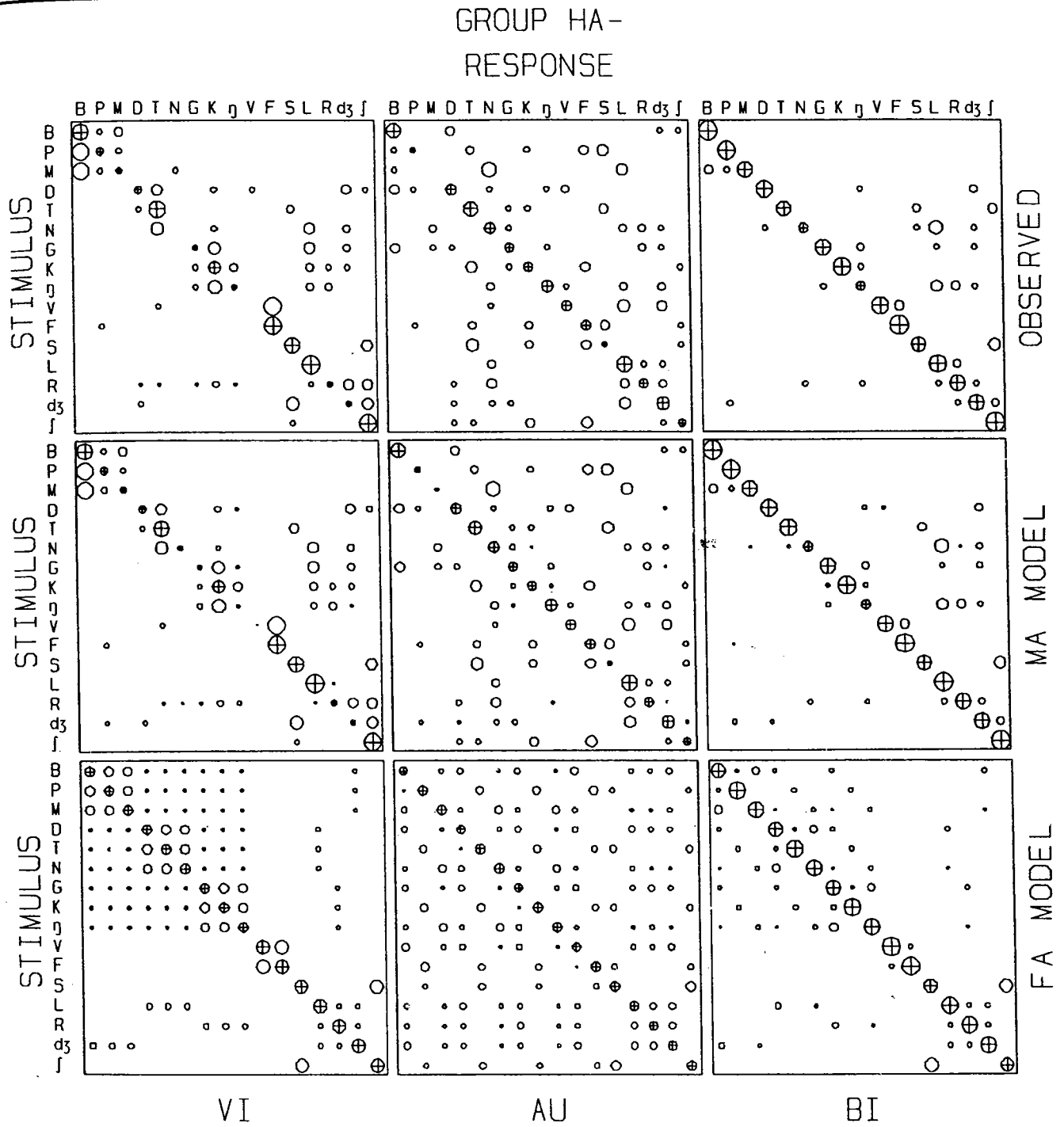
Figure 10. Observed (top panel) and predicted (bottom panels) confusion matrices for the eight observers with hearing aids and relatively good auditory recognition in the Agelfors study. The predictions are for the modality-analysis (MA) and the feature-analysis (FA) implementation of the FLMP.



feature-analysis FLMP is appropriate only for different categories of segments because they have different feature descriptions. A more discriminating test would have included a continuum between the visual stimuli in addition to the auditory stimuli. Symmetrical factorial designs have the greatest number of independent

observations relative to the number of free parameters. Only three responses were permitted in the Walden et al. study. Therefore 3×14 auditory parameters and 3×2 visual parameters are necessary. This gives a total of 48 free parameters to predict 126 data points. The predictions of the FLMP capture the joint influence of these

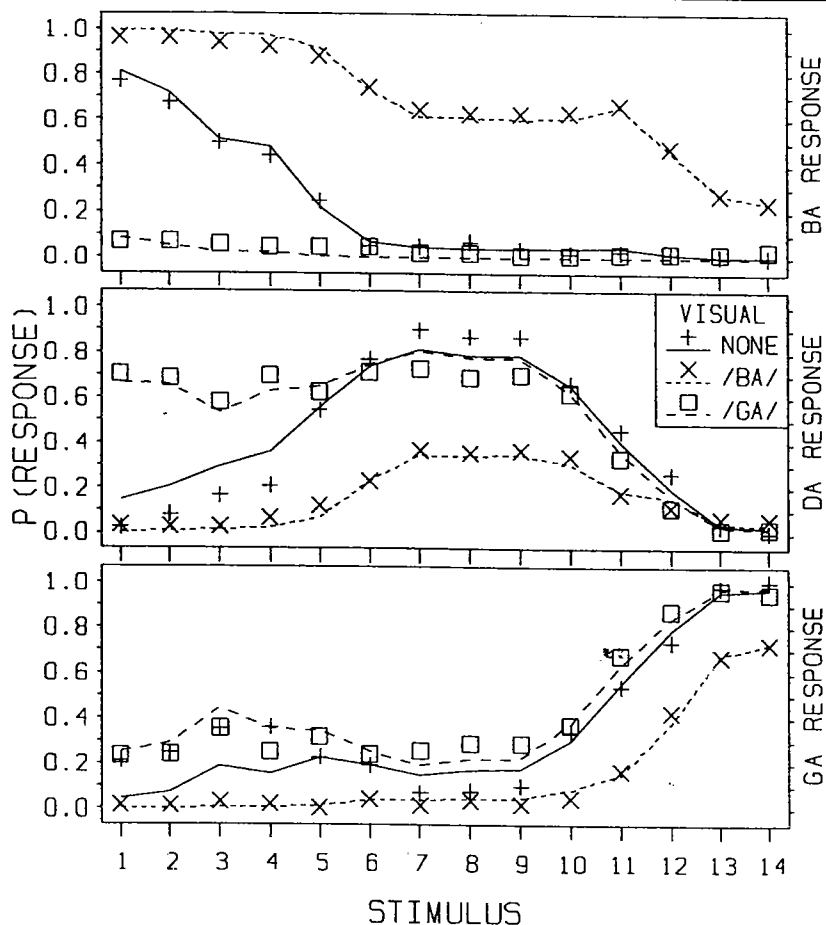
Figure 11. Observed (top panel) and predicted (bottom panels) confusion matrices for the eight observers with hearing aids and relatively poor auditory recognition in the Agelfors study. The predictions are for the modality-analysis (MA) and the feature-analysis (FA) implementation of the FLMP.



two modalities with an RMSD of .0544. The parameter values of the model are consistent with the observed and predicted results. The auditory parameter values more or less follow the curves marked visual-none in Figure 12. The visual /ba/ provided about 20 times more support for the alternative /ba/ than it did for /da/ and /ga/

combined. The visual /ga/ supported the alternative /ga/ to degree .544 and the alternative /da/ to degree .416. The similarity between these two values is consistent with the general finding that /d/ and /g/ belong in the same viseme class and are difficult to distinguish unless the inside of the talker's mouth is well illuminated.

Figure 12. Observed (points) and predicted (lines) proportion of /ba/, /da/, and /ga/ identifications for the hearing-impaired adults in the Walden et al. study as a function of auditory stimulus ranging from /ba/ to /da/ to /ga/ and the visual stimulus. Observed results from Walden et al. (1990). The lines give the predictions for the FLMP.



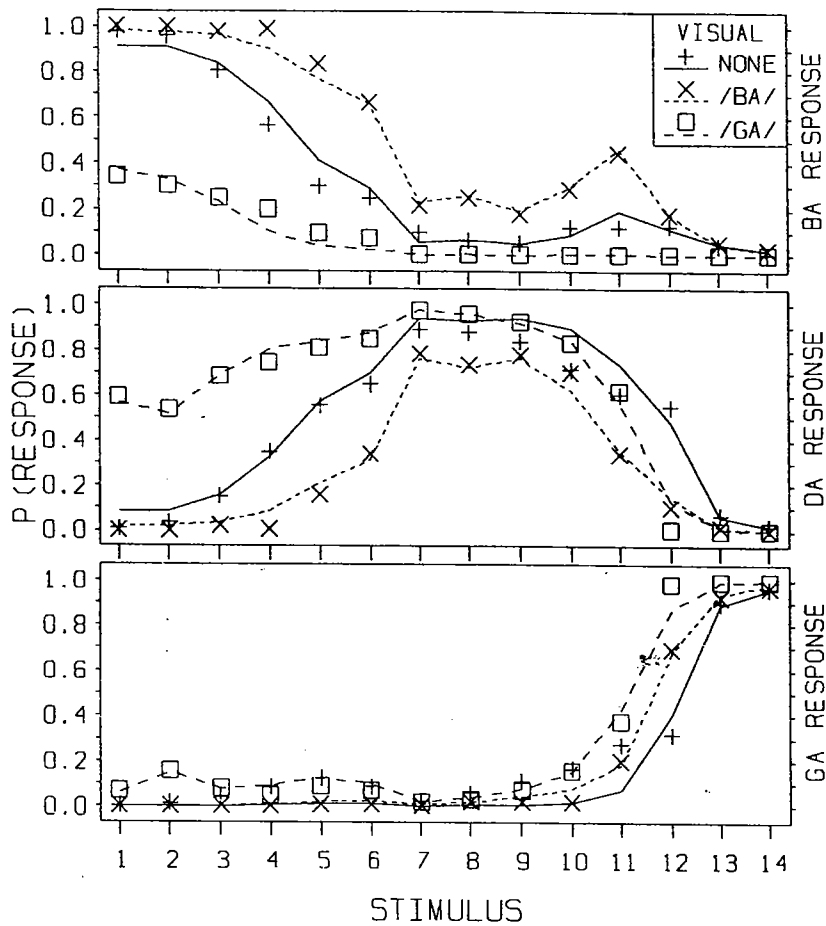
Walden et al. also tested a group of 15 individuals with normal hearing. These adults were about two decades younger than the hearing-impaired observers. The auditory stimuli for these participants were presented in a low-level broadband noise, with a +10 S/N ratio. The average observed results are shown in Figure 13 along with the predictions of the FLMP. These results with the normal-hearing observers showed a big impact of the visible speech, but not as big as that found for the perceivers with hearing loss. Of course, we would expect a bigger impact of visible speech if the noise was made even more intense for the control observers. The FLMP was fit to these results in the same manner as for the hearing-impaired observers. The RMSD was .0520, indicating that the FLMP does equally well describing the integration of audible and visible speech for noisy auditory speech or for hearing-impaired observers.

In another experiment, normal-hearing controls were tested with noise-free speech. In this case, the visible speech had much less of an influence. The major

impact was that a visual /ga/ greatly reduced /ba/ responses. These tended to be replaced by /da/ and /ga/ responses. Normal-hearing participants showed much less of an influence with normal auditory speech but a much larger influence when the auditory speech was degraded. According to our perspective, this result is entirely understandable. Hearing-impaired observers integrate information in the same manner as those with normal hearing, but they have less auditory information. One can be made to resemble the other by assigning the appropriate quality of information.

Our paradigm thus offers a potentially useful framework for the assessment and training of individuals with hearing impairment (Grant & Walden, 1995). The good fit of the FLMP illustrates that it accounts for speech perception in both normal and sensory-impaired individuals. We would expect visually impaired individuals to be less influenced by visible speech, but we do not know of any experiments that have examined this issue. Campbell and Massaro (1997) found that

Figure 13. Observed (points) and predicted (lines) proportion of /ba/, /da/, and /ga/ identifications for the normal-hearing adults in the Walden et al. study as a function of auditory stimulus ranging from /ba/ to /da/ to /ga/ and the visual stimulus. Broad band noise was added to the auditory stimulus. Observed results from Walden et al. (1990). The lines give the predictions for the FLM.



speechreading remained relatively good even when the visible speech was degraded somewhat by spatial quantization. Nonsighted children appear to have some difficulty learning those speech distinctions that are visibly salient and auditorily difficult (Mills, 1987). A potentially valuable study would be to manipulate the quality of both audible and visible speech in these types of tasks.

Discussion

This exercise in the discovery and analysis of confusion data from speech recognition experiments with normal and hearing-impaired individuals has confirmed many of the principles derived from recent experimental and theoretical studies of individuals with normal hearing (Grant et al., 1998). In addition, the experiments with individuals with hearing loss tend to be more ecologically valid in that many more stimuli and response

alternatives are used. The extension of the FLMP to these data sets was successful along several dimensions. First, the assumptions of the model appear to be equally powerful in describing the confusion matrices as they are in describing simpler experiments using expanded factorial designs. Second, the FLMP was extended to incorporate features as sources of information in speech perception. It should be stressed that we take no stand on the psychological reality of linguistic features. The implementation of the model simply assumes that there are sources of information from each modality that align themselves with traditional linguistic features. In this sense, the good fit of the feature-implementation of the FLMP reveals that the distinctions in spoken language can be well-described by traditional linguistic features. Future work should address the issue of whether alternative feature sets might give an even better description of the results.

The integration of degraded audible and visible speech by individuals with hearing loss addresses a new

research question. Performance of people with hearing loss appears to match that of people with normal hearing when the latter are presented with degraded auditory inputs (Campbell, 1974, discussed in Massaro, 1987, pp. 42–43). The FLMP predictions are that the information processing of people with normal hearing should remain invariant even though there is less auditory information. The good fit of the FLMP to these results indicate that normal hearing individuals evaluate and integrate visible speech with auditory speech in noise in the same way as they do with undegraded auditory speech. Furthermore, the good fit of the people with hearing loss in the present studies indicates similar processing for the two groups of individuals. In terms of the framework of the FLMP, the two groups differ in terms of information but are identical in terms of how the information is processed. Furthermore, the good fit of the FLMP indicates that the integration was optimal. The discrepancy with the Grant et al. (1998) finding that integration was not optimal remains to be resolved.

These positive findings encourage the use of multimodal environments for persons with hearing loss. Ling (1976, p. 51) reports that clinical experience seems to show that “children taught exclusively through a multisensory approach generally make less use of residual audition.” For these reasons, speech-language pathologists might use visible bimodal training less often than would be beneficial. There is some evidence that video feedback from their own speech production improved the speech production of adults with profound hearing loss (De Filippo & Sims, 1995). To evaluate multisensory control of speech production, the same type of research design used for the study of speech perception might be used to study speech production. It is well known that individuals with severe or profound hearing loss tend to have poorer speech production skills. An experiment could be carried out in which these individuals are asked to produce speech given auditory, visual, or bimodal speech input. The working hypothesis would be that speech production would be better (and learned more easily) given bimodal input relative to either source of information presented alone.

Acknowledgments

This research was supported, in part, by grants from the Public Health Service (PHS R01 DC00236), the National Science Foundation (23818), Intel Corporation, and the University of California, Santa Cruz. The authors are appreciative of the previous researchers who provided the detailed results in their original publications and also thank Eva Agelfors for providing the detailed results of the confusion matrices in order to make the present analyses of her results possible. The helpful comments of Christopher Turner, Sandra Gordon-Salant, and two anonymous reviewers are gratefully acknowledged.

References

- Agelfors, E. (1996). A comparison between patients using cochlear implants and hearing aids. Part I: Results on speech tests. “Royal Institute of Technology Speech, Music and Hearing.” *KTH Quarterly Progress and Status Report* (TMH-APSR 1), 63–75.
- Braida, L. D. (1991). Crossmodal integration in the identification of consonant segments. *Quarterly Journal of Experimental Psychology*, 43, 647–677.
- Campbell, C. S., & Massaro, D. W. (1997). Visible speech perception: Influence of spatial quantization. *Perception*, 26, 627–644.
- Campbell, H. W. (1974). *Phoneme recognition by ear and by eye: A distinctive feature analysis*. Unpublished doctoral dissertation, University of Nijmegen, Holland.
- Chandler, J. P. (1969). Subroutine STEPIT finds local minima of a smooth function of several parameters. *Behavioral Science*, 14, 81–82.
- De Filippo, C. L., & Sims, D. G. (1995). Linking visual and kinesthetic imagery in lipreading instruction. *Journal of Speech and Hearing Research*, 38, 244–256.
- Dowell, R. C., Martin, L. F. A., Tong, Y. C., Clark, G. M., Seligman, P. M., & Patrick, J. F. (1982). A 12-consonant confusion study on a multiple-channel cochlear implant patient. *Journal of Speech and Hearing Research*, 25, 509–516.
- Erber, N. P. (1972). Auditory, visual, and auditory-visual recognition of consonants by children with normal and impaired hearing. *Journal of Speech and Hearing Research*, 15, 413–422.
- Grant, K. W., & Walden, B. E. (1995). Predicting auditory-visual speech recognition in hearing-impaired listeners. *Proceedings of the XIIIth International Congress of Phonetic Sciences*, 3, 122–129.
- Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *Journal of the Acoustical Society of America*, 103, 2677–2690.
- Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *Journal of the Acoustical Society of America*, 97, 553–562.
- Ling, D. (1976). *Speech and the hearing-impaired child: Theory and practice*. Washington, DC: Alexander Graham Bell.
- Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.
- Massaro, D. W. (1972). Perceptual images, processing time, and perceptual units in auditory perception. *Psychological Review*, 79, 124–145.
- Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.
- Massaro, D. W., & Friedman, D. (1990). Models of

- integration given multiple sources of information. *Psychological Review*, 97, 225-252.
- McGurk, H., & MacDonald, J.** (1976). Hearing lips and seeing voices. *Nature*, 264, 746-748.
- Miller, G. A., & Nicely, P. E.** (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, 27, 338-352.
- Mills, A. E.** (1987). The development of phonology in the blind child. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 145-161). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Remez, R. E., Rubin P. E., Berns S. M., Pardo J. S., & Lang, J. M.** (1994). On the perceptual organization of speech. *Psychological Review*, 101, 129-156.
- Sumby, W. H., & Pollack, I.** (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26, 212-215.
- 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. 3-51). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Thurstone, L. L.** (1927). A law of comparative judgment. *Psychological Review*, 34, 273-286.
- Walden, B., Montgomery, A., Prosek, R. A., & Hawkins, D. B.** (1990). Visual biasing of normal and impaired auditory speech perception. *Journal of Speech and Hearing Research*, 33, 163-73.
- Zadeh, L. A.** (1965). Fuzzy sets. *Information and Control*, 8, 338-353.

Received February 20, 1998

Accepted September 16, 1998

Contact author: Dominic W. Massaro, PhD, Department of Psychology, University of California, Santa Cruz, Santa Cruz, CA 95064. Email: massaro@fuzzy.ucsc.edu