DOMINIC W. MASSARO

MULTIMODAL SPEECH PERCEPTION: A PARADIGM FOR SPEECH SCIENCE

1. INTRODUCTION

Speech science evolved as the study of a unimodal phenomenon. Speech was viewed as a solely auditory event, as captured by the seminal speech-chain illustration of Denes & Pinson (1963) shown in Figure 1.

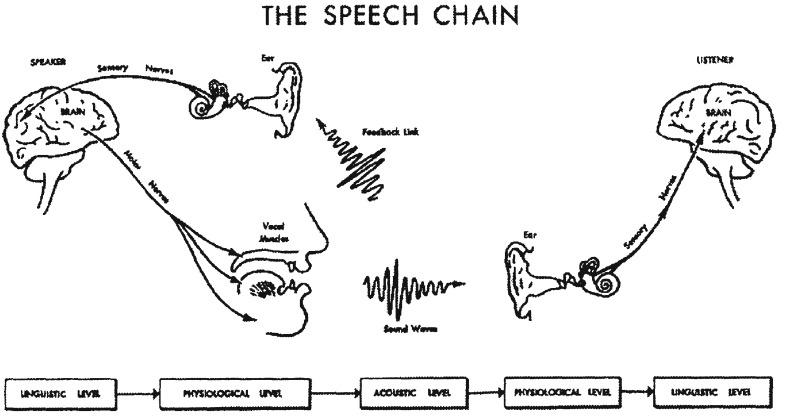

Figure 1. The classic speech-chain illustration of Denes & Pinson (1963).

This view is no longer viable as witnessed by this book as well as a burgeoning record of research findings. Although Denes & Pinson viewed speech as primarily an auditory phenomenon (rather than a multimodal one), they did acknowledge the important contribution of context to accurate recognition and understanding. In accepting the influence of both stimulus information and context on speech perception, the authors anticipated the approach taken in the present chapter. They stated,

> *"In speech communication, then, we do not actually rely on a precise knowledge of specific cues. Instead, we related a great variety of ambiguous cues against the background of the complex system we call our common language."* (Denes & Pinson, 1963, p. 8).

B. Granström et al. (eds.), Multimodality in Language and Speech Systems, 45–71.

Speech as a multimodal phenomenon is supported by experiments indicating that our perception and understanding are influenced by a speaker's face and accompanying gestures, as well as the actual sound of the speech (Massaro, 1987, 1998). Many communication environments involve a noisy auditory channel, which degrades speech perception and recognition. Visible speech from the talker's face (or from a reasonably accurate synthetic talking head) improves intelligibility in these situations. Visible speech also is an important communication channel for individuals with hearing loss.

The number of words understood from a degraded auditory message can often be doubled by pairing the message with visible speech from the talker's face. The combination of auditory and visual speech has been called super-additive because their combination can lead to accuracy that is much greater than accuracy on either modality alone. Furthermore, the strong influence of visible speech is not limited to situations with degraded auditory input. A perceiver's recognition of an auditory-visual syllable reflects the contribution of both sound and sight. For example, if the ambiguous auditory sentence, *My bab pop me poo brive*, is paired with the visible sentence, My *gag kok me koo grive*, the perceiver is likely to hear, *My dad taught me to drive*. Two ambiguous sources of information are combined to create a meaningful interpretation (Massaro, 1998).

There are several reasons why the use of auditory and visual information together is so successful. These include (a) robustness of visual speech, (b) complementarity of auditory and visual speech, and (c) optimal integration of these two sources of information. Speechreading, or the ability to obtain speech information from the face, is robust in that perceivers are fairly good at speech reading even when they are not looking directly at the talker's lips. Furthermore, accuracy is not dramatically reduced when the facial image is blurred (because of poor vision, for example), when the face is viewed from above, below, or in profile, or when there is a large distance between the talker and the viewer (Massaro, 1998).

Complementarity of auditory and visual information simply means that one of the sources is strong when the other is weak. A distinction between two segments robustly conveyed in one modality is relatively ambiguous in the other modality. For example, the place difference between /ba/ and /da/ is easy to see but relatively difficult to hear. On the other hand, the voicing difference between /ba/ and /pa/ is relatively easy to hear but very difficult to discriminate visually. Two complementary sources of information make their combined use much more informative than would be the case if the two sources were non-complementary, or redundant (Massaro, 1998, pp. 424-427).

The final characteristic is that perceivers combine or integrate the auditory and visual sources of information in an optimally efficient manner. There are many possible ways to treat two sources of information: use only the most informative source, average the two sources together, or integrate them in such a fashion in which both sources are used but that the least ambiguous source has the most influence. Perceivers in fact integrate the information available from each modality to perform as efficiently as possible. A wide variety of empirical results has been accurately predicted by a model that describes an optimally efficient process of combination.

In this chapter, I will analyze the multimodality of spoken language understanding within an information-processing framework. After describing the framework, a specific theoretical model is described to help organize the descriptions of experiments and theories. Several alternative theories are then presented and evaluated. To test among the theories, we discuss how the theories account for the influence of multiple sources of stimulus information in speech perception. To structure our information-processing analysis of spoken language understanding, we use a specific theoretical framework that has received substantial support from a variety of experiments in speech perception.

2. THEORETICAL FRAMEWORK

The general theoretical framework provided by the information-processing approach is based on the assumption that there is a sequence of processing stages in spoken language understanding. Stages of information processing have guided, for example, much of the research in visual perception (Palmer, 1999). Visual perception is assumed to occur in three stages of processing: retinal transduction, sensory cues (features), and perceived attributes (DeYoe & Van Essen, 1988). Visual input is transduced by the visual system, a conglomeration of sensory cues is made available, and attributes of the visual world are experienced by the perceiver. In visual perception, there is both a one-to-many and a many-to-one relationship between sensory cues and perceived attributes. The sensory cue of motion provides information about both perceived shape of an object and its perceived movement. A case of the many-to-one relationship in vision is that information about the shape of an object is enriched not only by motion, but also by perspective cues, picture cues, binocular disparity, and shading (e.g., chicariscuro).

We apply this same framework to speech perception and spoken language understanding. Speech perception via the auditory modality is characterized by a transduction of the acoustic signal along the basilar membrane, sensory cues, and perceived attributes. A single sensory cue can influence several perceived attributes. The duration of a vowel provides information about vowel identity (bit vs. beet), information such as lexical stress (the noun and verb pronunciations of the word *permit*), and syntactic boundaries in sentences. Another example is that the pitch of a speaker's voice is informative about both the identity of the speaker and intonation. The best-known example of multiple cues to a single perceived attribute in speech is the case of the many cues for the voicing of a medial stop consonant (Cohen, 1979; Lisker, 1978). These include the duration of the preceding vowel, the onset frequency of the fundamental, the voice onset time, and the silent closure interval. A multimodal example is the impressive demonstration that both the speech sound and the visible mouth movements of the speaker influence perception of place of articulation of a stop consonant (Massaro & Cohen, 1983; McGurk & MacDonald, 1976).

Our research and that of many others has demonstrated a powerful influence of visible speech in face-to-face communication. The influence of several sources of information from several modalities provides a new challenge for theoretical

accounts of speech perception. For theories that were developed to account for the perception of unimodal auditory speech (Diehl & Kluender, 1987, 1989), it is not obvious how they would account for the positive contribution of visible speech. Some extant theories view speech perception as a specialized process and not solely as an instance of pattern recognition (Liberman & Mattingly, 1985; Mattingly & Studdert-Kennedy, 1991). We take a different approach by envisioning speech perception as an instance of a more general process of pattern recognition (Massaro, 1998). In language processing, recognition is achieved via a variety of bottom-up and top-down sources of information. Top-down sources include contextual, semantic, syntactic, and phonological constraints; bottom-up sources include audible and visible features of the spoken word. A top-down source might be the overall frequency of a speech segment in the perceiver's language. A bottom-up source might be the degree of jaw rotation while talking.

3. THEORETICAL/EMPIRICAL INQUIRY

Our general framework documents the value of a combined experimental/theoretical approach. The research has contributed to our understanding of the characteristics used in speech perception, how speech is perceived and recognized, and the fundamental psychological processes that occur in speech perception and in pattern recognition in a variety of other domains.

We evaluate the contribution of visible information in face-to-face communication and how it is combined with auditory information in the ecologically valid condition of bimodal speech perception (face-to-face communication). Psychophysical and pattern-recognition tasks are carried out to analyze which audible and visible features are used by human observers in auditory, visual, and auditory-visual (bimodal) speech perception. Quantitative models of feature evaluation and integration are tested against identification judgments, ratings, and confusion matrices from perceptual tests. The results are used to determine which features influence performance.

The results are also to test formal models of speech perception. The models are formalized to make quantitative predictions of the judgments of the test items. Multiple models are tested to preclude a confirmation bias and to adhere to a falsification strategy of inquiry (Massaro, 1989, chapter 5). Each model is tested against the results of single subjects in order to avoid the pitfalls of averaging results across subjects. We also test a variety of participants to explore a broad variety of dimensions of individual variability. These include (1) life-span variability, (2) language variability, (3) sensory impairment, (4) brain trauma, (5) personality, (6) sex differences, and (7) experience and learning. In addition, a large variety of experimental procedures and test situations are used in our investigations (Massaro, 1998, Chapter 6). Generally, we need to know to what extent the processes uncovered in our research generalize across (1) sensory modalities, (2) environmental domains, (3) test items, (4) behavioural measures, (5) instructions, (6) and tasks.

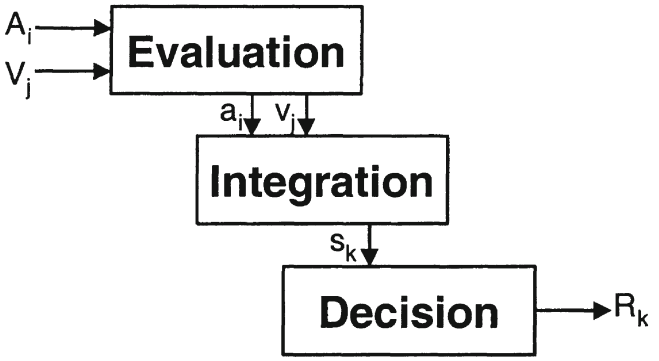

Figure 2. Schematic representation of the three processes involved in perceptual recognition. The three processes are shown to proceed left to right in time to illustrate their necessarily successive but overlapping processing. These processes make use of prototypes stored in long-term memory. The sources 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_j). 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.

We believe that our empirical work would be inadequate and perhaps invalid without the corresponding theoretical framework. Thus, the research addresses both empirical and theoretical issues. At the empirical level, experiments are carried out to determine how visible speech is combined with auditory speech for a broad range of individuals and across a wide variation of situational domains. At the theoretical level, the assumptions and predictions of several models are formalized, analyzed, contrasted, and tested. Various types of model fitting strategies have been employed, with similar outcomes. These model tests have been highly informative with respect to improving our understanding of how spoken language is perceived and understood.

4. FUZZY LOGICAL MODEL OF PERCEPTION

We have learned that a variety of empirical results can be successfully described within a framework of a fuzzy logical model of perception (FLMP). The FLMP assumes necessarily successive but overlapping stages of processing, as shown in Figure 2. The perceiver of speech is viewed as having multiple sources of information supporting the identification and interpretation of the language input. The model assumes that (1) each source of information is evaluated to give the continuous degree to which that source supports various alternatives, (2) the sources of information are evaluated independently of one another, (3) the sources are

integrated to provide an overall degree of support for each alternative, and (4) perceptual identification and interpretation follows the relative degree of support among the alternatives.

The paradigm that we have developed permits us to determine how visible speech is processed and integrated with other sources of information. The results also inform us about which of the many potentially functional cues are actually used by human observers (Campbell & Massaro, 1997; Massaro, 1987, Chapter 1; Massaro & Cohen, 1999). The systematic variation of properties of the speech signal combined with the quantitative test of models of speech perception enables the investigator to test the psychological validity of different cues. This paradigm has already proven to be effective in the study of audible, visible, and bimodal speech perception (Massaro, 1987, 1989, 1998). Thus, our research strategy not only addresses how different sources of information are evaluated and integrated, but can uncover what sources of information are actually used. We believe that the research paradigm confronts both the important psychophysical question of the nature of information and the process question of how the information is transformed and mapped into behaviour. Many independent tests point to the viability of the FLMP as a general description of pattern recognition. The FLMP is centered around a universal law of how people integrate multiple sources of information. This law and its relationship to other laws is developed in detail in Massaro (1998).

The assumptions of the FLMP are testable because they are expressed in quantitative form. The founding or keystone assumption of this model is the division of perception into the twin levels of information and information processing. Adhering to this fundamental dichotomy are a number of other testable assumptions. One is the idea that at the information level, sources are evaluated independently. Independence of sources is motivated by the principle of category-conditional independence (Massaro & Stork, 1998): it is not possible to predict the evaluation of one source on the basis of the evaluation of another, so the independent evaluation of both sources is necessary to make an optimal category judgment. While sources are thus kept separate at evaluation, they are then integrated to achieve perception and interpretation.

Multiplicative integration yields a measure of total support for a given category identification. This operation, implemented in the model, allows the combination of two imperfect sources of information to yield better performance than would be possible using either source by itself. However, the output of integration is an absolute measure of support; it must be relativized, due to the observed factor of relative influence (the influence of one source increases as other sources become less influential, i.e. more ambiguous). Relativization is effected through a decision stage, which divides the support for one category by the summed support for all other categories. An important empirical claim about this algorithm is that while information may vary from one perceptual situation to the next, the manner of combining this information – information processing – is invariant. With our algorithm, we thus propose an invariant law of pattern recognition describing how continuously perceived (fuzzy) information is processed to achieve perception of a category.

5. APPLIED VALUE OF RESEARCH

Many communication environments involve a noisy auditory channel, which degrades speech perception and recognition. Visible speech from the talker's face (or from a reasonably accurate synthetic talking head) improves intelligibility in these situations. Another applied value of visible speech is its potential to supplement other (degraded) sources of information for disabled individuals (Massaro & Cohen, 1999; Oerlemans & Blamey, 1998). Its use is important for hearing-impaired individuals because it allows effective communication within spoken language, the universal language of the community. Just as synthetic auditory speech has been of great importance for research on auditory speech perception, synthetic visual speech is important in studying visual speech perception. In addition, just as auditory speech synthesis has proved a boon to our visually impaired citizens in human machine interaction, visual speech synthesis may prove to be valuable for the hearing impaired. As just one example, cochlear implants have been shown to be successful in allowing implanted individuals to communicate via spoken language. In many situations, however, the electrical speech is not adequate, but the addition of visible speech allows successful communication (Schindler & Merzenich, 1985; Tyler et al., 1992).

It has been estimated by NIDCD that more than twenty-eight million Americans are hearing impaired. It is also the case that roughly three million Americans are estimated to have a corrected visual acuity of 20/40 or worse. With the rapidly increase in the number of elderly people, and the increase in visual and hearing impairment with aging, it is critical that we understand how people process multiple and somewhat ambiguous channels. There is also an unexplored positive potential of visible speech for (1) improving the quality of speech of persons with perception and production deficits, (2) enhancing second language learning and communication, (3) remedial training for poor readers, and (4) human-machine interactions.

6. DEMONSTRATION EXPERIMENT: VARYING THE AMBIGUITY OF THE SPEECH MODALITIES

An important manipulation is to systematically vary the ambiguity of each of the source of information in terms of how much it resembles each syllable. Synthetic speech (or at least a sophisticated modification of natural speech) is necessary to implement this manipulation. In a previous experimental task, we used synthetic speech to cross five levels of audible speech varying between /ba/ and /da/ with five levels of visible speech varying between the same alternatives. We also included the unimodal test stimuli to implement the expanded factorial design, as shown in Figure 3.

6.1. Prototypical Method

The properties of the auditory stimulus were varied to give an auditory continuum between the syllables /ba/ and /da/. In analogous fashion, properties of our animated face were varied to give a continuum between visual /ba/ and /da/. Five levels of

audible speech varying between /ba/ and /da/ were crossed with five levels of visible speech varying between the same alternatives. In addition, the audible and visible speech also were presented alone for a total of 25 + 5 + 5 = 35 independent stimulus conditions. Six random sequences were determined by sampling the 35 conditions without replacement giving six different blocks of 35 trials. An experimental session consisted of these six blocks preceded by six practice trials and with a short break between sessions. There were four sessions of testing for a total of 840 test trials (35 x 6 x 4). Thus there were 24 observations at each of the 35 unique experimental conditions. Subjects were instructed to listen and to watch the speaker, and to identify the syllable as /ba/ or /da/. This experimental design was used with 82 participants and their results have served as a database for testing models of pattern recognition (Massaro, 1998).

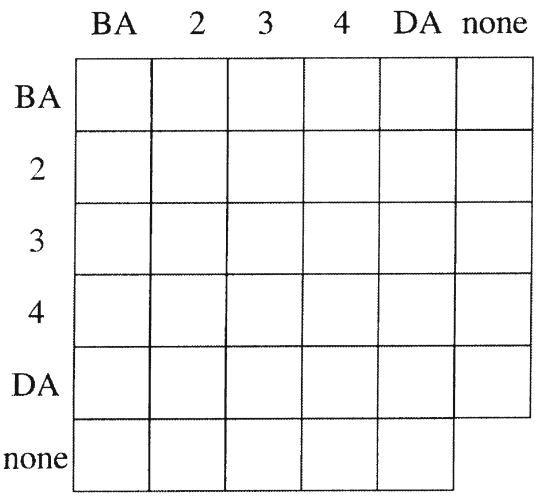

Figure 3. Expansion of a typical factorial design to include auditory and visual conditions presented alone. The five levels along the auditory and visible continua represent auditory and visible speech syllables varying in equal physical steps between /ba/ and /da/.

6.2. Prototypical Results

We call these results prototypical because they are highly representative of many different experiments of this type. The mean observed proportion of /da/ identifications was computed for each subject for the 35 unimodal and bimodal conditions. For this tutorial, we present the results for three participants who can be considered typical of the others in this task. The points in Figure 4 give the observed proportion of /da/ responses for the auditory alone (left plot), the bimodal (middle plot), and the visual alone (right plot) conditions as a function of the five levels of the synthetic auditory and visual speech varying between /ba/ and /da/. For the unimodal plots,

the degree of influence of a modality is indicated by the steepness of the response function. By this criterion, both the auditory and the visual sources of information had a strong impact on the identification judgments. As illustrated in the left and right plots, the identification judgments changed systematically with changes in the audible and visible sources of information. The likelihood of a /da/ identification increased as the auditory speech changes from /ba/ to /da/, and analogously for the visible speech.

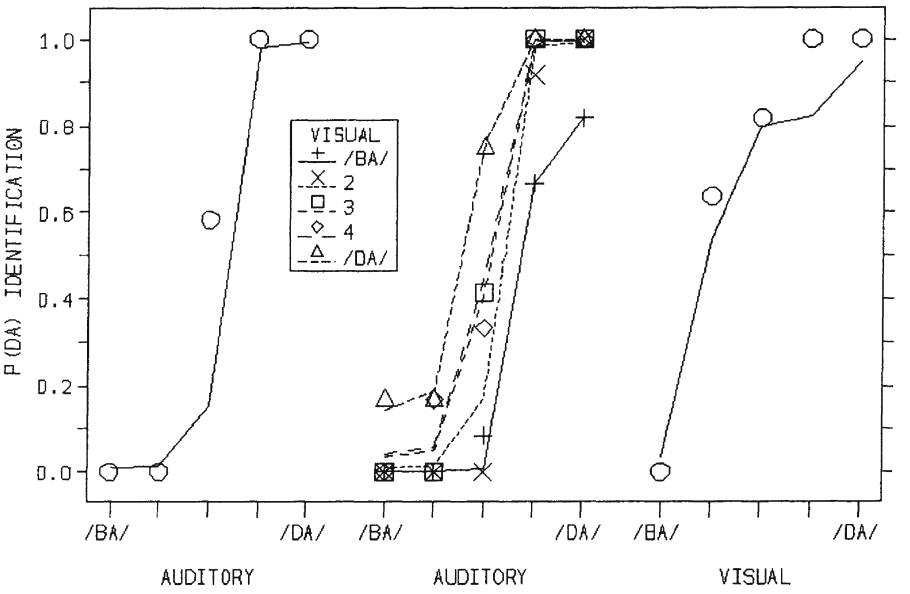

Figure 4. The points give the observed proportion of /da/ identifications for a typical observer in the auditory-alone (left panel), the factorial auditory-visual (center panel) and the visual-alone (right panel) conditions as a function of the five levels of the synthetic auditory and visual speech varying between /ba/ and /da/. The lines give the predictions of the FLMP.

For the bimodal results in the middle plot, the degree of influence is again indexed by the slope of the function for the variable plotted on the x-axis, and by the spread among the curves for the variable described in the key or legend. By these criteria, both sources had a large influence in the bimodal conditions. The curves across changes in the auditory variable are relatively steep and also spread out from on another with changes in the visual variable.

Finally, the auditory and visual effects were *not* additive in the bimodal condition, as demonstrated by a significant auditory-visual interaction. The interaction is indexed by the change in the spread among the curves across changes in the auditory variable. This vertical spread between the curves is about four times greater in the middle than at the end of the auditory continuum. It means that the influence of one source of information is greatest when the other source is neutral or ambiguous. We now address how the two sources are used in perception.

6.3. Evaluation of How Two Sources are Used

Of course, an important question is how the two sources of information are used in perceptual recognition. An analysis of several results informs this question. Figure 5 gives the results for another participant in the task. Three points are circled in the figure to highlight the conditions in which the third level of auditory information is paired with the first (/ba/) level of visual information. When presented alone, P(/ba/ | A_3) is about .2 whereas P(/ba/| V_1) is about .8. When these two stimuli occur together, P(/ba/| V_1 A_3) is about .5. This subset of results is consistent with just about any theoretical explanation, for example, one in which only a single source of information is used on a given trial. Similarly, a simple averaging of the audible and visible speech predicts this outcome.

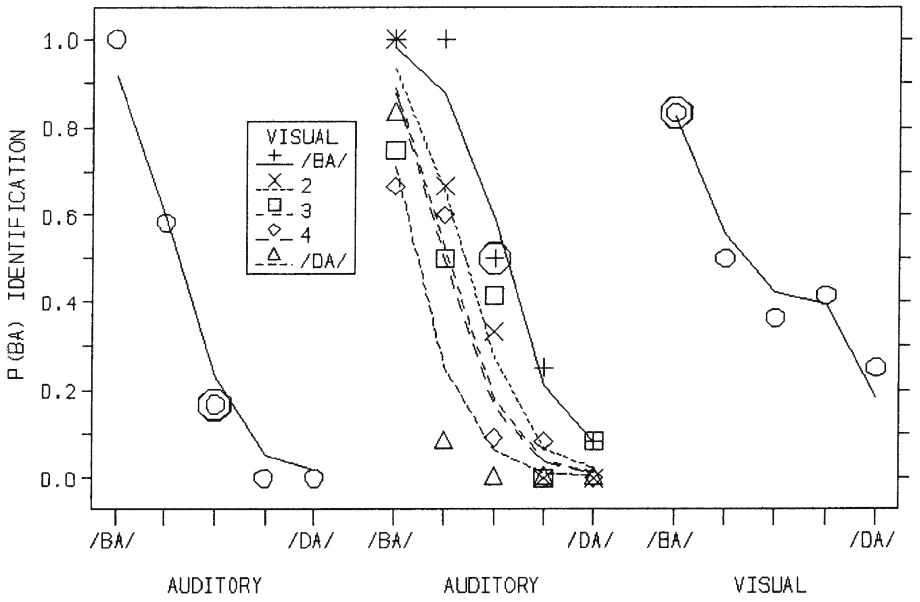

Figure 5. The points give the observed proportion of /ba/ identifications for a typical observer in the auditory-alone (left panel), the factorial auditory-visual (center panel) and the visual-alone (right panel) conditions as a function of the five levels of the synthetic auditory and visual speech varying between /ba/ and /da/. The three circled points A_3 V_1 give two unimodal conditions and the corresponding bimodal condition. The relationship among the three points can be explained by the use of a single modality, an averaging of the two sources, or a multiplicative integration of the two sources. The lines are the predictions of the FLMP.

Other observations, however, allow us to reject these alternatives. Figure 6 gives the results for yet another participant in the task. Three points are circled in the figure to highlight the conditions in which the second level of auditory information is paired with the second level of visual information. When presented alone, P(/ba/ | A_2) is about .8 and P(/ba/| V_2) is about .8. When these two stimuli occur together,

P(/ba/| V_2 A_2) is about 1. This so-called super-additive result (the bimodal is more extreme than either unimodal response proportion) is not easily explained by either the use of a single modality or a simple averaging of the two sources, but is well described by the FLMP. The quantitative predictions of the FLMP have been formalized in a number of different publications (e.g., Massaro, 1987, 1998). In a two-alternative task with /ba/ and /da/ alternatives, the degree of auditory support for /da/ can be represented by a_i, and the support for /ba/ by $(1 - a_i)$. Similarly, the degree of visual support for /da/ can be represented by v_j, and the support for /ba/ by $(1 - v_j)$. The probability of a response to the unimodal stimulus is simply equal to its feature value. For bimodal trials, the predicted probability of a response given auditory and visual inputs, P(/da/|A_iV_j) is equal to

$$P(/da/ \mid A_iV_j) = \frac{a_i v_j}{a_i v_j + (1-a_i)(1-v_j)} \qquad (1)$$

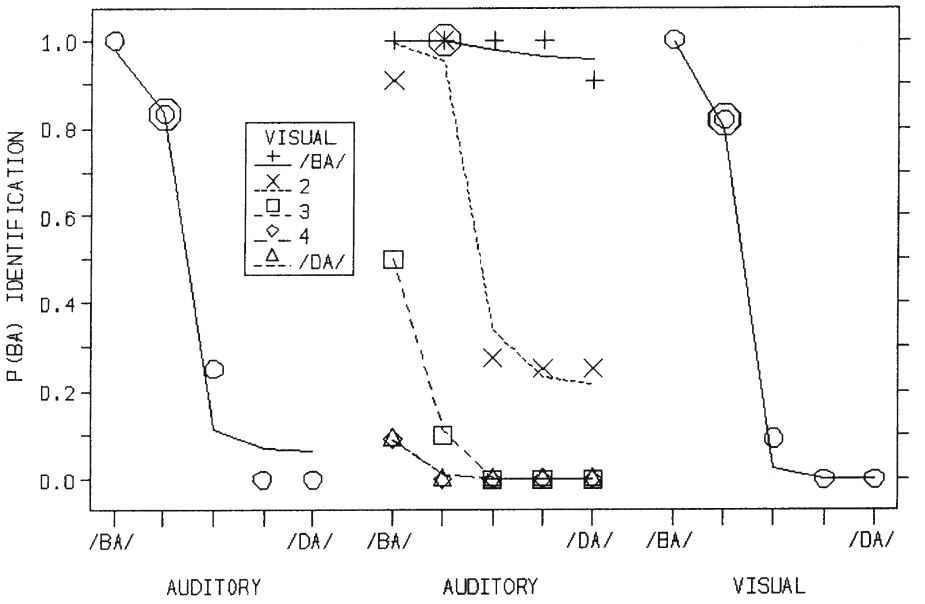

Figure 6. The points give the observed proportion of /ba/ identifications for a typical observer in the auditory-alone (left panel), the factorial auditory-visual (center panel) and the visual-alone(right panel) conditions as a function of the five levels of the synthetic auditory and visual speech varying between /ba/ and /da/. The three circled points A_2V_2 give two unimodal conditions and the corresponding bimodal condition. The relationship among the three points cannot *be explained by the use of a single modality or an averaging of the two sources, but can be described by a multiplicative integration of the two sources. The lines are the predictions of the FLMP.*

Given that these results using an expanded factorial design and tests of formal models, it is important to replicate this task under a broader set of conditions. These

basic findings hold up under a variety of experimental conditions (Massaro, 1998, Chapter 6). In one case, subjects were given just two alternatives, and in the other the same subjects were allowed an open-ended set of alternatives. When tested against the results, the FLMP gives a good description of performance, even with the constraint that the same parameter values are used to describe performance when the number of response alternatives is varied (see Massaro, 1998, pp. 265-268).

7. TESTS OF THE FLMP

We have found that the FLMP has provided the best description of a variety of results in bimodal speech perception. We have contrasted this model against a large number of alternative models. Our criterion for model selection has been the root mean square deviation (RMSD) between predicted and observed values. The RMSD provides an easily understood measure of the agreement between the actual and the theoretical outcomes. If we observe an RMSD of .04, then we know the average difference between the observed versus predicted values was .04. Recently, this measure has been called into question during a resurgence of interest in model testing and selection from both researchers in various domains of performance and also the mathematical modeling community (Cutting et al., 1992; Massaro, 1998; Myung & Pitt, 1997, 1998).

Myung & Pitt (1997) explored the predictive power of three extant models by simulating hypothetical data from a 2 by 8 factorial design, with 20 observations at each of the 16 experimental conditions. They began with three sets of hypothetical parameter values and simulated results from 100 subjects for each of the three sets. The models used to simulate the results were (1) a linear model (LIM) in which the values from the two independent variables are simply averaged, (2) the fuzzy logical model of perception (FLMP, Massaro, 1998), and (3) a model based on signal detection theory (TSD, Massaro & Friedman, 1990). In our earlier work (Massaro, 1987; Massaro & Friedman, 1990), we found that these models made different predictions from one another and that one model could not mimic another when the hypothetical results had no variability. The FLMP and TSD made very similar predictions, however, and are probably indistinguishable in practice. With sampling variability, Myung & Pitt demonstrated that the RMSD measure of goodness of fit was not always sufficient to recover the model that actually generated the original data. They found that FLMP appeared to be more powerful than LIM in that it sometimes gave a better account of the simulated results even when LIM was used to generate the data. When FLMP or TSD was used to generate the hypothetical results, the LIM model never provided a better fit than the other two models. Myung & Pitt (1997) proposed the Bayes factor (Kass & Raftery, 1995) for model selection, which incorporates both functional form and model complexity as criteria for selecting the best model. When applied to the simulated results, this new technique usually provided a recovery of the "correct" model. These results implied that the LIM model might have been erroneously rejected in our previous work (see also Cutting et al., 1992). This is obviously an undesirable state of affairs and challenges our previous conclusions in this arena.

This important analysis and potential solution provided by the Bayes factor alerted us to the possibility that our previous tests between alternative models may have been inadequate. Given the more powerful ability of the FLMP to fit results, even results that were not generated by that model, our conclusions might have been invalid. However, there were several aspects of the Myung & Pitt simulation that did not mirror our prototypical experimental situations. First, the authors simulated data from an unweighted averaging model (LIM) rather than a weighted averaging model (WTAV) that we have tested in all of our research (Massaro, 1998; Massaro & Cohen, 1976). The WTAV is more psychologically realistic in that it is unlikely that each factor is weighted equally in pattern recognition tasks. (This differential weighting in the FLMP emerges from the nonlinear combination of the two sources of information corresponding to the two factors.) Second, the authors simulated data from an asymmetrical factorial design whereas we usually carry out symmetrical expanded factorial designs. The latter are much more powerful than the former in discriminating among different models. A symmetrical design has the best ratio of independent observations to free parameters, and the expanded design provides an additional set of recognition probabilities whose expected values are assumed to be equal to the actual parameter values. Third, the authors used only three hypothetical sets of parameter values whereas we have contrasted the models in literally hundreds of independent tests.

To explore these differences, we carried out a series of comparisons of the use of RMSD versus Bayesian selection in the evaluation of extant models (Massaro et al., 2001). We used a database from the task described in the Demonstration experiment and shown in Figures 4-6 (Massaro et al., 1993; Massaro et al., 1995). This experimental design was used with 82 participants and their results also served as a database for testing models of pattern recognition (Massaro, 1998, Chapters 2 and 10).

For these 82 participants, the FLMP gave a better description than the WTAV model for 94% of the real subjects. To analyze the robustness of the RMSD measure, we created a set of hypothetical subjects who behaved according to either one model or the other. These montecarlo simulations involved creating 20 simulated subjects for each model for each real subject. By using the same number of trials, the simulation should have the same sampling variability as was present in the data set being modelled. For these simulated participants, the RMSD measure was sufficient to recover the original model that generated the data. For both data sets, the incorrect model was recovered only 1% of the time. These same results were used to test the models on the basis of the Bayes factor; Kass & Raftery, 1995) for model selection, which incorporates both functional form and model complexity as criteria for selecting the best model. When applied to these empirical results, this new technique did not change the conclusions that were reached. The FLMP maintained its significant descriptive advantage over the WTAV with this new criterion (Massaro et al., 2001). The outcomes support the conclusion that the RMSD measure yields similar outcomes to the Bayes factor for the conditions of our prototypical design. Thus, the validity of the FLMP holds up under even more demanding methods of model selection.

As in all things, there is no holy grail of model evaluation for scientific inquiry. As elegantly concluded by Myung & Pitt (1997), the use of judgment is central to model selection. Extending their advice, we propose that investigators should make use of as many techniques as feasible to provide converging evidence for the selection of one model over another. More specifically, both RMSD and the Bayes factor can be used as independent metrics of model selection. Inconsistent outcomes should provide a strong caveat for the validity of selecting one model over another in the same way that conflicting sources of information create an ambiguous speech event for the perceiver.

8. CLARIFIYING THE MCGURK EFFECT

It has been well over two decades since the publication of the McGurk effect (McGurk & MacDonald, 1976), which has obtained widespread attention in many circles of psychological inquiry and cognitive science (Green, 1998; Schwartz et al., 1998). The classic McGurk effect involves the situation in which an auditory /ba/ is paired with a visible /ga/ and the perceiver reports hearing /da/, called a fusion response. The reverse pairing, an auditory /ga/ and visual /ba/, tends to produce a perceptual judgment of /bga/, called a combination response. The finding that auditory experience is influenced by the visual input stimulated many students of speech perception to carry out similar investigations. However, many previous studies used just a few experimental conditions in which the auditory and visual sources of information are made to mismatch. Many experiments failed to test the unimodal conditions separately so that there is no independent index of the perception of the single modalities. The experiments also tend to take too few observations under each of the stimulus conditions. The data analysis is also usually compromised because investigators analyze the data with respect to whether or not there was a McGurk effect, which often is simply taken to mean whether the visual information dominated the judgments. This analysis can be highly misleading because we have seen in Figures 4-6 that one modality does not dominate the other. Both modalities contribute to the perceptual judgment with the outcome that the least ambiguous source of information has the most influence. I propose a better understanding of the McGurk effect by enhancing the database and testing formal models of the perceptual process.

To explore the McGurk effect more fully, we carried out a series of experiments in which the auditory syllables /ba/, /da/, and /ga/ were crossed with these same visible syllables in an expanded factorial design. Subjects are either limited to these three response alternatives or given a larger set of response alternatives. Why does auditory /ba/ paired with a visible /ga/ produce a perceptual report of hearing /da/ rather than /ga/? Initial explanation of this outcome has been to expect it to follow from the psychophysical properties of the audible and visible sources of information. This means that visual da/ and visual /ga/ are virtually indistinguishable and that auditory /ba/ must be somewhat more similar to an auditory /da/ than to an auditory /ga/.

Another possibility is that there are other sources of information (or constraints) contributing to the preference of /da/ over /ga/. One of the themes of our research is that there are multiple influences (both top-down and bottom-up) on perceptual processing. One potential top-down source is the frequency of occurrence of these segments in the language. Previous studies have shown that transitional probability contributes to perceptual processing (Massaro & Cohen, 1983; Pitt & McQueen, 1998), and word frequency has been shown to be highly functional in word recognition. Top-down context might be functional in the McGurk effect because the segment /d/ appears to be more frequent in initial position than the segment /g/ (Denes, 1963). This a priori bias for /d/ over /g/ (and /t/ over /k/) could be an important influence contributing to the "fusion" response that is observed.

To explore these two contributions, the natural auditory syllables /ba/, /da/, and /ga/ were crossed with the synthetic visual syllables /ba/, /da/, and /ga/. Participants also identified the unimodal syllables. Ten participants were tested for two sessions of 216 trials each, for a total of roughly 29 observations under each of the 15 conditions. Subjects were given the response alternatives /ba/, /da/, /ga/ or were permitted to make combination responses involving these alternatives (e.g., /bga/).

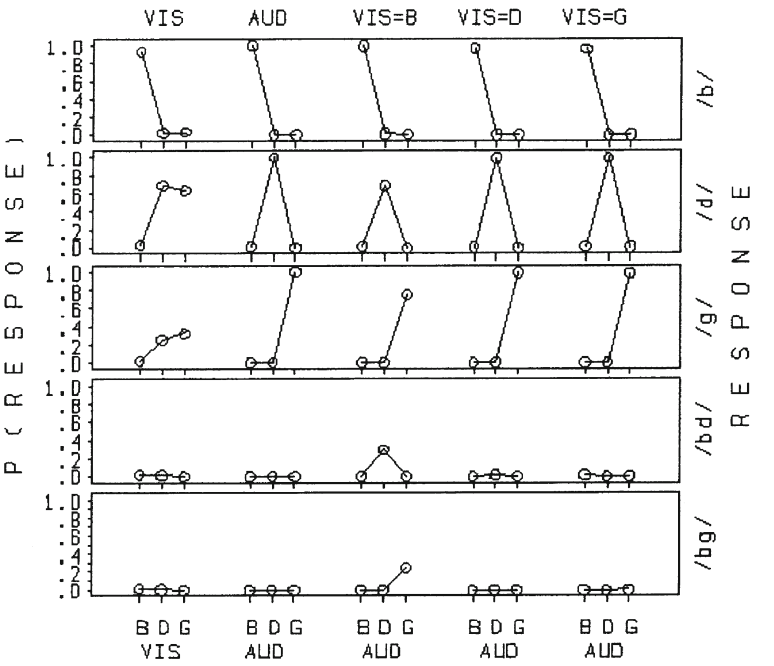

Figure 7. The percentage of /b/, /d/, /g/, /bd/, and /bg/ responses as a function of the three test stimuli in the unimodal visual (VIS), unimodal auditory (AUD), and bimodal conditions.

Figure 7 gives the probability of /ba/, /da/, /ga/, /bda/, and /bga/ responses for each of the 15 experimental conditions. Several results are of interest. As expected, there were confusions between visible /da/ and /ga/, because these syllables tend to look the same. Their major differences in articulation occur inside the mouth, which are hidden to the perceiver. As can be seen in the left plot of Figure 6, the visual syllable /da/ was identified as /d/ about as often as it was identified as /g/. The same result occurred for the syllable /ga/. What is important for our purposes, however, is that the participants respond to both visual /da/ and visual /ga/ about twice as often with the alternative /da/ than with the alternative /ga/. This observation is novel because previous investigators had not tested these unimodal visual conditions. This result offers a new explanation of why an auditory /ba/ paired with a visual /ga/ produces the response /da/. Apparently, people are biased to report /d/ over /g/ because /d/ occurs much more often than /g/ in spoken language (Denes, 1963).

Much to our dismay, however, we failed to find a strong McGurk fusion effect. Neither a visual /da/ or /ga/ biased the response to auditory /ba/. For whatever reason, the auditory information tended to dominate the perceptual judgment. One possibility is that observers were not permitted to make other responses, such as /va/ or /tha/, which are frequently given to these conflicting syllables. Another possibility is that the quality of the natural auditory speech was much greater than the quality of the synthetic visual speech. To solidify our interpretation of the prototypical fusion effect, however, we will have to observe the traditional McGurk effect in the same situation in which a bias for /d/ over /g/ is observed.

Our experiment shows that perceivers are biased to perceive /d/ rather than /g/, perhaps because the alveolar segment occurs more frequently than the velar one (Denes, 1963). Participants have difficulty perceiving differences between visual /d/ and visual /g/, and tend to label both of these segments as /d/. One surprising outcome of this experiment was that there were relatively few McGurk Illusions. We believe there are several explanations for this finding. First, participants were limited to the judgments /b/, /d/, and /g/ and in many cases, auditory /b/ and visual /d,g/ produce /v/ and /th/ as responses (Massaro, 1998). Second, our natural auditory speech was long-duration citation speech, which was necessarily very high quality. Third, our visual speech was computer animated, which is slightly less intelligible than natural speech (Massaro, 1998, Chapter 13). The contribution of visible speech (i.e., the McGurk effect) will tend to be smaller as the quality of the auditory speech is increased and the quality of the visual speech is decreased (see Sekiyama, 1998).

9. INTEGRATING WRITTEN TEXT AND SPEECH

An important issue concerns whether sensory fusion of auditory and visual inputs is limited to speech stimuli. We carried out a series of experiments that compared the perception of auditory speech paired with visible speech versus auditory speech paired with written language. The results from this study can help inform us about which theories of bimodal speech perception are viable. Knowing or seeing the words to a rock song while hearing the song creates the impression of hearing a highly intelligible rendition of the words. Without this knowledge of the words, the

listener cannot make heads or tails of the message. The first demonstration of this kind that we know of was by John Morton, who played a song by the Beatles. Members of the audience could not perceive clearly the words of the song until they were written on the viewing screen. Another variation on this type of illusion is the so-called phonemic restoration effect in which we claim to hear the /s/ in the word legislatures even though it is replaced by a cough, a buzz or even a pure tone (Warren, 1970).

Frost et al. (1988) found that when a spoken word is masked by noise having the same amplitude envelope, subjects report that they hear the word much more clearly when they see the word in print at the same time. This result supports the idea that written text can influence our auditory experience. To show effects of written information on auditory judgment at the perceptual level, Massaro et al. (1988) compared the contribution of lip-read information to written information. Subjects were instructed to watch a monitor and listen to speech sounds. The sounds were randomly selected from nine synthetic speech sounds along a /ba/ to /da/ continuum. On each trial, the subjects were presented with either (1) a visual representation of a man articulating the sound /ba/ or /da/, or (2) a written segment BA or DA. Although there was a large effect of visible speech, there was only a small (but significant) effect of the written segments on the judgments. Both the speech and written-text conditions were better described by the FLMP than by an alternative additive or single channel model.

To better test for the possible influence of text on speech perception, our study tested whether we could obtain a larger effect of written text. Given that letters of the alphabet have a strict spelling-to-sound mapping and are pronounced automatically and effortlessly, the letters B and D were used. The letter sequences BA and DA are not necessarily pronounced /ba/ and /da/. The letters B and D are only pronounced /bi/ and /di/ – as they are named in the alphabet.

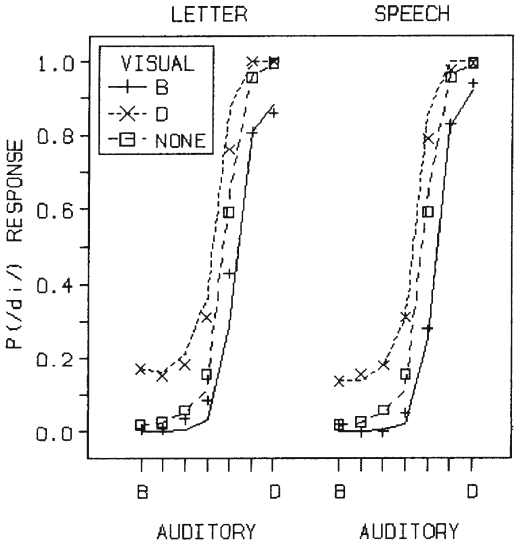

Figure 8. Observed (points) and predicted (lines) by the FLMP probability of a /di/ response as a function of the auditory and visual stimuli for the letter and word conditions.

Nine participants from the University of California, Santa Cruz, were tested. This experiment employed a within-subjects expanded factorial design. There were seven auditory levels between the syllables /bi/ and /di/. There were four visual levels – two letter conditions (the letters B and D) and two speech conditions (the visual syllables /bi/ and /di/), for a total of 39 trial types. The observers were specifically instructed to both watch the screen and listen for a sound and to report what they heard. On those trials in which only a visual stimulus was presented, they were to report the visual stimulus. On each trial, subjects identified stimuli as B or D by typing the appropriately marked keys. The stimuli were presented in 6 blocks of the 39 trial types, for a total of 234 trials per session. The test conditions were selected at random without replacement. A practice block of 10 trials occurred prior to the experimental trials. Subjects had approximately three seconds to respond on each trial. Each subject participated on two days with two sessions per day. Thus there were 24 observations per subject per condition. The dependent measure was the proportion of /di/ judgments.

Figure 8 displays the average results for the letter and speech conditions. The proportion of /di/ responses as a function of the seven auditory levels is shown with the visual B or D stimulus or no visual information (NONE) as the curve parameter. The average proportion of /di/ responses increased significantly as the auditory syllable went from the most /bi/-like to the most /di/-like level. There was also a significant effect on the proportion of /di/ responses as a function of the visual stimulus; there were fewer /di/ responses for visual B than for a visual D. The interaction of these two variables was also significant: the influence of the visual

variable was larger at the more ambiguous regions of the auditory continuum. Not shown in Figure 8, the visual alone trials gave essentially perfect performance for both the speech and letters.

The result of interest here is the difference between the visible speech and the letter conditions. As can be seen in the figure, the visual effect was substantial and of similar size for the letter and for the speech condition. The FLMP was fit to the average proportion of /di/ responses for each of the nine participants. The FLMP gave a very good description of the observations. Thus, it appears that written text, as well as visible speech, can influence our auditory experience and that the FLMP accounts for both types of influence. Given these results, it is important to explore a number of important variables to test whether there are any qualitative differences between the integration of written text or visual speech with auditory speech. These conclusions hold up in additional studies with a larger number of response alternatives and without visual-alone trials. Unless we are completely wedded to the idea that speech is special, an influence of written language on our perceptual experience should not be surprising.

10. WORD RECOGNITION

We believe that visual input is a strong influence on spoken language perception in face-to-face communication. An important issue is a possible concern that research with syllables might not generalize to words and sentences. Experimental results with syllables should be compared with those with words and sentences to determine if the same model can be applied to these different test items. To move beyond syllables, we assessed the processing of auditory and visual speech at the word level. Settling on an experimental task for evaluation is always a difficult matter. Even with adequate justification, however, it is important to see how robust the conclusions are across different tasks. In one experiment, we used a gating task, in which successively longer portions of a test word are presented (Grosjean, 1980; Munhall & Tohkura, 1998).

Following our theoretical framework, we tested observers under auditory, visual, and bimodal conditions. The test words were monosyllabic CVCs. Eight gating durations were tested. We expected performance to improve with increases in the duration of both the auditory and visual components of the test word. We expect the auditory information to be more informative than visual, but most importantly bimodal performance should be significantly better than either unimodal condition.

The results were as expected. The FLMP and competing models were fit to both the accuracy of identification of the test words, as well as to the identification of the individual segments of the word. The FLMP gave the best description of the results. This extension of our paradigm to words is an important test of how well our theoretical framework applies beyond the syllable level.

11. PERCEPTION OF PARALINGUISTIC INFORMATION

Laypersons and researchers agree that communication involves much more than simply the linguistic message. Paralinguistic as well as linguistic information is necessary for optimal communication and understanding. Our research has been directed primarily at the linguistic dimensions of speech but it is important that we explore the paralinguistic ones in parallel. In collaboration with Jonas Beskow from KTH, we studied the joint influence of F0, loudness, eye widening and eyebrow movements on the perception of stress. A stressed word tends to be somewhat longer in duration, somewhat greater amplitude, and somewhat higher in pitch. Cave et al. (1996) found that rapid rising-falling eyebrow movements occurred with F0 rises about 70% of the time. There is some other unpublished evidence that eye widening might occur on stress words. Using our factorial design methodology, we manipulate these sources of information independently of one another to determine their relative contributions to the perception of stress. Participants were asked to indicate the degree to which a given word in a sentence was stressed. The analyses included tests of formal models of speech perception and language processing (see Massaro, 1996, 1998).

In one experiment carried out in collaboration with Jonas Beskow, participants were given sentences of the form noun-verb-noun, and asked to indicate whether the first or last word was emphasized. Four independent variables were orthogonally varied in a factorial design. Using an animated talking head and synthetic speech, the eyebrows were raised during either the first or last word, the eyes were widened during either the first or last word, the amplitude was increased during either the first or last word, and the pitch was raised during the first or last word or was held constant during both of the words. For the all four variables, the noun that did not receive emphasis for a given variable was set at the neutral value for that variable. For example, if the eyebrows are raised during the last noun, the eyebrows are kept still during the first noun of that sentence. For the F0 variable, there was a third condition in which the pitch could be kept neutral during both nouns. This gives a 2 by 2 by 2 by 3 factorial design for a total of 24 experimental conditions. The experiment consisted of 20 noun-verb-noun sentences, each presented under each of the 24 experimental conditions, yielding a total of 480 trials. Stimuli were presented in random order, with a short break half way through the experiment. Nine subjects were tested in the experiment, which took place in front of a computer screen, where stimuli were presented by the animated face. Subjects entered their responses using the mouse by clicking on the noun that they perceived to be more stressed in a text representation of the sentence that was presented below the face.

Figure 9 presents the results of the experiment in terms of the proportion of times the first noun in the sentence was categorized as stressed. As can be seen in the figure, although all four independent variables had some influence on the judgments, the amplitude of the noun was the most influential factor.

This situation is slightly more complicated than the prototypical experiment and it will be worthwhile to describe how the FLMP and alternative models can be

applied to the results. It is assumed that perceivers evaluate and integrate a variety of cues to perceive word stress. As a working hypothesis, it is assumed that the perceivers evaluate the four cues eyebrow raising (ER), eye widening (EW), amplitude increase (AI), and F0 raising (F0) as cues to stress. A stressed word S(word), can be represented by

$$S(word): ER \ \& \ EW \ \& \ AI \ \& \ F0 \ \& \ O$$

where O corresponds to other potential cues that are not being systematically manipulated in the experiment. An unstressed word, ~S(word), would be represented by

$$\sim S(word): \sim ER \ \& \ \sim EW \ \& \ \sim AI \ \& \ \sim F0 \ \& \ \sim O$$

It is assumed that the perceiver determines the degree of stress and unstress for both the first and last nouns in the sentences. The probability of choosing the first noun as stressed is determined by the degree to which the first noun is stressed and the degree to which the second noun is unstressed. It is possible to eliminate the O term corresponding to the other cues since they do not change under the different conditions.

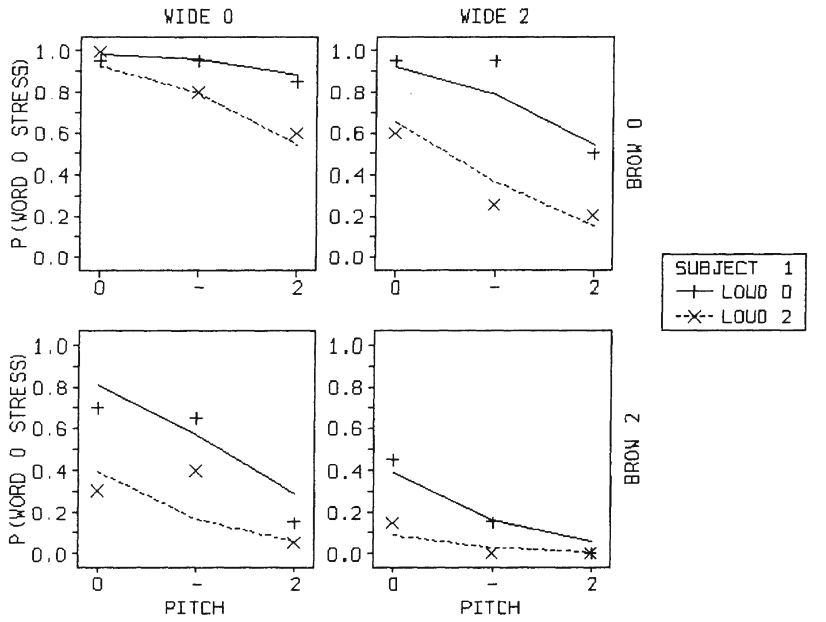

Figure 9. Proportion of times the first word was categorized as stressed, P(Word 0 Stress), as a function of whether the eyebrows were raised during either the first (0) or last word (2), the eyes were widened during either the first or last word, the amplitude was increased during either the first or last word, and the pitch was raised during the first or last word or was held constant (-) during both of the words.

Because the properties of the second noun are simply the unmarked complements of those in the first noun, it is sufficient to simply determine the amount of

support for the first noun being stressed or unstressed. The degree to which the first noun is stressed is therefore directly related to having the four cues marking stress in the first noun. If a stimulus cues matches its representation in the prototype for stress, we give it a feature value of f_i; if it mismatches the representation for stress, we give it a feature value of $(1 - f_i)$. The index i stands for the four cues manipulated in the experiment. The value f_i should be greater than .5 if the stress marking is a functional cue to stress and its value can be interpreted as its cue strength. The F0 neutral condition in which both nouns have neutral stress should be given the value .5, which represents the complete absence of support in one direction or the other. Imposing this constraint also insures that the estimated parameter values given the model fit are identifiable or unique (see Crowther et al., 1995; Massaro, 1998, Chapter 11).

The probability of a stressed judgment is equal to the support for a stressed alternative divided by the sum of support for stress and for unstressed alternatives.

$$P(S) = \frac{S(s)}{S(s) + S(\sim s)} \qquad (2)$$

where s(s) and S(~s) are equal to the total support for the stressed and unstressed alternatives, respectively. As an example, consider the case in which the first noun has raised eyebrows, no eye widening, no amplitude change, and a higher F0. The support for the stressed alternative, S(s), would be

$$S(s) = f_{ER} \,\&\, (1 - f_{EW}) \,\&\, (1 - f_{AI}) \,\&\, f_{F0})$$

The support for the unstressed alternative, S(~s), would be

$$S(\sim s) = (1 - f_{ER}) \,\&\, f_{EW}) \,\&\, f_{AI} \,\&\, (1 - f_{F0})$$

Integration models differ primarily in terms of how conjunction (&) is implemented in determining the total support. Conjunction is multiplicative in the FLMP and additive in the additive model of perception (AMP). All other aspects of the models are equivalent.

Finally, it is probably likely that the sentence context itself differentially supports one of the nouns as the stressed one. If this is the case, it is necessary to assume an additional source of information from the sentence context. The fit of models against these results tests not only the models but also determines which are the informative cues and how they are used in sentence prosody perception. As in our other studies, the FLMP gave the best description of performance.

12. PERCEIVING EMOTION FROM THE FACE AND THE VOICE

We have also carried out a set of studies of how people evaluate and integrate information in the recognition of emotion. Our research has varied multiple sources of paralinguistic information, facial expressions and vocal cues, to investigate perception of a speaker's emotion. We have undertaken a set of studies to assess what properties of the face and voice people actually use to infer emotional content. As for speech, we operate under the assumption that multiple sources of information are

also used to perceive a talker's emotion. A variety of signals are used in addition to the verbal content of the speech. The emotion may be interpreted in different ways depending on the voice quality, facial expression, and body language used. In order to study the degree to which emotional sources of information are used, it is important to define these sources and then determine how they are used. In our research, two sources of emotional information, facial expressions and vocal cues, are chosen to be as analogous to the speech situation as possible.

In previous research (Ellison & Massaro, 1996), we used an expanded-factorial design where the affective categories happy and angry were chosen because they represent two of the basic categories of emotion. We chose two features that seem to differ somewhat in happy and angry faces. An important criterion for manipulating two features is that they can be varied independently of one another. Thus, varying one cue in the upper face and one cue in the lower face was an ideal solution. Five levels of the upper face and five levels of the lower face were factorially combined, along with the ten half-face conditions. The feature values were obtained by comparison to features displayed in exemplar photographs in Ekman & Friesen (1975). The features varied were brow deflection (BD) and mouth-corner deflection (MD). BD was varied from fully elevated and arched for a prototypically happy affect to fully depressed and flattened for a prototypically angry affect. MD was varied from fully curled up at corners for a prototypically happy affect to fully curled down at corners for a prototypically angry affect. The FLMP gave a good fit to individual fits in binary judgments, judgments with six response alternatives, and in a rating task in which with instructions to rate the affect on a scale from 1 to 9. The independent variables influenced performance in the same manner in all tasks.

In another experiment, we examined how emotion is perceived by using facial and vocal cues of a speaker (Massaro & Egan, 1996). Three levels of facial affect were presented using a computer-generated face. Three levels of vocal affect were obtained by recording the voice of a male amateur actor who spoke a semantically neutral word in different simulated emotional states. These two independent variables were presented to participants of the experiment in all possible permutations, i.e. visual cues alone, vocal cues alone and visual and vocal cues together, which gave a total set of 15 stimuli. The participants were asked to judge the emotion of the stimuli in a two-alternative forced choice task (either HAPPY or ANGRY). The results indicate that participants evaluate and integrate information from both modalities to perceive emotion. The influence of one modality was greater to the extent that the other was ambiguous (neutral). The FLMP fit the judgments significantly better than an additive model, which weakens theories based on an additive combination of modalities, categorical perception, and influence from only a single modality.

We have extended these studies to the other basic emotions, including fear, surprise, disgust, and sadness. These studies not only tested how well the theoretical framework holds up with these other emotions, they also provide an understanding of which facial features are informative in conveying these emotions. As an example, there are cues in the upper and lower face that can be varied to create surprised and fearful emotions. A surprised face is characterized by a wide-eyed look with very raised eyebrows and an open mouth. A fearful face, on the other

hand, has a somewhat less raising of the eyebrows and less opening of the mouth. Given that the FLMP provides a good description of performance for the full set of basic emotions, the perception of emotion appears to be well described by our theoretical framework.

13. TESTS OF DYNAMIC INFORMATION IN VISIBLE SPEECH PERCEPTION

It has been proposed that speech perception is based on higher order dynamic properties of the spoken language. Point-light displays have been used as support for this idea. Rosenblum & Saldana (1996, 1998) using hybrid visual-auditory tests have argued that a point-light display using 28 lights attached to the face is effective as a visual speech stimulus. The authors claim that point-light displays were effective for the /b/-/v/ distinction. However, they did not convincingly demonstrate that it was as informative as a real moving face. We have tested this idea by comparing point-light displays to the perception of displays of our synthetic talker (Cohen, et al., 1996). This test provides a more extensive assessment of how much information can be transmitted by point-light displays, as well as a better control to equate the stimuli for the two conditions. Rosenblum & Saldana's original study (1) used only /ba/ and /va/ tokens, and (2) the facial and point-light displays were made separately under differing conditions, which makes it difficult to say whether the stimuli had equivalent articulations.

The experimental procedure was identical to that used in Massaro (1998, chapter 13), except that the synthetic point-light face replaced the natural one. The point-light display was made by putting tiny spheres on 28 of the polygon vertices of the face, positioned on the face, lips, teeth, and tongue identically to those used by Rosenblum & Saldana. Performance on the initial consonant visemes of the test words was significantly worse for the point-light display compared to our synthetic facial display. Analysis of performance on final consonant and vowel visemes also showed a significant advantage for the synthetic face over the point-light display. It is evident that with equivalent display geometries, the point-light display does provide some valuable information for speech reading, although performance was significantly worse than the full synthetic facial display. Although it might be the case that kinematic properties are informative for speech reading, our results indicate that they are not sufficient.

14. CONCLUSION

If the reader has persisted to this stage, it is hoped that they have obtained a good understanding of our approach to the study of speech perception and understanding. Like multimodality, which provides several roads to understanding speech, our framework exploits multiple ways of knowing about how this magnificent and magical process works. One goal was to illustrate the value of an information-processing framework for the study of multimodality in spoken language understanding. Our theoretical framework and the FLMP were used to impose coherence on a complex set of experiments and results. I look forward to the future in which

the concept of multimodality is being applied fruitfully in theoretical development, empirical research, applied situations, and commercial endeavors.

15. REFERENCES

Campbell, C.S. & D.W. Massaro. "Perception of visible speech: influence of spatial quantization", *Perception, 26*, 627-644, 1997.

Cave, C., I. Guaitella, R. Bertrand, S. Santi, F. Harlay & R. Espesser. "About the relationship between eyebrow movements and FO variations". *Proceedings of the International Conference on Spoken Language Processing* (pp. 2175-2178), Wilmington: University of Delaware, 1996.

Cohen, M.M., R.L. Walker & D.W. Massaro. "Perception of synthetic visual speech". In: D.G. Stork & M.E. Hennecke (Eds.), *Speechreading by humans and machines* (pp. 153-168). New York: Springer, 1996.

Cole, R., T. Carmell, P. Connors, M. Macon, J. Wouters, J. deVilliers, A. Tarachow, D.W. Massaro, M.M. Cohen, J. Beskow, J. Yang, U. Meier, A. Waibel, P. Stone, G. Fortier, A. Davis, C. Soland. "Intelligent Animated Agents for Interactive Language Training". *Proceedings of Speech Technology in Language Learning*. Stockholm, Sweden, 1998.

Crowther, C.S., W.H. Batchelder & X. Hu. "A measurement-theoretical analysis of the Fuzzy Logical Model of Perception". *Psychological Review, 102*, 396-408, 1995.

Cutting, J.E., N. Bruno, N.P. Brady & C. Moore. "Selectivity, scope, and simplicity of models: A lesson from fitting judgments of perceived depth". *Journal of Experimental Psychology: General*, 121, 364-381, 1992.

Denes, P.B. "On the statistics of spoken English". *Journal of the Acoustical Society of America, 35*, 892-904, 1963.

Diehl, R.L. & K.R. Kluender. "On the categorization of speech sounds". In: S. Harnad (Ed.), *Categorical perception* (pp. 226-253). Cambridge: Cambridge University Press, 1987.

Diehl, R.L. & K.R. Kluender. "On the objects of speech perception". *Ecological Psychology,* 121-144, 1989.

DeYoe, E.A. & D.C. Van Essen. "Concurrent processing streams in monkey visual cortex". *Trends in Neurosciences, 11,* 219-226, 1988.

Ekman, P. & W. Friesen. *Pictures of facial affect*. Palo Alto, CA: Consulting Psychologists Press, 1975.

Ellison, J.W. & D.W. Massaro. "Featural evaluation, integration, and judgement of facial affect", *Journal of Experimental Psychology: Human Perception and Performance, 2*, 213-226, 1997.

Fowler, C.A. "Listeners do hear sounds, not tongu". *Journal of the Acoustical Society of America, 99,* 1730-1741, 1996.

Frost, R., B.H. Repp & L. Katz. "Can speech perception be influenced by simultaneous presentation of print?" *Journal of Memory and Language, 27*, 741-755, 1988.

Green, K.P. "The use of auditory and visual information during phonetic processing: Implications for theories of speech perception". In: Campbell, R., B. Dodd & D. Burnham (Eds.), *Hearing by Eye II* (pp. 3-25). East Sussex, UK: Psychology Press Ltd, 1998.

Grosjean, F. "Spoken word recognition processes and the gating paradigm". *Perception & Psychophysics, 28*, 267-283, 1980.

Kass, R.E. & A.E. Raferty. "Bayes factors". *Journal of the American Statistical Association, 90,* 773-795, 1995.

Liberman, A.M. & I.G. Mattingly. "The motor theory of speech perception revised". *Cognition, 21*, 1-33, 1985.

Lisker, L. "Rabid vs rapid: A catalog of acoustic features that may cue the distinction". *Haskins Laboratories, Status Report on Speech Research, SR-54*, 127-132, 1978.

Massaro, D.W. *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1987.

Massaro, D.W. Multiple book review of *Speech perception by ear and eye: a paradigm for psychological inquiry*, by D.W. Massaro. *Behavioral and Brain Sciences, 12*, 741-794, 1989.

Massaro, D.W. "Integration of multiple sources of information in language processing". In: T Inui & J.L. McClelland (Eds.), *Attention and Performance XVI: Information integration in perception and communication* (pp. 397-432). Cambridge, MA: MIT Press, 1996.

Massaro, D.W. *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. MIT Press: Cambridge, MA, 1998.

Massaro, D.W. & M.M. Cohen. "Evaluation and integration of visual and auditory information in speech perception". *Journal of Experimental Psychology: Human Perception and Performance, 9*, 753-771, 1983.

Massaro, D.W. & M.M. Cohen. "Perception of synthesized audible and visible speech". *Psychological Science, 1*, 55-63, 1990.

Massaro, D.W. & M.M. Cohen. "Speech Perception in Perceivers with Hearing Loss: Synergy of Multiple Modalities". *Journal of Speech, Language, and Hearing Research,* 42: 21-41, 1999.

Massaro, D.W. & P.B. Egan. "Perceiving affect from the voice and the face". *Psychonomic Bulletin and Review, 3,* 215-221, 1996.

Massaro, D.W. & D. Friedman. "Models of integration given multiple sources of information", *Psychological Review, 97(2),* 225-252, 1990.

Massaro, D.W. & D.G. Stork. "Speech recognition and sensory integration". *American Scientist, 86,* 236-244, 1998.

Massaro, D.W., M.M. Cohen & P.M.T. Smeele. "Cross-linguistic Comparisons in the Integration of Visual and Auditory Speech," *Memory and Cognition, 23,(1)* 113-131, 1995.

Massaro, D.W., M.M. Cohen & L.A. Thompson. "Visible language in speech perception: Lipreading and reading," *Visible Language, 22,* 9-31, 1988.

Massaro, D.W., M.M. Cohen, C.S. Campbell & T. Rodriguez. "Bayes factor of model selection validates FLMP". *Psychonomic Bulletin & Review*, 8, 1-17, 2001.

Massaro, D.W., M. Tsuzaki, M.M. Cohen, A. Gesi & R. Heredia. "Bimodal Speech Perception: An Examination across Languages", *Journal of Phonetics, 21*, 445-478, 1993.

Mattingly. I.G. & M. Studdert-Kennedy, (Eds). *Modularity and the motor theory of speech perception.* Hillsdale, NJ: Lawrence Erlbaum, 1991.

McGurk, H. & J. MacDonald. "Hearing lips and seeing voices". *Nature, 264,* 746-748, 1976.

Munhall, K.G. & Y. Tohkura. "Audiovisual gating and the time course of speech perception". *Journal of the Acoustical Society of America, 104,* 530-539, 1998.

Myung, I.J. & M.A. Pitt. "Applying Occam's razor in modeling cognition: A Bayesian approach". *Psychonomic Bulletin & Review, 4,* 79-95, 1997.

Oerlemans, M. & P. Blamey. "Touch and auditory-visual speech perception". In: Campbell, R., B. Dodd, & D. Burnham (Eds), *Hearing by Eye II* (pp. 267-281). East Sussex, UK: Psychology Press Ltd, 1998.

Palmer, S.E. *Vision Science: Protons to Phenomenology*. Cambridge, MA: MIT Press, 1999.

Pitt, M.A. & J. M. McQueen. "Is Compensation for Coarticulation Mediated by the Lexicon?" *Journal of Memory and Language, 39*, 347-370, 1998.

Rosenblum, L.D. & H.M. Saldana. "An audio-visual test of kinematic primitives for visual speech perception". *Journal of Experimental Psychology: Human Perception and Performance, 22*, 318-331, 1996.

Rosenblum, L.D. & H.M. Saldana. "Time-varying information for visual speech perception". In: Campbell, R., B. Dodd, & D. Burnham (Eds), *Hearing by Eye II* (pp.61-81). East Sussex, UK: Psychology Press Ltd, 1998.

Schindler, R.A. & M.M. Merzenich. *Cochlear Implants*. New York: Raven, 1985.

Schwartz, J., J. Robert-Ribes, & P. Escudier."Ten years after Summerfield: A taxonomy of models for audio-visual fusion in speech perception". In: Campbell, R., B. Dodd & D. Burnham (Eds), *Hearing by Eye II* (pp. 85-108). East Sussex, UK: Psychology Press Ltd, 1998.

Sekiyama, K. "Face or voice? Determinant of compellingness to the McGurk effect". *Proceedings of AVSP'98*. Terrigal – Sydney, Australia, 1998.

Tyler, R.S., J.M. Opie, H. Fryauf-Bertschy & B.J. Gantz. "Future directions for cochlear implants". *Journal of Speech-Language Pathology and Audiology, 16,* 151-164, 1992.

Warren, R.M. "Perceptual restoration of missing speech sounds". *Science, 167*, 392-393, 1970.

16. AFFILIATION

Dr. Dominic W. Massaro
Department of Psychology
University of California
Santa Cruz, CA 95064 USA
URL: http://mambo.ucsc.edu/psl/dwm/

17. ACKNOWLEDGEMENT

The research and writing of the paper were supported by grants from National Science Foundation (Grant No. CDA-9726363, Grant No. BCS-9905176, Grant No. IIS-0086107), Public Health Service (Grant No. PHS R01 DC00236), Intel Corporation, the University of California Digital Media Program, and the University of California, Santa Cruz.