From Multisensory Integration to Talking Heads and Language Learning

Dominic W. Massaro

Department of Psychology, University of California, Santa Cruz,

Santa Cruz, CA 95060 U.S.A.

1-831-459-2330

FAX 1-831-459-3519

massaro@fuzzy.ucsc.edu

Send correspondence to:

Dr. Dominic W. Massaro

Department of Psychology

University of California

Santa Cruz, CA 95064 USA

work: 831-459-2330

FAX: 831-459-3519

email: massaro@fuzzy.ucsc.edu

URL: http://mambo.ucsc.edu/psl/dwm/

Chapter to be published in Handbook of Multisensory Processes

Edited by: Gemma Calvert, Charles Spence and Barry E. Stein

Cambridge, MA: MIT Press.

Introduction

Given this handbook of multisensory processes, we learn that perceptual and behavioral outcomes are influenced by simultaneous inputs from several senses. In this chapter, we present theoretical and empirical research on speech perception by eye and ear, and address the question of whether speech is a special case of multisensory processing. Our conclusion will be that speech perception is indeed an ideal or prototypical situation in which information from the face and voice is seamlessly processed to impose meaning in face-to-face communication.

Scientists are often intrigued with questions whose answers necessarily pigeonhole some striking phenomenon. One question about language is whether speech perception is uniquely specialized for processing multisensory information or whether it is simply a prototypical instance of crossmodal processing that occurs in many domains of pattern recognition. Speech is clearly special, at least in the sense that (as of now) only we big-mouthed biped creatures can talk. Although some chimpanzees have demonstrated remarkable speech perception and understanding of spoken language, they seem to have physiological and anatomical constraints that preclude them from assuming bona fide interlocutor status (Lieberman, 2000; Savage-Rumbaugh et al., 1998). An important item of debate, of course, is whether they also have neurological, cognitive or linguistic constraints that provide an impenetrable barrier for language use (Arbib, 2002). We begin with a short description of the idea that speech is special.

Speech is Special (SiS)

Noam Chomsky (1980) envisioned language ability as dependent on an independent language organ (or module), analogous to other organs such as our digestive system. This organ follows an independent course of development in the first years of life and allows the child to achieve a language competence that cannot be elucidated in terms of traditional learning theory. This mental organ, responsible for the human language faculty and our language competence, matures and develops with experience, but the mature system does not simply mirror this experience. The language user inherits rule systems of highly specific structure. This innate knowledge allows us to acquire the rules of the language, which cannot be induced from normal language experience because (advocates argue) of the paucity of the language input. The data of language experience

are so limited that no process of induction, abstraction, generalization, analogy, or association could account for our observed language competence. Somehow, the universal grammar given by our biological endowment allows the child to learn to use language appropriately without learning many of the formal intricacies of the language. Developmental psychologists, however, are finding that infants are exposed to a rich sample of their mother tongue, and they are highly influenced by this experience (e.g., Marcus, 2000; Saffran et al., 1996, 1999). Moreover, the frequency and ordering of speech inputs have immediate and strong influences on perceptual processing and these influences are similar for speech and nonspeech (Aslin et al., 1998; Gomez & Gerken, 2000). Linguists are also documenting that the child's language input is not as sparse as the nativists had argued (Pullum & Scholz, , 2002).

Although speech has not had a spokesperson as charismatic and influential as Chomsky, a similar description is given for speech perception. In addition, advocates of the special nature of speech are encouraged by Fodor's (1983) influential proposal of the modularity of mind. Some of our magnificent capabilities result from a set of innate and independent input systems, such as vision, hearing, and language (Fodor, 1983, 2000). Speech-is-special theorists now assume that a specialized biological speech module is responsible for speech perception (Liberman & Mattingly, 1985; Mattingly & Studdert-Kennedy, 1991; Trout, 2001). Given the environmental information, the speech module analyzes this information in terms of possible articulatory sequences of speech segments. The perceiver of speech uses his or her own speech-motor system to achieve speech recognition.

In some ways, it is ironic the multisensory processing should serve as a touchstone for advocates that speech is special, and for articulatory mediation of speech perception. It all began with the McGurk's discovery (McGurk & MacDonald, 1976), which has obtained widespread attention in many circles of psychological inquiry and cognitive science. The classic McGurk effect involves the situation in which an auditory /ba/ is paired with a visible /ga/ and the perceiver reports *hearing* /da/. The reverse pairing, an auditory /ga/ and visual /ba/, tends to produce a perceptual judgment of *hearing* /bga/. It was apparently unimaginable at that time that this type of crossmodal influence would occur in other domains. As documented in several chapters in this

handbook, multisensory integration is the rule rather than the exception (see Lederman & Klatzky, this volume; Làdavas & Farnè, this volume). As an example, both sound and sight contribute to our localization of an event in space, and the visual input can distort our experience such as when we hear the puppet's coming from the puppet rather than the ventriloquist. This similarity to other domains dictates a more general account of sensory fusion and modality specific experience rather than one unique to speech perception.

It should be noted, however, that the perceiver might have a unimodal experience even though multisensory integration contributed to the experience. This is clearly an unintuitive outcome, and one requiring explanation. Speech information from the auditory and visual modalities provides a situation in which the brain combines both sources of information to create an interpretation that is easily mistaken for an auditory one. An exactly analogous outcome is found when our perceived taste is influenced by smell, as in the pleasurable taste of coffee accompanied by smell. If the nose is pinched, the taste becomes either indistinguishable or bitter (see Stevenson & Boakes, this volume). For spoken language, we that believe we hear the speech because perhaps audition is the most informative modality for spoken language. A caveat is, therefore, that we cannot trust a modality-specific experience as implying that only that modality played a role.

We turn to a short review of existing theories of speech perception before turning to relevant empirical evidence. The powerful influence that visible speech has been shown to have in face-to-face communication speaks to both traditional and extant theoretical accounts. The influence of several sources of information from several modalities provides a new challenge for theoretical accounts of speech perception. Most theories were developed to account for the perception of unimodal auditory speech, and it is not always obvious how they would account for the positive contribution of visible speech.

Theories of Speech Perception

Psychoacoustic Accounts

One class of theory seems to be either contradicted or at least placed outside the domain of bimodal speech perception. Psychoacoustic accounts of speech perception are grounded in the idea that speech is nothing more than a complex auditory signal, and its processing can be understood by the psychophysics of complex sounds, without any

reference to language specific processes. This chapter reinforces the conclusion that a psychoacoustic account of speech perception is not sufficient because speech perception is not strictly a function of auditory information. Advocates of the psychoacoustic account have modified their stance accordingly and now acknowledge the influence of visible speech (for example, Diehl & Kluender, 1987) They have not specified, however, how visible speech makes its contribution but it would appear that visible speech would somehow have to be secondary to audible speech. If psycholacoustic theorists propose that visible speech need not be secondary in their framework, then we might ask what is uniquely psychoacoustic about it.

Motor Theory

The motor theory assumes that the perceiver uses the sensory input to best determine the set of articulatory gestures that produced this input (Liberman & Mattingly, 1985; Mattingly & Studdert-Kennedy, 1991). The main motivation and support for this theory is that phoneme perception is putatively more easily predicted on the basis of articulation than in terms of acoustic cues. Speech scientists learned that there did not appear to be a one-to-one correspondence between a set of acoustic properties and a phonetic segment. On the other hand, the phonetic segment could be more adequately described in terms of articulation. The best known example is the difference between /di/ and /du/. The onset of these two syllables have very different acoustic properties but have similar articulatory gestures, which involves a constriction of the tongue against the alveolar ridge of the hard palate. The syllables with different vowels differ in their sound even at onset because the consonant and vowel are coarticulated. Thus motor theory appeared to solve the mapping problem from stimulus to percept by viewing articulation as mediating representation.

According to motor theory, the inadequate auditory input is assessed in terms of the articulation, and it is only natural that visible speech could contribute to this process. The motor theory is consistent with a contribution of visible speech because visible speech can be considered to be an integral part of the sensory input reflecting the talker's articulatory gestures. In a related proposal, Robert-Ribes et al. (1995a) advocate an amodal motor representation to account for the integration of audible and visible speech.

The motor theory has not been sufficiently formalized, however, to account for the vast set of empirical findings on the integration of audible and visible speech.

The motor theory found new life in Fodor's notion of modularity in which an input system operates in an encapsulated manner. Speech perception is viewed as a module with its own unique set of processes and information. As stated succinctly by Liberman (1996, p. 29), "the phonetic module, a distinct system that uses its own kind of signal processing and its own primitives to form a specifically phonetic way of acting and perceiving." To me, this statement implies that not only the information but the information processing should be qualitatively different in the speech domain than in other domains of perceptual and cognitive functioning. We will see, however, that this expectation does not hold up to experimental tests. For example, perceiving emotion from the face and voice follows the same processing algorithm as speech perception.

As we have argued, it is very difficult to determine the representation medium in which integration occurs. We see no reason, however, to postulate a motor representation for integration. Integration occurs in a variety of other domains, such as object recognition, that involves no analogous motor medium.

Direct Perception

In contrast to the motor theory and consistent with our view, the direct perception theory assumes that speech perception is not special (Fowler, 1996; this handbook). Thus, although gestures are the objects of speech perception, the speech motor system does not play a role. Furthermore, speech perception is just one of many different perceptual domains in which direct perception occurs. The direct perception theory states that persons directly perceive the causes of sensory input. In spoken language, the cause of an audible-visible speech percept is the vocal tract activity of the talker. Accordingly, it is reasoned that visible speech should influence speech perception because it also reveals the vocal-tract activity of the talker. Speech perceivers therefore obtain direct information from integrated perceptual systems from the flow of stimulation provided by the talker (Best, 1993). The observed influence of visible speech is easily predicted by this theory because visible speech represents another source of stimulation, providing direct information about the gestural actions of the talker. However, we know of no convincing evidence for the gesture as the primary object of speech perception (see Massaro, 1998b,

Chapter 11). For now, it seems most parsimonious to assume that the objects of speech perception are relatively abstract symbols (Nearey, 1992).

On the basis of just this short review of extant theories of speech perception, it is apparent that they are stated in verbal rather than quantitative form. Although no one can deny that a qualitative fact is more informative than a quantitative one, qualitative theories do not seem to be sufficiently precise to be distinguished from one another. Very different theories make very similar predictions. Some quantitative refinement of the theories is usually necessary to create a chance for falsification and strong inference (Platt, 1964; Popper, 1959). Therefore, our strategy has been to quantify and test a family of specific models that represent the extant theories and also other reasonable alternatives (Massaro, 1987b, 1996, 1998b).

Pattern Recognition

We envision speech perception as a prototypical instance of pattern recognition. The term pattern recognition describes what is commonly meant by recognition, identification, or categorization. Although these terms have different meanings, they are all concerned with roughly the same phenomenon. Recognition means re-cognizing something we experienced previously. Identification involves mapping a unique stimulus into a unique response. Categorization means placing several noticeably different stimuli into the same class. For example, a child perceives a dog, recognizes it as a dog she has seen before, identifies it as Fido, and categorizes it as a dog. Recognition, identification, and categorization appear to be central to perceptual and cognitive functioning (Quinn, 2002). They entail the same fundamental processes to allow a person, given some input, settles on one of a set of alternative interpretations. Pattern recognition has been found to be fundamental in such different domains as depth perception, playing chess, examining X-rays, and reading text (Quinn, 2002). It involves similar operations regardless of the specific nature of the patterns, the sensory inputs, and the underlying brain structures, and is thus equally appropriate for an informative description of speech perception.

There is a growing consensus to view speech perception as an instance of a general form of pattern recognition (). To understand speech perception, the researcher only has to describe how pattern recognition works in this domain. Questions include the ecological and functional properties of audible and visible speech, as well as other

influences such as top-down constraints on what can occur when and where—that is, those sources of information that influence speech perception. Although one can discover a variety of frameworks to describe pattern recognition, their similarities far exceed the differences. Reading about one framework will certainly prepare the reader to better understand other frameworks. In this chapter, I will describe speech perception within a specific framework, one that is representative of a prototypical framework for pattern recognition. Central to this framework is the natural ease of crossmodal perception, particularly the value of visible speech when it is presented with auditory speech.

This chapter addresses both empirical and theoretical issues. At the empirical level, experiments are reviewed to illustrate 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, analysed, contrasted, and tested. Various types of model fitting strategies have been employed in variety of experimental tests. These model tests have been highly informative about how crossmodal spoken language is perceived and understood. We begin with an experimental study of the processing of unimodal and bimodal speech.

A Paradigm for Psychological Inquiry

We are attracted to bimodal speech perception as a paradigm for psychological inquiry for several reasons (Massaro & Cohen, 1983). It offers a compelling example of how processing information from one modality (vision) appears to influence our experience in another modality (audition). Second, it provides a unique situation in which multiple modalities appear to be combined or integrated in a natural manner. Third, experimental manipulation of these two sources of information is easily carried out in pattern recognition tasks. Finally, conceptualizing speech as crossmodal has the potential for valuable applications for individuals with hearing loss, person with language challenges, learners of a new language, and for other domains of language learning.

The study of speech perception by ear and eye has been and continues to be a powerful paradigm for uncovering fundamental properties of the information sources in speech and how speech is perceived and understood. 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 pattern recognition in a variety of other domains.

We believe that our empirical work would be inadequate and perhaps invalid without the corresponding theoretical framework. Thus, the work continues to address both empirical and theoretical issues. At the empirical level, experiments have been carried out to determine how visible speech is used alone and 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 have been analyzed, contrasted, and tested. In addition, a general framework for inquiry and a universal principle of behavior has been proposed, as described in the next section.

Demonstration Experiment: Varying the Ambiguity of the Speech Modalities

Most experiments of multimodal speech perception have been carried out in the context of the McGurk (1976) effect, a striking demonstration of how visual speech can influence the perceiver's perceptual experience. It has been well over two decades since the The classic McGurk effect involves the situation in which an auditory /ba/ is paired with a visible /ga/ and the perceiver reports hearing /da/. The reverse pairing, an auditory /ga/ and visual /ba/, tends to produce a perceptual judgment of /bga/. Most studies of the McGurk effect, however, use just a few experimental conditions in which the auditory and visual sources of information are made to mismatch. Investigators also sometimes fail to test the unimodal conditions separately so that there is no independent index of the perception of the single modalities. 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 auditory speech was accurately perceived. Investigators also tend to take too few observations under each of the stimulus conditions, which precludes an analysis of individual behavior and limits the analyses to group averages. A better understanding of the McGurk effect will occur when we have a better account of speech perception more generally. Our approach involves enhancing the database and testing formal models of the perceptual process.

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 systematic 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 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 6 blocks preceded by 6 practice trials and with a short break between sessions. There were 4 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. Participants 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, 1998b).

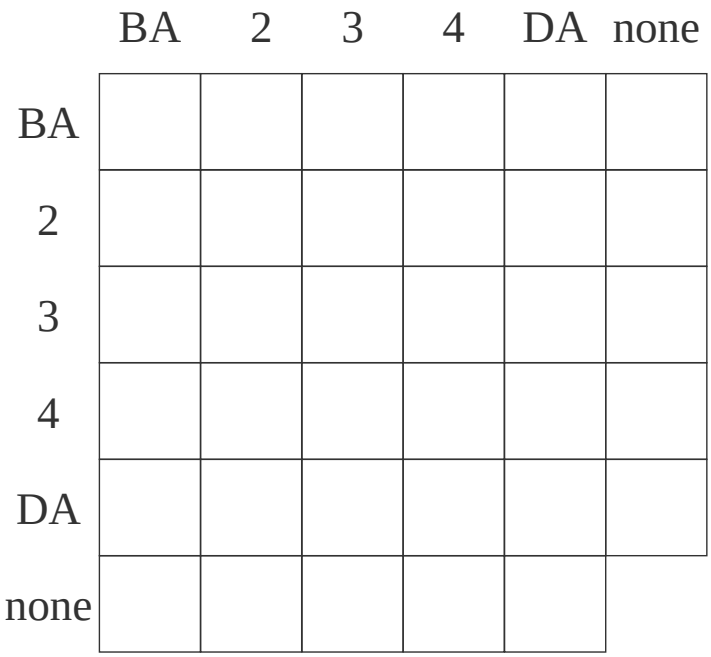
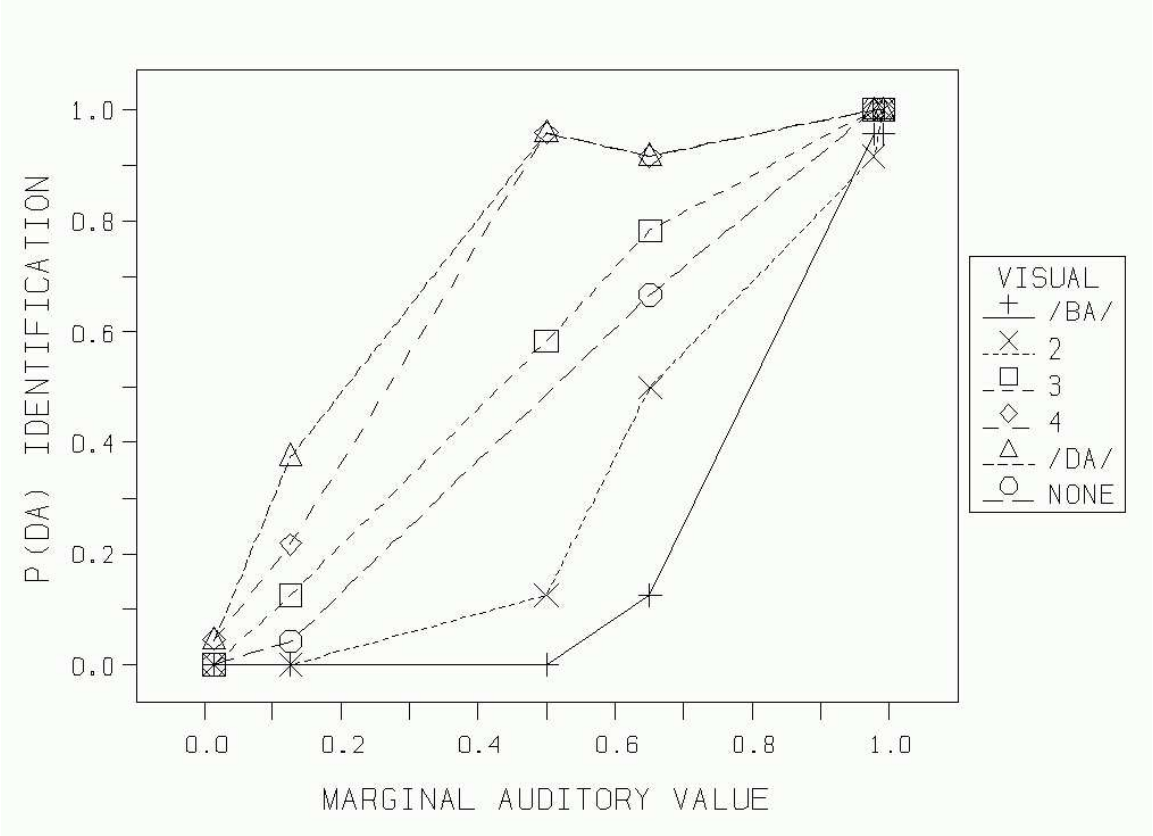


Figure 1. Expansion of a typical factorial design to include auditory and visual conditions presented alone. The five levels along the auditory and visible continua represent auditory and visible speech syllables varying in equal physical steps between /ba/ and /da/.

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 of the 82 participants for the 35 unimodal and bimodal conditions. Although it is not feasible to present the results of each of the participants, we will be able to show the outcomes for 5 different individuals. For this tutorial, we begin with the results for a single participant who can be considered typical of the others in this task.

Figure 2. The points give the observed proportion of /da/ identifications in the unimodal and factorial auditory-visual conditions as a function of the five levels of synthetic auditory and visual speech varying between /ba/ and /da/. The columns of points are placed at a value corresponding the marginal probability of a /da/ judgment for each auditory level on the independent variable. The auditory alone conditions are given by the open circles. The unimodal visual condition is plotted at .5 (completely neutral) on the auditory scale. Results for participant 9.



The points in Figure 2 give the observed proportion of /da/ responses for the auditory alone, the bimodal, and the visual alone conditions as a function of the five levels of the synthetic auditory and visual speech varying between /ba/ and /da/. Although this plot of the results might seem somewhat intimidating at first glance, I believe a graphical analysis of this nature can facilitate understanding dramatically. Notice that the columns of points are spread unevenly along the x-axis. The reason is that they are placed at a value corresponding the marginal probability of a /da/ judgment for each auditory level on the independent variable. This spacing reflects relative influence of adjacent levels of the auditory condition.

The unimodal auditory curve (indicated by the solid circles) shows that the auditory speech had a large influence on the judgments. More generally, the degree of influence of this modality when presented alone would be indicated by the steepness of the response function. The unimodal visual condition is plotted at .5 (which is considered to be completely neutral) on the auditory scale. The influence of the visual speech when presented alone is indexed by the vertical spread among the five levels of the visual condition.

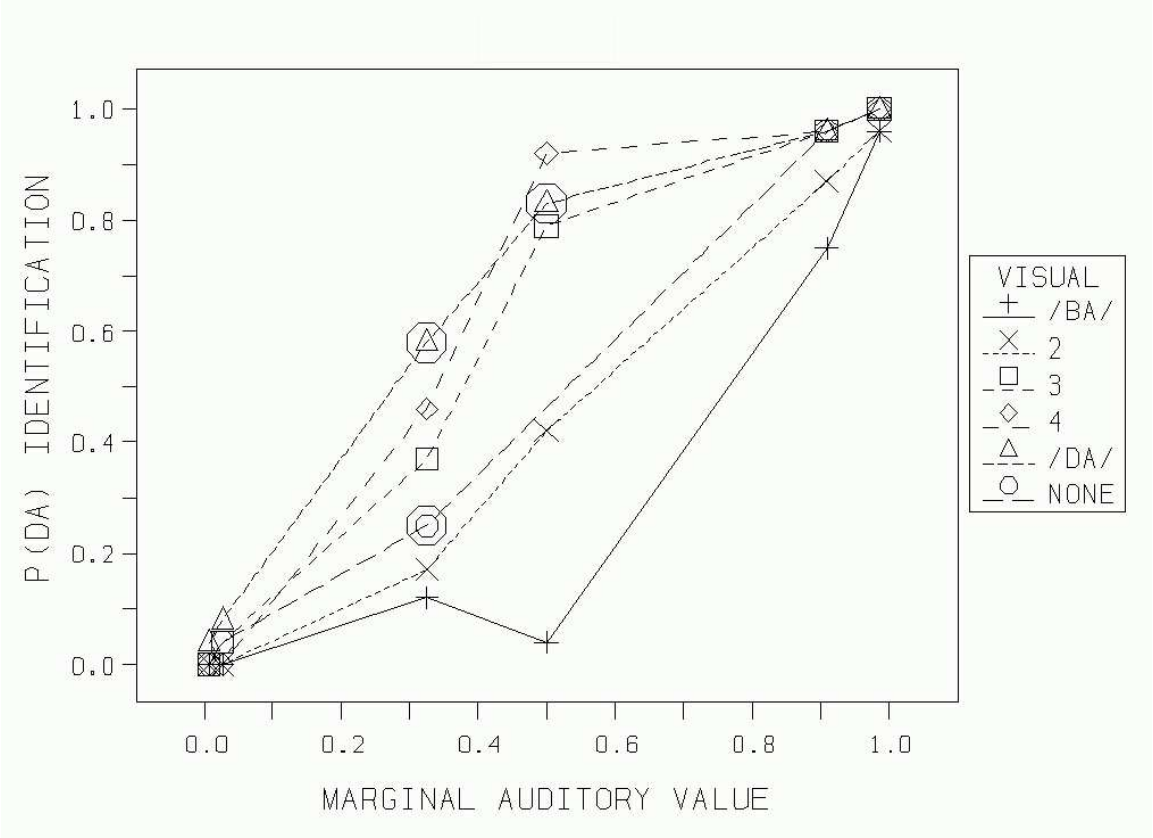
The other points give performance for the bimodal conditions. This graphical analysis shows that both the auditory and the visual sources of information had a strong impact on the identification judgments. The likelihood of a /da/ identification increased as the auditory speech changed from /ba/ to /da/, and analogously for the visible speech. The curves across changes in the auditory variable are relatively steep and also spread out from on another with changes in the visual variable. By these criteria, both sources had a large influence in the bimodal conditions.

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 of information are used in perception.

Evaluation of How Two Sources are Used.

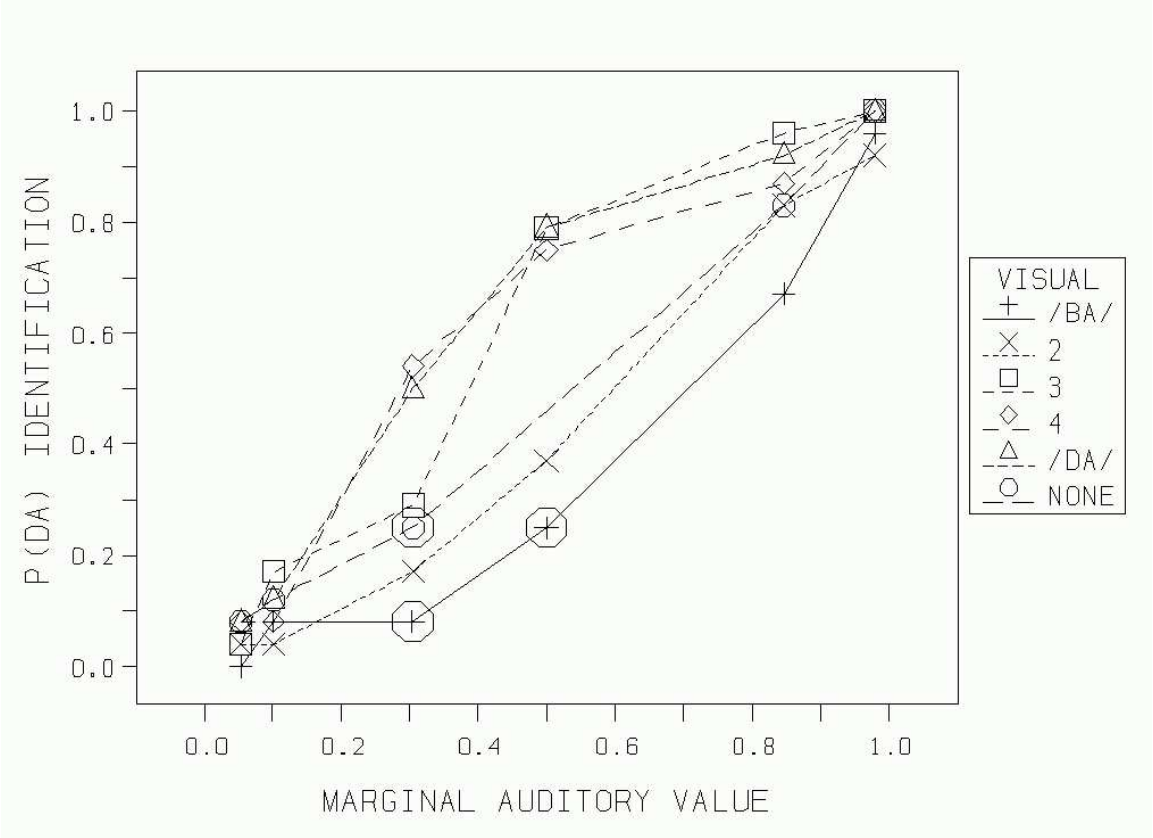
Of course, an important question is how the two sources of information are used in perceptual recognition. An analysis of several results informs this question. Figure 3 gives the results for another participant in the task. Three points are circled in the figure to highlight the conditions in which the second level of auditory information is paired with the fifth (/da/) level of visual information. When presented alone, P(/da/ | A_2) is about .25 whereas P(/da/| V_5) is about .8. When these two stimuli occur together, P(/da/| A_2 V_5) is about .6. 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 3. The points give the observed proportion of /da/ identifications in the unimodal and factorial auditory-visual conditions as a function of the five levels of synthetic auditory and visual speech varying between /ba/ and /da/. The columns of points are placed at a value corresponding the marginal probability of a /da/ judgment for each auditory level on the independent variable. The auditory alone conditions are given by the open circles. The unimodal visual condition is plotted at .5 (completely neutral) on the auditory scale. Results for participant 41. The lines are drawn through the observed points. The three large-circled points A_2V_5 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, a weighted averaging of the two sources, or a multiplicative integration of the two sources.



Other observations, however, allow us to reject these alternatives. Figure 4 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. Recall that in this forced-choice task, P(/ba/) is equal to one minus P(/da/). When presented alone, P(/ba/ | A_3) and P(/ba/| V_1) are both about .75. When these two stimuli occur together, P(/ba/| A_3 V_1) is about 9. This so-called super-additive result (the bimodal is more extreme than either unimodal response proportion) does not seem to be easily explained by either the use of a single modality or a simple averaging of the two sources. In order to evaluate theoretical alternatives, however, formal models must be proposed and tested against all of the results, not just selected conditions. We now formalize two competing models and test them against the results.

Figure 4. The points give the observed proportion of /da/ identifications in the unimodal and factorial auditory-visual conditions as a function of the five levels of synthetic auditory and visual speech varying between /ba/ and /da/. The columns of points are placed at a value corresponding the marginal probability of a /da/ judgment for each auditory level on the independent variable. The auditory alone conditions are given by the open circles. The unimodal visual condition is plotted at .5 (completely neutral) on the auditory scale. Results for participant 25. The lines are drawn through the observed points. The three large-circled points A_3V_1 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 a weighted averaging of the two sources, but can be described by a multiplicative integration of the two sources.



Tests of Competing Models

To explain pattern recognition, representations in memory are an essential component. The current stimulus input has to be compared to the pattern recognizer's memory of previous patterns. One type of memory is a set of summary descriptions of the meaningful patterns. These summary descriptions are called prototypes and they contain a description of features of the pattern. The features of the prototype correspond to the ideal values that an exemplar should have if it is a member of that category. To recognize a speech segment, the evaluation process assesses the input information relative to the prototypes in memory. Given this general theoretical framework, we consider whether or not integration of auditory and visual speech occurs. It might seem obvious that integration occurred in our experiment because there were strong effects of both auditory and visual speech. In fact, this outcome is logically possible even if integration did not occur. Most experiments using the McGurk effect paradigm were not able to demonstrate conclusively that integration occurred. It is possible, for example, that only the visual speech was used and simply dominated the judgments on some of the trials. This type of nonintegration is the simpler account of pattern recognition and we begin with a formalization of this type of model.

Nonintegration Models of Bimodal Speech Perception

According to nonintegration models, any perceptual experience results from only a single sensory influence. Thus the pattern recognition of any crossmodal event is determined by only one of the modalities, even though the influential modality might vary. Although this class of models involves a variety of alternatives that are worthy of formulation and empirical test (see Massaro, 1998b), we will formulate and test just one for illustrative purposes.

Single Channel Model (SCM)

Although there are multiple inputs, it is possible that only one of them is used. This idea is in the tradition of selective attention theories according to which only a single channel of information can be processed at any one time ((Pashler 1998). According to the single channel model (SCM), only one of the two sources of information determines the response on any given trial. Given a unimodal stimulus, it is

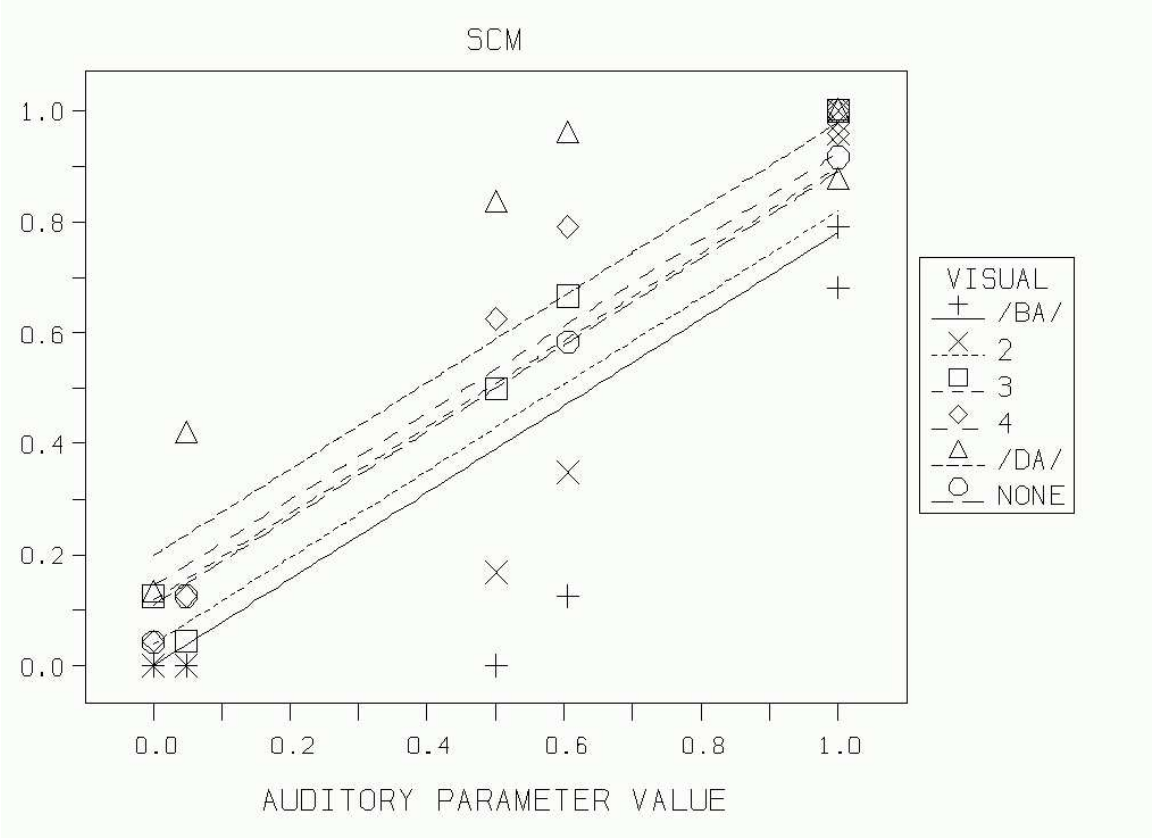
assumed that the response is determined by the presented modality. A unimodal auditory stimulus will be identified as /da/ with probability a_i, and, analogously, the unimodal visual stimulus will be identified as /da/ with probability v_j. The value i simply indexes the ith level along the auditory continuum and j indexes the level of the visual input.

Given that only one of the auditory and visual inputs can be used on any bimodal trial, it is assumed that the auditory modality is selected with some bias probability p, and the visual modality with bias 1 - p. If only one modality is used, it is reasonably to assume that it will be processed exactly as it is on unimodal trials. In this case, for a given bimodal stimulus, the auditory information will be identified as /da/ with probability a_i, and the visual information with probability v_j. Thus, the predicted probability of a /da/ response given the ith level of the auditory stimulus, a_i, and the jth level of the visual stimulus, v_j, is .

$$P(/da/ \mid A_i V_j) = pa_i + \{(1-p)v_j$$

Equation 1 predicts that a /da/ response can come about in two ways: 1) the auditory input is selected and is identified as /da/, or 2) the visual input is selected and is identified as /da/. This formalization of the SCM model assumes a fixed p across all conditions, an a_i value that varies with the auditory information and a v_j value that varies with the visual information.

Figure 5. The points give the observed proportion of /da/ identifications in the unimodal and factorial auditory-visual conditions as a function of the five levels of synthetic auditory and visual speech varying between /ba/ and /da/. The columns of points are placed at a value corresponding the marginal probability of a /da/ judgment for each auditory level on the independent variable. The auditory alone conditions are given by the open circles. The unimodal visual condition is plotted at .5 (completely neutral) on the auditory scale. Results for participant 7. The lines give the predictions of the SCM, with an RMSD of .115.



We can assess the predictive power of the SCM and other models using the 5 by 5 expanded factorial design. The points in Figure 5 gives the proportion of /da/ identifications for a prototypical participant in the task. Figure 5 also shows the predictions of the SCM, as represented by Equation 1. Equation 1 is a linear function and it predicts a set of parallel functions with this type of plot. The equation and graph illustrate how a constant increase in a_i and v_j lead to a constant increase in P(/da/). The mismatch between the observations and predictions illustrates that this model appears to be inadequate. Even so, a formal test is required. Before we present this test of the SCM, it is necessary to discuss estimation of the free parameters in a model.

Testing a Model's Predictions

We cannot expect a model's predictions of behavior to be exact or even very accurate without first taking into account what results are being predicted. As an example, we cannot know exactly how often a given person will identify one of the visible speech syllables as a particular alternative. As can be seen in a comparison among Figures 2-4, individual participants give similar but not identical results for the same experiment. We can know that one syllable might be more likely to be identified as /ba/ but we cannot predict ahead of time the actual probability of a /ba/ response by an individual participant. This uncertainty would preclude the quantitative test of models if we were not able to determine (estimate) the values of free parameters.

Free Parameters and Their Estimation

When applied to empirical data, most computational or quantitative descriptions have a set of free parameters. A free parameter in a model is a variable whose values cannot be exactly predicted in advance. We do not know what these values are, and we must use the observed results given to find them. The actual performance of the participant is used to set the value of this variable. This process is called parameter estimation.

In parameter estimation, we use our observations of behavior to estimate the values of the free parameters of the model being tested. Because we want to give every model its best shot, the goal is to find the values of the parameters that maximize how accurately the model is able to account for the results. The optimal parameter values can be found with an iterative search algorithm to find those parameter values that minimize

the differences between the predicted and observed results. The parameters and parameter space must be specified for the search. In the SCM, for example, the parameters are p, a_i, and v_j. These values are probabilities and thus must be between 0 and 1.

Equation 1 predicts P(/da/) for each of the 35 conditions in the expanded factorial experiment. The SCM does not predict in advance how often the syllable in each modality will be identified as /ba/ or /da/. According to the model, there can be a unique value of a_i for each unique level of audible speech. Similarly, there can be a unique value of v_j for each level of visual speech. We also do not know the value of p on bimodal trials, which requires another free parameter. For unimodal trials, we assume that the presented modality is always used. We have 35 equations with 11 free parameters: the p value, the 5 a_i and 5 v_j values. Finding values for these 11 unknowns allows us to predict the 35 observations.

RMSD Measure of Goodness-of-Fit

A factor that is often used to maximize the goodness-of-fit is 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. (Squaring the differences makes all differences positive and also magnifies large deviations compared to small ones.) The RMSD can be thought of as a standard deviation of the differences between the 35 predicted and observed values. The RMSD would increase as the differences increase. In general, the smaller the RMSD value, the better the fit of the model.

The quantitative predictions of the model are determined by using any mimimization routine such as 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 maximizes the accuracy of the predictions by minimizing the RMSD. The outcome is a set of parameter values which, when put into the model, come closest to predicting the observed results.

Data Base and Model Tests

The results for the present model tests come from the results from 82 participants, with 24 observations from each participant under each of the 35 conditions (Massaro, 1998b). The model fit was carried out separately on each participant's results. We have learned that individuals differ from one another and averaging the results across individuals can be hazardous. The free parameters of a model should be capable of handling the individual differences. Fitting a model to single individuals should permit the model to describe individual participants while also accounting for between-participant differences, insofar as they can be captured by the differences among the 11 parameters.

The observations and predictions of the SCM for a representative participant are given in Figure 5. The data points in the figures are the observations, and the lines correspond to the model predictions. We use lines for the predictions so one can see the form of a model's predictions. The distance between the observed points and these predictions gives a graphical measure of goodness-of-fit. The predictions of the SCM do not capture the trends in the data. The predictions are a set of parallel lines whereas the observations resemble an American football--wide in the middle and narrowing at the ends.

The RMSD is also used to evaluate the goodness-of-fit of a model both in absolute terms and in comparison to other models. Of course, the smaller the RMSD the better the fit of a model. The RMSD for the fit of the SCM for the participant shown in Figure 5 was .137. The RMSDs for the fit of the SCM across all 82 participants averaged .097. We now formalize an integration model, called the fuzzy logical model of perception (FLMP).

The Fuzzy Logical Model of Perception (FLMP)

The FLMP is shown in Figure 6. Consider the case in which the perceiver is watching the face and listening to the speaker. Although both the visible and audible speech signals are processed, each source is evaluated independently of the other source. The evaluation process consists of determining how much that source supports various alternatives. The integration process combines these sources and outputs how much their combination supports the various alternatives. The perceptual outcome for the perceiver will be a function of the relative degree of support among the competing alternatives.

More generally, multiple sources of information contribute to the identification and interpretation of the language input. The assumptions central to the model are 1) each source of information is evaluated to give the continuous degree to which that source specifies various alternatives, 2) the sources of information are evaluated independently of one another, 3) the sources are integrated multiplicatively 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 quantitative predictions of the FLMP have been derived and formalized in a number of different publications (e.g., Massaro, 1987b, 1998b). 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. The predicted probability of a /da/ response given an auditory input, P(/da/|A_i) is equal to

$$P(/da/ \mid A_i) = \frac{a_i}{a_i + (1 - a_i)} = a_i \qquad (2)$$

Similarly, the predicted probability of a /da/ response given an visual input, P(/da/|V_j) is equal to

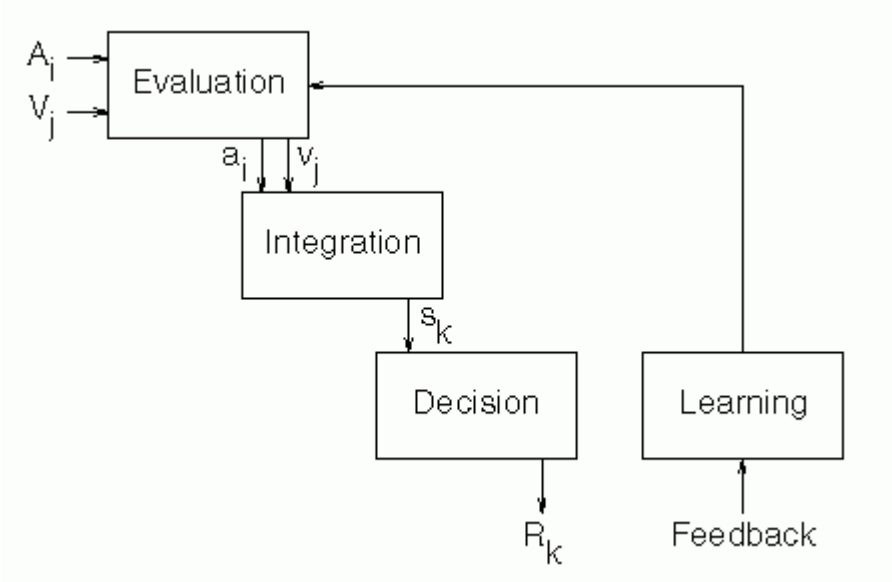


Figure 6. Schematic representation of the three processes involved in perceptual recognition. The three processes are shown to precede left to right in time to illustrate their necessarily successive but overlapping processing. These processes make use of prototypes stored in long-term memory. The sources 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. The feedback is assumed to tune the prototypical values of the features used by the evaluation process.

$$P(/da/|V_j) = \frac{v_j}{v_j + (1 - v_j)} = v_j \qquad (3)$$

For bimodal trials, the predicted probability of a /da/ response given auditory and visual inputs, $P(/da/|A_iV_j)$ is equal to

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

Equations 2-4 assume independence between the auditory and visual sources of information. Independence of sources at the evaluation stage is motivated by the principle of category-conditional independence (Massaro, 1998b; Massaro & Stork, 1998). Given that it isn't possible to predict the evaluation of one source on the basis of the evaluation of another, the independent evaluation of both sources is necessary to make an optimal category judgment. Although the sources are kept separate at evaluation, they are integrated to achieve perception, recognition, and interpretation. The FLMP assumes multiplicative integration, which 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, which is effected through a decision stage, which divides the support for one category by the summed support for all categories.

Underlying Neural Mechanism

A natural question is what is the neural mechanism postulated by the integration algorithm specified in the FLMP. An important set of observations from single cell recordings in the cat could be interpreted in terms integration of the form specified by the FLMP (Meredith, this volume; Stein & Meredith, 1993; Stein et al., this volume). A single hissing sound or a light spot can activate neurons in the superior colliculus. A much more vigorous response is produced, however, when both signals are simultaneous presented from the same location. This results parallels the outcomes we have observed in unimodal and bimodal speech perception.

As proven elsewhere, the FLMP is mathematically equivalent to Bayes theorem (Massaro, 1998b, Chapter 4). Anatasio and Scheier (this volume) propose that the brain

can implement a computation analogous to Bayes' rule, and that the response of a neuron in the superior colliculus is proportional to the posterior probability that a target is present in its receptive fields, given its sensory input. The authors also assume that the visual and auditory inputs are conditionally independent given the target. This implies that the visibility of the target indicates nothing about the audibility of the target, and vice-versa. This assumption corresponds to our assumption of category-conditional independence. They show that the target-present posterior probability computed from the impulses from the auditory and visual neurons is higher given sensory inputs of two modalities than it is given input of only one modality. In addition, when only one modality is activated, the target-present posterior probability computed from the impulses from the auditory and visual neurons is less than the modality specific posterior probability from the activated modality. Given the value of neurons that evaluate input from several modalities.

Anatasio and Scheier ask why all neurons don't have this property. The answer is that inputs from two modalities can actually produce more uncertainty than an input from just one of the modalities. This situation occurs when one of the inputs has very little resolution, which can degrade their joint occurrence. We have observed similar results, particularly in the perception of emotion in which adding information from the voice with the face can actually decrease accurate identification relative to the face alone.

Bernstein, Auer, and Moore (this volume) distinguish whether speech perception is best described by convergence or by an association of modality-specific speech representations. These two alternatives bear some similarity to two of the three alternatives that I have proposed as possible mechanisms of the joint influence of audible and visible speech (Massaro, 1998b; 1999). These alternatives are shown in Figure 7. Bernstein et al. claim that the FLMP might be interpreted as claiming convergent integration. In my discussion of these alternatives (Massaro, 1999), I indicated that "convergent integration offers a potential implementation of the FLMP" but did not mean to imply that I favored this type of

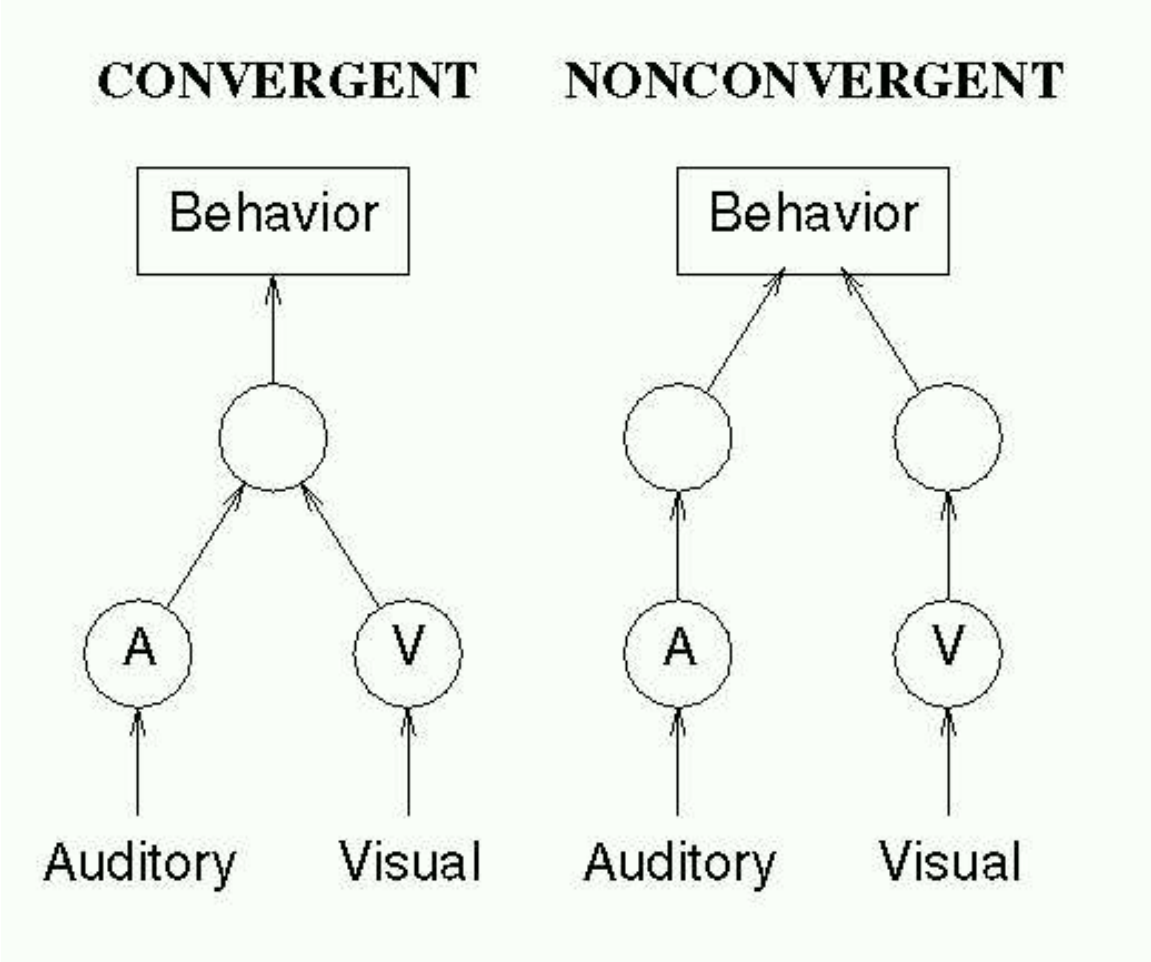


Figure 7. Neural representations of convergent integration and non-convergent integration.

integration over non-convergent integration. In fact, I observed that non-convergent integration "was most consistent with the findings …", findings that we reviewed in Section *The Relationship between Identification and Discrimination* in the present chapter.

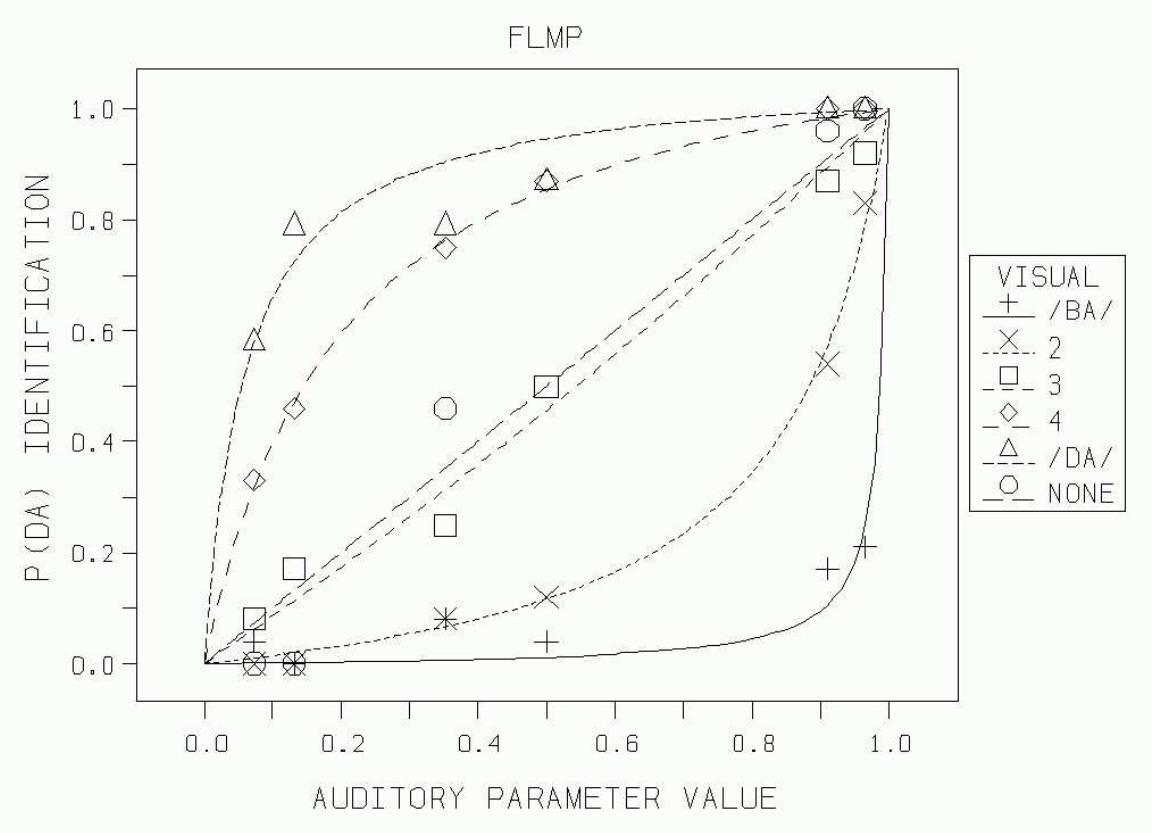
When considering the important relationship between psychological models like the FLMP and underlying neural mechanisms, one has to keep in mind that information-processing algorithms are not going to be easily observable in the underlying hardware. As I have stated in another discussion of this issue, "Only biology is found in living systems, not algorithms … a biological explanation cannot represent and therefore replace the algorithm. Biology is only biology. Similarly, we do not expect to find our law of pattern recognition in the brain. We expect to observe only chemical and electrical

activity, not the algorithm. Of course, this activity can be interpreted in different ways (Massaro, 1998b, p. 105).

Before addressing the issue of neural mechanism, a couple of attributes of the FLMP should be emphasized. The FLMP takes a strong stance on the question of discrete versus continuous information processing. Information input to a stage or output from a stage is continuous rather than discrete. Furthermore, the transmission of information from one stage to the next is assumed to occur continuously rather than discretely. The three processes shown in Figure 6 are offset to emphasize their temporal overlap. Evaluated information is passed continuously to integration while additional evaluation is taking place. Although it is logically the case that some evaluation must occur before integration can proceed, the processes are assumed to overlap in time. Similarly, integrated information is continuously made available to the decision process.

It is important to emphasize that information transformed from one stage to another does not obliterate the information from the earlier stage. Thus, evaluation maintains its information even while simultaneously passing it forward to the integration process. There is evidence that information can be maintained in memory at multiple levels and in various forms. As observed by Mesulam (1998) in a review of the neural underpinnings of sensation to cognition, in "The transfer of information …, several (synaptic) levels remain active as the pertinent information is conveyed from one node to the other." (Mesulam, 1998, p. 1041). This parallel storage of information does not negate the sequential stage model in Figure 6. What is important to remember is that transfer of information from one stage to another does not require that the information is lost from the earlier stage. Integrating auditory and visual speech does not necessarily compromise or modify the information at the evaluation stage. Thus, given that multiple representations can exist in parallel, there may be both convergence and association operative in the perception of auditory –visual speech. There appears to be strong neural evidence for two types processes: "the establishment, by local neuronal groups, of convergent cross-modal associations related to a target event; and (ii) the formation of a directory pointing to the distributed sources of information." (Mesulam, 1998, p. 1024). These can be interpreted to correspond to convergent and non-convergent integration (association), respectively. We believe that the FLMP algorithm can be implemented by

Figure 8. The points give the observed proportion of /da/ identifications in the unimodal and factorial auditory-visual conditions as a function of the five levels of synthetic auditory and visual speech varying between /ba/ and /da/. The columns of points are placed at a value corresponding the marginal probability of a /da/ judgment for each auditory level on the independent variable. The auditory alone conditions are given by the open circles. The unimodal visual condition is plotted at .5 (completely neutral) on the auditory scale. Results for participant 30. The lines give the predictions of the FLMP, with an RMSD of .051.



both of these neural mechanisms. It might be the case that auditory-visual speech processing follows just non-convergent integration, but there are other domains such as localizing an event given sound and sight that follow convergent integration (Anatasio, this volume; Meredith, this volume).

We don't really know how well the SCM performs without contrasting it with other models. We favor the FLMP, which is an integration model. The FLMP was fit to these same results, using Equations 2-4 with 10 free parameters. Like the SCM, the FLMP also requires 5 a_i and 5 v_j values. In the FLMP, however, these are not probabilities but fuzzy truth values between 0 and 1 indicating the degree to which the information supports the alternative /da/ (see Equations 2-4). The RMSD for the fit of the FLMP for the participant shown in Figure 8 was .051, and the RMSDs for the fit of the FLMP for the 82 individual participants averaged .051.

As in all areas of scientific inquiry, 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, 1998b, Chapter 6). In one case, participants were given just two alternatives, and in the other the same participants 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, 1998b, pp. 265-268).

We have explored alternative methods of model testing. The first involves the match between the goodness-of-fit of a model and a benchmark measure that indexes what the goodness of fit should be if indeed the model was correct. Because of sampling variability, we cannot expect a model to give a perfect description of the results. Second we have used a model selection procedure suggested by Myung and Pitt (1987; Massaro et al., 2001). The advantage of the FLMP over the SCM and other competing models holds up under these alternative procedures of model testing (Massaro, 1998b, Chapter 10; Massaro et al., 2001). 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 and 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.

Broadening the Domain of Inquiry

We have broadened our domain of inquiry in several directions. The first direction involves the development of a framework for understanding individual differences. One of the first impressions a researcher obtains is how differently individuals will perform in the same experimental task. This variability is not surprising once we consider that each of us has unique life histories and genetics. Given the FLMP framework, however, 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's and v_j's) indicating the degree of support for each alternative 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.

Once individual variability is accounted for, by estimating free parameters in the fit of the model, we are able to provide a convincing description of how the information is processed and mapped into a response. Although we cannot predict a priori how /ba/-like a particular audible or visible speech syllable is for a given individual, we can predict how these two sources of information are integrated and a decision is made. In addition, the model does take a stand on the evaluation process in the sense that it is assumed that the auditory and visual sources of information are evaluated independently of one another.

Our research has made important progress by analysing the results of individual participants rather than being dependent on average data. As is well known, it is possible that average results of an experiment do not reflect the results of any individual making up that average. Our research has adapted the sophisticated methodology developed in psychophysics and the theory of signal detectability to provide a framework for the study of individual participants (Green & Swets, 1966).

Given this framework, we have explored a broad variety of dimensions of individual variability in terms of .the distinction between information and information processing. 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. The methodology of a set of cross-linguistic experiments allowed us to separate information differences from information processing differences. Earlier cross-linguistic results had led investigators to conclude that the *processing* of bimodal speech differed for Japanese and English speakers. Although the results of experiments with native English, Spanish, Japanese, and Dutch talkers showed substantial differences in performance across the different languages (Massaro et al., 1993, 1995), the application of the FLMP indicated that these differences could be completely accounted for by information differences with no differences in information processing. The information in a speech segment made available by the evaluation process naturally differs for talkers of different languages whereas the information processing appears to be invariant. The differences that are observed are primarily the different speech categories used by the different linguistic groups, which can be attributed to differences in the phonemic repertoires, phonetic realizations of the syllables, and phonotactic constraints in these different languages. In addition, talkers of different languages are similarly influenced by visible speech, with its contribution largest to the extent the other source is ambiguous. The details of these judgments are nicely captured in the predictions of the FLMP.

A second direction of our research concerns ecological variability, which refers to different perceptual and cognitive situations involving pattern recognition and to variations in the task itself. Generally, we have asked to what extent the processes uncovered in bimodal speech perception generalize across 1) sensory modalities, 2) environmental domains, 3) test items, 4) behavioral measures, 5) instructions, 6) and

tasks. We found, for example, that written information can influence speech perception in an analogous manner to visible speech (Massaro, 2002). Participants tend to perceive /di/ when an ambiguous auditory stop consonant is paired with the written letter D, similar to the influence of a visible spoken /di/.

Pursuing the question of whether our model of pattern recognition is valid across different domains, we examined how emotion is perceived given facial and vocal cues of a speaker (Massaro, 1998b; 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. Similar results have been found in other laboratories (see de Gelder, Vroomen, & Pourtois, this volume). The perception of emotion appears to be well-described by our theoretical framework. Analogous to speech perception, we find a synergistic relationship between the face and the voice. Messages communicated by both of the modalities is more informative than either one alone (Massaro, 1998b).

A Universal Principle

In the course of our research, we have developed a universal principle of perceptual cognitive performance to explain pattern recognition (Campbell et al., 2001; Massaro, 1998b; Massaro et al., 2001). As illustrated by this handbook of multisensory integration, animals are influenced by multiple sources of information in a diverse set of situations. In multisensory texture perception, for example, there appears to be no fixed sensory dominance by vision or haptics, and the bimodal presentation yields higher

accuracy than either of the unimodal conditions (Lederman & Klatzky, this volume). In many cases, these sources of information are ambiguous and any particular source alone does not usually specify completely the appropriate interpretation. According to the FLMP, the perceiver evaluates these multiple sources of information in parallel, and determines the degree to which each source supports various interpretations. The sources are then integrated to derive the overall support for each alternative interpretation. Finally, the relative support for each alternative determines the perceptual judgment. Parenthetically, it should be emphasized that these processes are not necessarily conscious or under deliberate control.

Advantages of Bimodal Speech Perception

There are several reasons why the use of auditory and visual information together is so successful, and why they hold so much promise for educational applications such as language tutoring. These include a) robustness of visual speech, b) complementarity of auditory and visual speech, and c) optimal integration of these two sources of information.

Empirical findings show that speech reading, or the ability to obtain speech information from the face, is robust. Research has shown 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 (Jordan & Sergeant, 2000; Massaro, 1998b; Munhall & Vatikiotis-Bateson, this volume). These findings indicate that speechreading is highly functional in a variety of nonoptimal situations.

Another example of the robustness of the influence of visible speech is that people naturally integrate visible speech with audible speech even when the temporal occurrence of the two sources is displaced by about a 1/5 of a second. Given that light and sound travel at different speeds and that the dynamics of their corresponding sensory systems also differ, a crossmodal integration must be relatively immune to small temporal asynchronies. To assess the robustness of the integration process across relatively small temporal asynchronies, the relative onset time of the audible and visible

sources was systematically varied (Massaro & Cohen, 1993). In the first experiment, bimodal syllables composed of the auditory and visible syllables /ba/ and /da/ were presented at five different onset asynchronies. The second experiment replicated the same procedure but with the vowels /i/ and /u/. The results indicated that perceivers integrated the two sources at asynchronies of 200 ms or less.

More recently, two experiments were carried out to study whether integration would be disrupted by differences in the temporal arrival of the two sources of information (Massaro et al., 1996). Synthetic visible speech and natural and synthetic auditory speech were used to create the syllables /ba/, /va/, /tha/, and /da/. An expanded factorial design was used to present all possible combinations of the auditory and visual syllables, as well as the unimodal syllables. The tests of formal models made it possible to determine when integration of audible and visible speech did occur. The FLMP, an additive model, and an auditory dominance model were tested. The FLMP gave the best description of the results, when the temporal arrival of the two sources of information was within 250 ms. Results indicated that integration was not severely disrupted with asynchronies of 250 ms or less. These results are in agreement with similar experiments reviewed by Munhall and Vatikiotis-Bateson (this volume). The findings support the conclusion that integration of auditory and visual speech is a robust process and is not easily precluded by offsetting the temporal occurrence of the two sources of information.

Complementarity of auditory and visual information simply means that one of the sources is most informative in those cases in which the other is weakest. Because of this, a speech distinction is differentially supported by the two sources of information. That is, two segments that are robustly conveyed in one modality are relatively ambiguous in the other modality. For example, the difference between /ba/ and /da/ is easy to see but relatively difficult to hear. On the other hand, the difference between /ba/ and /pa/ is relatively easy to hear but very difficult to discriminate visually. The fact that two sources of information are complementary makes their combined use much more informative than would be the case if the two sources were non-complementary, or redundant (Massaro, 1998b, Chapter 14).

The final characteristic is that perceivers combine or integrate the auditory and visual sources of information in an optimally efficient manner (Massaro, 1987b; Massaro

& Stork, 1998). 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.

One might question why perceivers integrate several sources of information when just one of them might be sufficient. Most of us do reasonably well in communicating over the telephone, for example. Part of the answer might be grounded in our ontogeny. Integration might be so natural for adults even when information from just one sense would be sufficient because, during development, there was much less information from each sense and therefore integration was all the more critical for accurate performance (see Lewkowicz, this volume).

Additional Tests of the Fuzzy Logical Model of Perception (FLMP)

Perceivers who are hard of hearing obviously have less auditory information, but we can also ask whether they differ in terms of information processing. We can ask whether the integration process works the same way regardless of the degree of hearing loss. 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 possible that integration of the two modalities is more difficult with cochlear implants than with hearing aids. It should be noted that addressing this question does not depend on controlling for individual characteristics such as the level of hearing loss, when it occurred, how long it has persisted, when the hearing aid or implant was received, and so on. Given our distinction between information and information processing within the FLMP framework, we can address the nature of information processing across the inevitable differences that will necessarily exist among the individuals in the study.

Study of Hard of Hearing Children

Erber (1972) tested three populations of children (adolescents and young teenagers): normal hearing (NH). severely impaired (SI), and profoundly deaf (PD). 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 four years of experience with the oral method of crossmodal speech perception. The hearing-impaired children used their hearing-assisted devices during the test. None of the children with normal hearing had any training in speechreading. The 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. It is important to note that the talkers face was intensely illuminated so that the inside of the oral cavity was visible. The test was presented under auditory, visual, and bimodal conditions.

The results for the SI group under the three presentation conditions are shown in Figure 9 in the form of confusions matrices. These data are not as overwhelming as they might seem at first glance. The confusion matrix provides for each of the 8 stimuli the proportions of each of the 8 possible responses. Although the SI group made many errors on the auditory speech, they revealed a tremendous performance in the bimodal condition relative to either of the unimodal conditions.

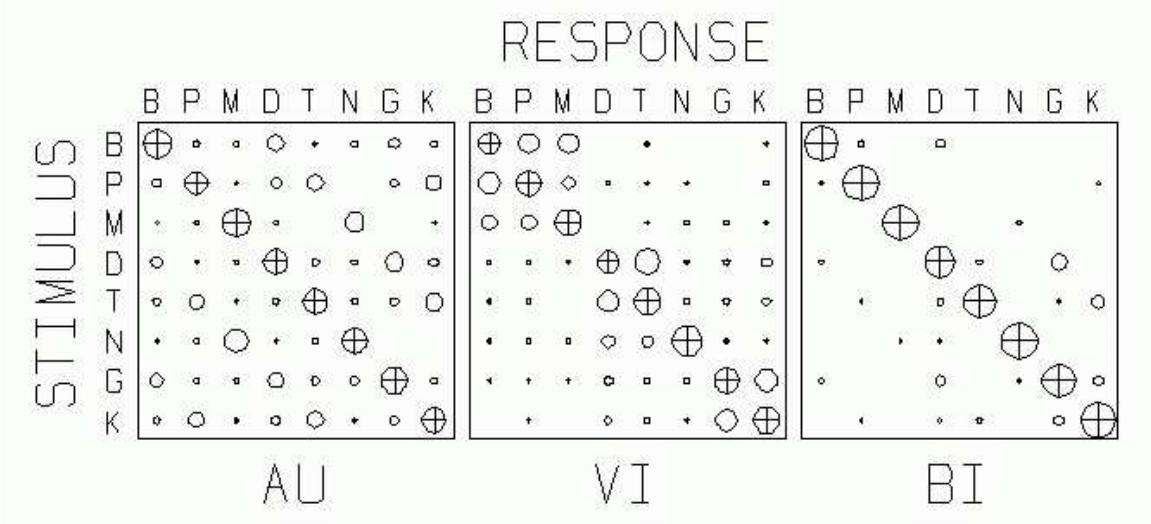


Figure 9. Confusion matrix for children with severely impaired (SI) hearing. The area of the circle is proportional to response probability. The results should be interpreted as both the observations and the predictions of the FLMP because they were essentially equivalent to one another; the small differences are not noticeable in this type of plot.

The FLMP was applied to the results of all three groups and gave an excellent description of the confusion errors of all three groups of children. The predicted values are not plotted in Figure 9 because they would not be noticeably different from the observed. Or equivalently, one can say that the predictions are also plotted but they are perceptually identical to the observations. Erber's results also reveal a strong complementarity between the audible and visible modalities in speech, which is discussed more fully in Massaro (1998b, Chapter 14).

Study of Hard of Hearing Adults

Many individuals with hearing aids (HA) or cochlear implants (CI) are able to understand auditory speech. In a substantial number of cases, however neither device provides a sufficiently rich information source. 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 robust communication environment for these individuals with one of these two assistive devices. Solid evidence for this conclusion comes from a study by Agelfors (1996). She studied persons with HA and CI in several speech tests under auditory, visual, and bimodal presentations. One test involved the identification of 16 Swedish consonants presented in a /aCa/ context preceded by a carrier phase. The 16 syllables were /p,b,m,t,d,n,g,ng,f,v,s,sh,r,l,j/. 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.

Massaro & Cohen (1999) evaluated these results in the context of the FLMP and other competing models. According to the FLMP, there should be a superadditive effect of the bimodal presentation relative to the unimodal conditions. The superadditivity results from both complementarity and an optimal integration algorithm (Massaro, 1998b, Chapter 14). The FLMP analysis of Agelfors' study addresses an interesting question. Does bimodal presentation give the same synergy advantage for HA and CI? Perhaps integration does not occur as optimally with CI relative to HA, for example. To address this question, 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 three with poorer hearing (HA-). For the CI

group, there were eight participants with better auditory recognition (CI+) and seven with poorer auditory recognition (CI-).

Given the experimental design, a confusion matrix gives the results to be predicted. The confusion matrix provides for each of the 16 stimuli the proportions of each of the 16 possible responses. A modality-analysis FLMP can be tested against this confusion matrix by estimating the amount of support that a modality-specific syllable presentation provides for each of the 16 consonants. Thus, 16 times 16 = 256 parameters are necessary to describe the auditory information and the same number to describe the visual, for a total of 512. Given the three confusion matrices in each condition, there is a total of 3 times 256 = 768 independent data points. Thus, the ratio of data points to free parameters is thus 3 to 2.

The results showed that all individuals performed more accurately in the bimodal condition relative to the unimodal conditions; i.e., superadditivity was obtained. Furthermore, the FLMP gave a good description of performance of each of the four subgroups. The SCM with an additional weight parameter was also tested and performed much more poorly than the FLMP in that its RMSD was about six to eight times larger than RMSD for the FLMP. To reduce the number of free parameters in the model tests, we also tested the models by describing the auditory and visual speech in terms of features.

Feature Analysis Implementation

The model test we have presented in the previous section makes no assumptions about the psychophysical properties of the 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/. A description of the features of the speech segments can save a large number of free parameters, because it is assumed that a given feature in a given modality has the same impact regardless of what segment it is in. Following the tradition begun with Miller and Nicely (1955), we can define each segment by five features: voicing, nasality, place, frication, and duration. The feature values for one modality are assumed to 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.

In this implementation, each of the test syllables is described by the conjunction of five features for unimodal speech and the conjunction of ten features for bimodal speech. Even though each feature is defined as a discrete category or its complement (e.g., voiced or voiceless), its influence in the perception of visible speech is represented by a continuous 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 interpretation would be that the nasal feature is not functional at all. Because of the definition of negation as one minus the feature value, a feature value of .5 would give the same degree of support for a segment that has the feature as it would for a viseme that doesn't have the feature. If the calculated parameter value is .20, however, then the nasal feature is functional but 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 non-nasal response alternatives to degree .9, then a non-nasal stimulus also supports a non-nasal alternative to degree .9.

The overall match of the feature set to the prototype was calculated by combining the features according to the FLMP. These assumptions dictate that 1) the features are the sources of information that are evaluated independently of one another, and 2) the features are integrated multiplicatively (conjoined) to give the overall degree of support for a viseme alternative, and 3) the stimulus is categorized according to the relative goodness decision rule. Thus, this implementation parallels modality-analysis FLMP in all aspects except for the featural description of the stimulus and response alternatives. The SCM was also implemented with this same featural description. The FLMP and SCM were tested against the confusion matrices by estimating the amount of information in each feature and the featural correspondence between the stimulus and response prototypes. Thus, 5 parameters are necessary to describe the auditory information and the same number to describe the visual. The SCM requires an additional weight parameter.

The fit of the FLMP to the four different groups gave an average RMSD of about half of that given for the fit of the SCM.

The Relationship between Identification and Discrimination

One of the themes of research from the perspective that speech is special concerns how speech perception differs from prototypical perceptual phenomena (Fowler, this volume). As an example, two stimuli that differ in two ways are easier to discriminate than if they differ in just one of the two ways. Advocates of the speech is special persuasion have claimed to provide evidence that this is not always the case in speech (see Fowler, this volume). Consider two speech categories /ba/ and /da/ cued by auditory and visual speech. A visual /ba/ paired with an auditory /da/ might give a similar degree of overall support for the category /da/ as an auditory /ba/ paired with a visual /da/. The speech is special claim is that these two items should be difficult to discriminate from one another. However, the research that has been cited as support for this outcome has a number of theoretical and methodological limitations (Massaro, 1987a, 1989, 1998b), limitations similar to those existing in claims for categorical perception. Basically, these studies are simply another variant of categorical perception in disguise, and vulnerable to a host of criticisms (Massaro, 1998a).

To illustrate that speech is not special, it is worthwhile to review a test (Massaro & Ferguson, 1993). Participants performed both a perceptual identification task and a same-different discrimination task. There were 3 levels (/ba/, neutral, /da/) of visual speech and 2 levels (/ba/, /da/) of auditory speech. This design gives 2 times 3 = 6 unique bimodal syllables for the identification task. In the identification task, participants identified these syllables as /ba/ or /da/. For the same-different task discrimination task, two of the bimodal syllables were presented successively, and the task was indicate whether the two syllables differed on either the auditory or visual channels. There were 20 types of discrimination trials: 6 "same" trials, 6 trials with auditory different, 4 trials with visual different, and 4 trials with both auditory and visual different.

The predictions of the FLMP were derived for both tasks, and the observed results of both tasks were described with the same set of parameter values. The predictions for the identification task were derived in the standard manner. At the evaluation stage, truth values (of fuzzy logic) are assigned to the auditory and visual sources of information

indicating the degree of support for each the response alternatives /ba/ and /da/. The truth values lie between zero and one, with zero being no support, one being full support, and .5 being completely ambiguous. Integration computes the overall support for each alternative. The decision operation in the identification task determines the support for the /da/ alternative relative to the sum of support for each of the /ba/ and /da/ alternatives, and translates relative support into a probability.

Given the FLMP's prediction for the identification task, its prediction for a same-different task can also be derived. Participants are instructed to respond "different" if a difference is perceived along either or both modalities. Within the framework of fuzzy logic, this discrimination task is a disjunction task. The perceived difference along the visual dimension is given by the difference in their truth values assigned at the evaluation stage, and analogously for the auditory dimension. The perceived difference given two bimodal speech syllables can be derived from the assumption of a multiplicative conjunction rule for integration in combination with DeMorgan's Law. It is also assumed that the participant computes the degree of sameness from the degree of difference, using the fuzzy logic definition of negation. The participant is required to select a "same" or "different" response in the discrimination task, and the actual "same" or "different" response is derived from the relative goodness rule used at the decision operation.
The predictions of the FLMP were determined for both the identification and discrimination tasks. There were 6 unique syllables in identification, and there were 14 types of different trials and 6 types of same trials. These 26 independent observations were predicted with just 5 free parameters, corresponding to the 3 levels of the visual

Figure 10. Observed and predicted probability of a /da/ identification in the identification task and the observed and predicted probability of a different judgment in the discrimination task, as a function of the different test events. The points are given by letters: The letters A through T give the discrimination performance, and the letters U through Z give identification. The conditions are listed on the right of the graph. For example, A corresponds to a visual /ba/ auditory /ba/ followed by a visual /ba/ auditory /ba/. Predictions are of the FLMP, which assumes maintenance of separate auditory and visual feature information at the evaluation stage.

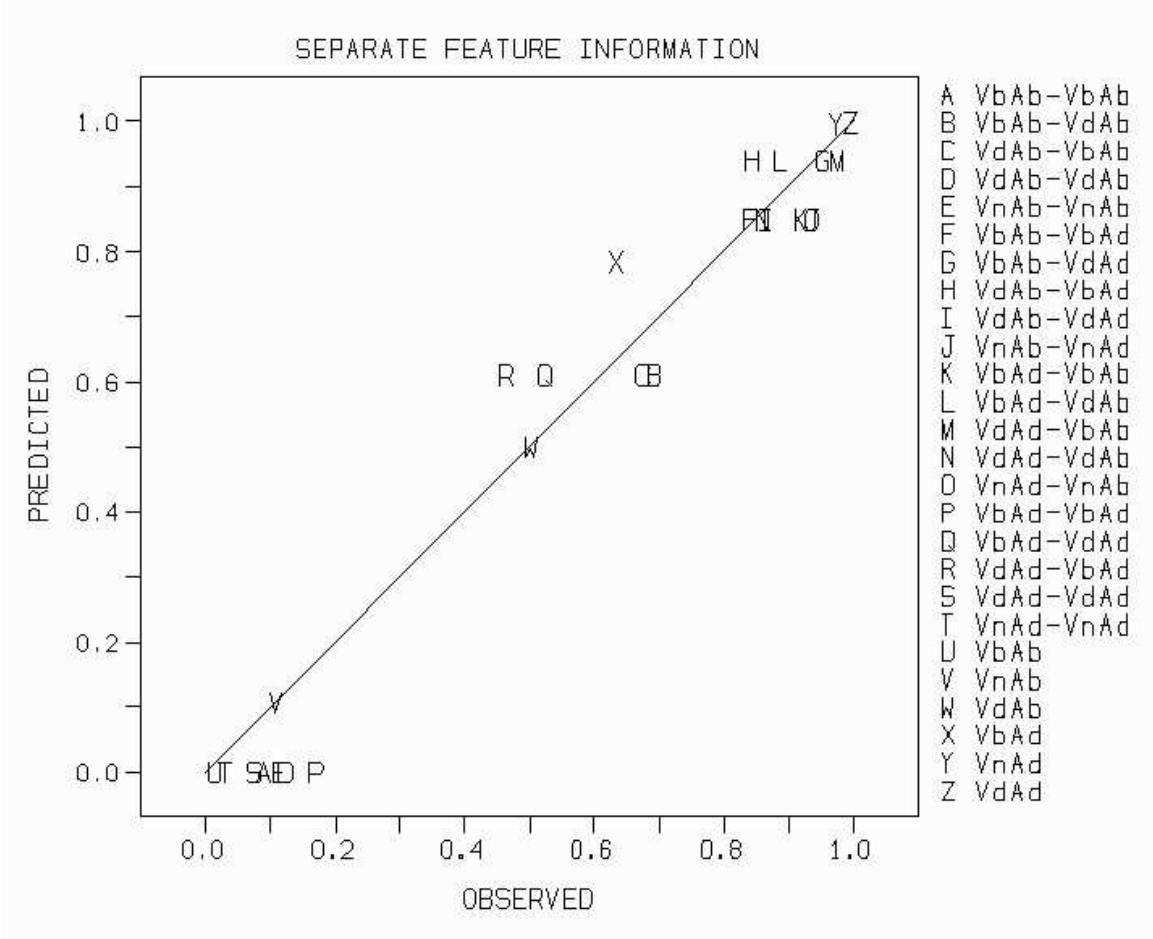
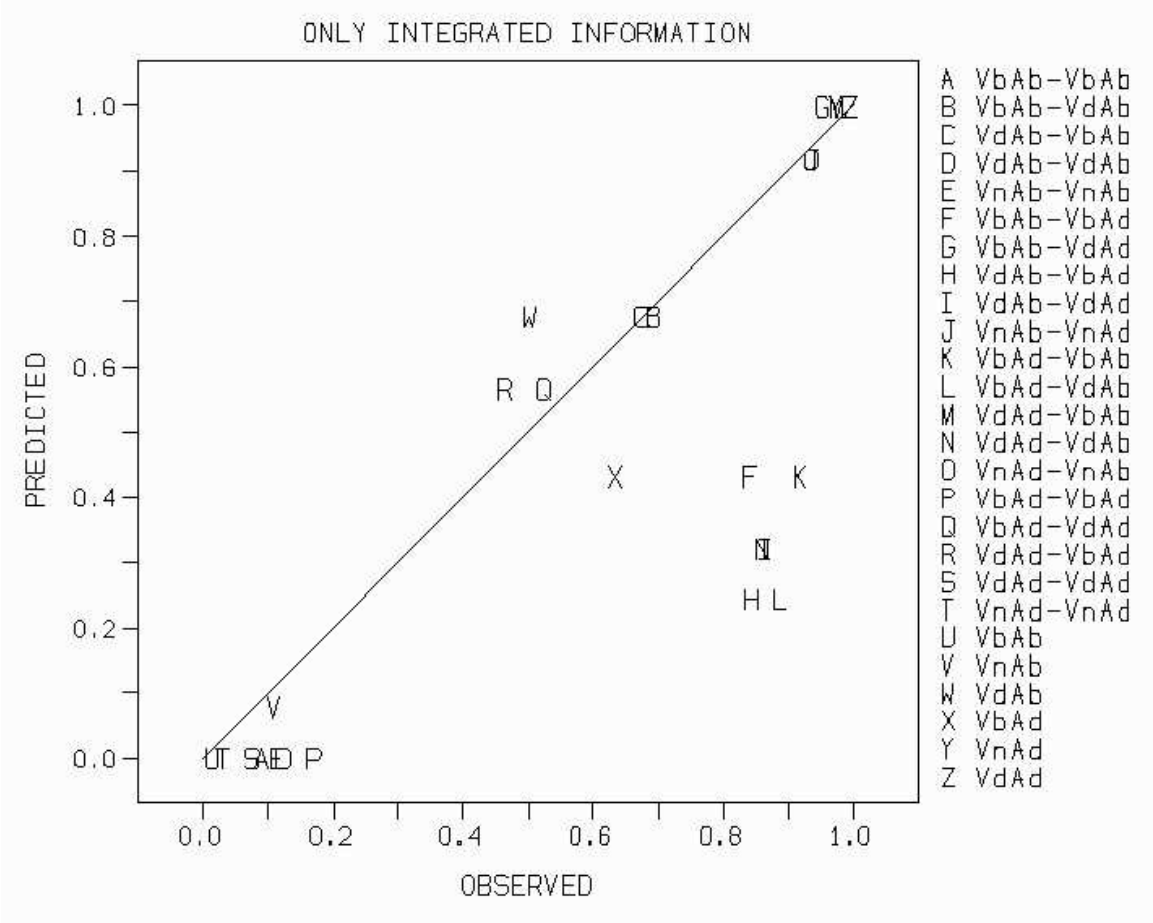


Figure 11. Observed and predicted probability of a /da/ identification in the identification task and the observed and predicted probability of a different judgment in the discrimination task, as a function of the different test events. The points are given by letters: The letters A through T give the discrimination performance, and the letters U through Z give identification. The conditions are listed on the right of the graph. For example, A corresponds to a visual /ba/ auditory /ba/ followed by a visual /ba/ auditory /ba/. Predictions are of the speech is special model, which assumes no maintenance of separate auditory and visual feature information.



factor and the 2 levels of the auditory factor. Values of the 5 parameters were estimated to give the optimal predictions of the observed results, with the goodness of fit based on the RMSD between predicted and observed values. The model was fit to the average results (pooled across the 20 participants). The best fit of the FLMP to the average results gave an RMSD of .0805, a good fit considering that 26 data points are being predicted with just 5 free parameters. Figure 10 plots the observed versus the predicted outcomes of the FLMP for these 26 observations.

As noted, the application of the FLMP to the results carries the assumption that the output of the evaluation stage is identical in both the identification and discrimination tasks. This assumption captures the proposal that integration of the audible and visible sources does not modify or eliminate their representations given by the feature evaluation stage. If it did, then the model could not have accurately predicted the results with the same parameter values for identification and discrimination. According to the application of the model, the only difference between the two tasks is how the truth values provided by evaluation are combined. They are conjoined in the identification task and disjoined in the discrimination task.

To further test the assumption that the feature values produced by evaluation are maintained throughout integration and decision, we formulated an alternative model carrying the opposite assumption, and tested it against the same data. This speech is special model assumes that auditory and visual sources are blended into a single representation, without separate access to the auditory and visual representations. According to this model, the only representation that remains after a bimodal syllable is presented is the overall degree of support for the response alternatives. What is important for this model is that the overall degree of support for /da/ is functional independently of how much the auditory and visual modalities individually contributed to that support. It is possible to have two bimodal syllables made up of different auditory and visual components, but with the same overall degree of support for /da/. For example, a visual /ba/ paired with an auditory /da/ might give a similar degree of overall support for /da/ as an auditory /ba/ paired with a visual /da/. The FLMP predicts that these two bimodal syllables could be discriminated from one another. On the other hand, the speech is special model predicts that only the output of integration is available and, therefore, these

two different bimodal syllables could not be discriminated from one another. Figure 11 plots the observed versus the predicted outcomes for this model for these 26 observations. When formulated, this speech is special model gave a significantly poorer ($p < .001$) description of the results, with an RMSD of .1764.

These results substantiate the claim that information at evaluation maintains its integrity, and can be used independently of the output of integration and decision. Thus, it is inappropriate to believe that perceivers are limited to the output of integration and decision. Perceivers can also use information at the level of evaluation when appropriate. A related result consistent with this conclusion is the observed difference between the detection of temporal asynchrony between auditory and visual speech and the interval over which integration occurs. An observer can detect asynchrony at relatively short asynchronies whereas integration can occur across much longer asynchronies (Massaro & Cohen, 1993; Massaro et al., 1996).

Fowler (this volume) reviews other experiments exploring the relationship between identification and discrimination given conflicting and cooperating cues. Her gesture theory interpretation of the results from unimodal auditory speech experiments are opposite of what we concluded from the auditory-visual speech perception experiments. It is possible that the unimodal versus crossmodal conditions are responsible for the different conclusions (Massaro, 1987b, p. 110). More importantly, however, is the possibility that participants in the discrimination task were actually basing their judgments on the categorizations of the speech stimuli. In this case, observed discrimination of stimuli with cooperating cues would be poorer relative to stimuli with conflicting cues because the integrated percepts would be much more similar with conflicting cues than cooperating cues. Most importantly, however, is that quantitative model tests of the gesture theory were not carried out in the unimodal auditory speech experiments. Given that Fowler's gesture theory would predict the same outcome as the speech-is-special formulation, we can at least reject this theory in favor of the FLMP for the auditory-visual experiments.

Given the results of the identification/discrimination task, observers appear to maintain access to information at evaluation even though integration has occurred. Furthermore, the integration process does not modify the representation corresponding to

evaluation. Given a perceptual identification judgment that reflects the influence of both audible and visible speech, it is often concluded that a new representation has somehow supplanted the separate auditory and visual codes. However, we learned that we can tap into these separate codes with the appropriate type of psychophysical task. This result is similar to the finding that observers can report the degree to which a syllable was presented even though they categorically labelled it as one syllable or another (Massaro, 1987a; Massaro & cohen, 1993). If we grant that integration of audible and visible speech produced a new representation, then we see that multiple representations can be held in parallel. On the one hand, this result should not be surprising because a system is more flexible when it has multiple representations of the events in progress, and can draw on the different representations when necessary. On the other hand, we might question the assumption of representation altogether and view the perceiver as simply using the information available to act appropriately given the demands of the current situation (O'Regan & Noe, 2001; Dennett, 1991).

One might question why we have been so concerned about current theories of speech and language when the emphasis here is multisensory fusion or integration. The reason is that the theoretical framework we accept has important ramifications about how we can understand how information from several senses can be combined in speech perception. If indeed speech is special and categorically perceived, then it precludes many reasonable kinds of crossmodal integration (Massaro, 1987b, 1998a).

Learning in the FLMP

Figure 6 also illustrates how learning is conceptualized within the model by specifying exactly how the feature values used at evaluation change with experience. Learning in the FLMP can be described by the following algorithm (Friedman et al., 1995; Kitzis et al., 1999). The initial feature value representing the support for an alternative is initially set to .5 (since .5 is neutral in fuzzy logic). A learning trial consists of a feature (such as closed lips at onset) occurring in a test item followed by informative feedback (such as the syllable /ba/). After each trial, the feature values would be updated according to the feedback, as illustrated in Figure 6. Thus, the perceiver uses the feedback to modify the prototype representations and these in turn will become better tuned to the informative characteristics of the patterns being identified. This algorithm is

highly similar to many contemporary views of language acquisition (Best, 1993; Best et al., 2001; Werker & Logan, 1985)

Learning Speechreading

Given the importance of the visual modality for spoken language understanding, a significant question is to what extent skill in speechreading can be learned. In addition, it is important to determine whether the FLMP can describe speech perception at several levels of skill. Following the strategy of earlier training studies (e.g., Walden et al., 1977), long-term training paradigm in speechreading was used to test the FLMP across changes in experience and learning (Massaro, Cohen, & Gesi, 1993). The experiment provided tests of the FLMP at several different levels of speechreading skill.

Participants were taught to speechread 22 initial consonants in three different vowel contexts. Training involved a variety of discrimination and identification lessons with the

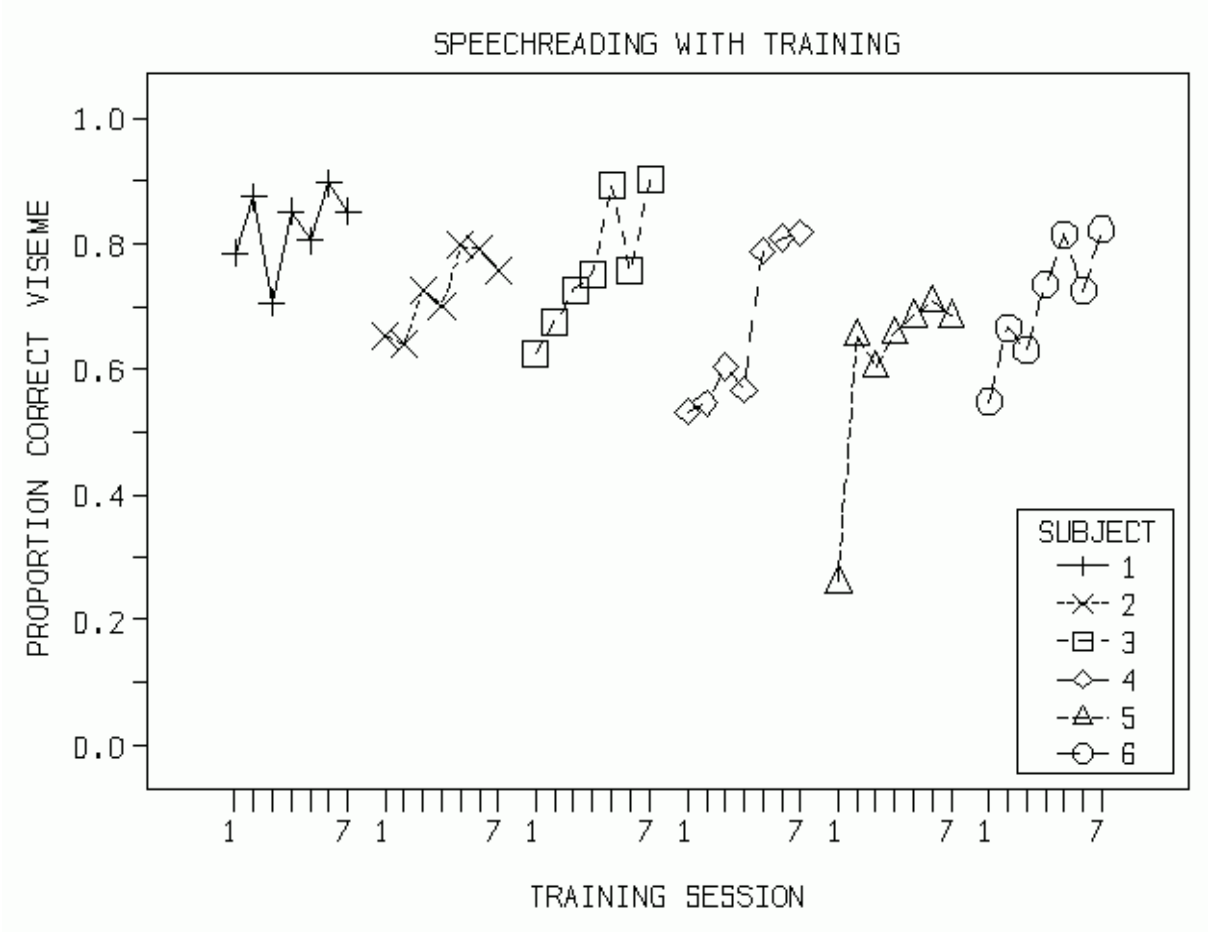


Figure 12. Proportion of correct viseme recognition of the initial consonant in the visible presentation of consonant-vowel syllables, as a function of the seven sessions of training in speechreading for each of the six participants.

consonant-vowel syllables. Throughout their training, participants were repeatedly tested on their recognition of syllables, words, and sentences. The test items were presented visually, auditorily, and bimodally, and presented at normal rate or three times normal rate. Participants improved in their speechreading ability across all three types of test items. Figure 12 gives their individual performance on the syllables across 7 sessions. The results are plotted in terms of correct viseme classifications, which groups similar visible consonants together. As can be seen in the figure, all six participants improved over training. Replicating previous results (Walden et al., 1977), the present study illustrates that substantial gains in speechreading performance are possible.

The FLMP was tested against the results at both the beginning and end of practice. According to the model, a participant would have better information after

training than before. To implement this gain in information, we simply assume more informative feature values before and after training. However, the audible and visible sources should be combined in the same manner regardless of training level. Consistent with these assumptions, the FLMP gave a good description of performance at both levels of speechreading skill. Thus, the FLMP was able to account for the gains in bimodal speech perception as the participants improved their speechreading and listening abilities. This success suggests that the FLMP and its distinction between information and information processing would provide a valuable framework for the study of language learning.

We have seen that speechreading can be taught and one important consideration involves the best method for instruction. Different models predict different outcomes for training participants to speechread in unimodal versus bimodal paradigms. Visible speech is presented alone in the unimodal paradigm followed by feedback whereas visible speech is paired with auditory speech in the bimodal paradigm. The Single Channel Model predicts that unimodal learning would be better, the Fuzzy Logical Model predicts no difference, and an extension of the less-is-more hypothesis predicts that bimodal learning would be better. The results of two recent experiments show that participants learn the same amount during unimodal and bimodal learning, supporting the FLMP (Geraci and Massaro, unpublished).

Language Learning

The FLMP paradigm offers a potentially useful framework for the assessment and training of individuals with language delay due to various factors such as the hard of hearing, autism, or specific language impairment (Massaro et al., 2000). An important assumption is that while information may vary from one perceptual situation to the next, the manner of combining this information—called 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.

These positive findings encourage the use of crossmodal environments for persons with hearing loss. Ling (1976), however, 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 profesionals might use bimodal training less often than would be beneficial. We have carried out two recent studies to evaluate multisensory instruction of speech perception and production. The working hypothesis is that speech perception and production will be better (and learned more easily) given bimodal input relative to either source of information presented alone.

Although there is a long history of using visible cues in speech training for individuals with hearing loss, these cues have usually been abstract or symbolic rather than direct representations of the vocal tract and articulators. Our goal is to create an articulatory simulation as accurate as possible, and to assess whether this information can guide speech production. We know from children born without sight that the ear alone can guide language learning. Our question is whether the eye can do the same, or at least the eye supplemented with degraded auditory information from the ear.

An Innovative and Valuable Application: Speech Training

One of the most promising features of our pedagogy and technology is that they can be easily applied to carry out tutoring of speech perception and speech production. Because of its accuracy and flexibility, our visible speech synthesis could be highly effective when implemented in a Speech Training Tutor to teach second language learners to perceive and produce spoken words, the skills needed for ordinary communication in everyday contexts. In addition, the same application can be used to carry out accent training for students across a wide range of second language competency. For example, beginning students would focus on perception and production of segments, words, and short phrases whereas advanced students might focus on accent neutralization. This spectrum of training is particularly important since training a second language is a labor-intensive task, traditionally involving significant one-on-one interaction with a teacher. Our automated system relieves the burden of finding a large number of teachers from a limited pool of qualified individuals.

Internal Anatomical Structures for “Visible Speech”

We have added internal structures to Baldi both for improved accuracy of visible speech and to pedagogically illustrate correct articulation (Cohen et al., 1998; Massaro et al., in press). Our immediate motivation for developing a hard palate, velum, teeth and tongue is their potential utility in language training. Many of the subtle distinctions

among segments are not visible on the outside of the face. The skin of our talking head can be made transparent or eliminated so that the inside of the vocal tract is visible, or we can present a cutaway view of the head along the sagittal plane. The articulators can also be displayed from different vantage points so that the subtleties of articulation can be optimally visualized. The goal is to instruct the student by revealing the appropriate articulation via the hard palate, velum, teeth and tongue, in addition to views of the lips and perhaps other aspects of the facial structure.

Second Language Learning

There is recent evidence that speech tutoring using the Baldi technology is effective. Light and Massaro (submitted) investigated its effectiveness against simulated methods of previous training procedures, for teaching the perception and production of non-native phonetic contrasts (Japanese learning English). The effects of training when presented with instruction revealing internal articulatory processes of the oral cavity versus instruction simulating the capabilities of a human tutor (providing just the normal view of the tutor's face) were contrasted.

In the articulators condition, Baldi first gave the participant verbal instructions on how to produce the /r/ segment (e.g. where to position the tongue with respect to the teeth, the shape of the tongue and lips, etc.). Baldi then showed the participant how to produce the word in the test pair involving the phoneme /r/, by allowing him/her to view the inside of Baldi's oral cavity during his production. Baldi asked the participant to try and produce the word on his/her own but the participant was not given feedback about his/her production ability at this time. The same procedure was carried out for the phoneme /l/.

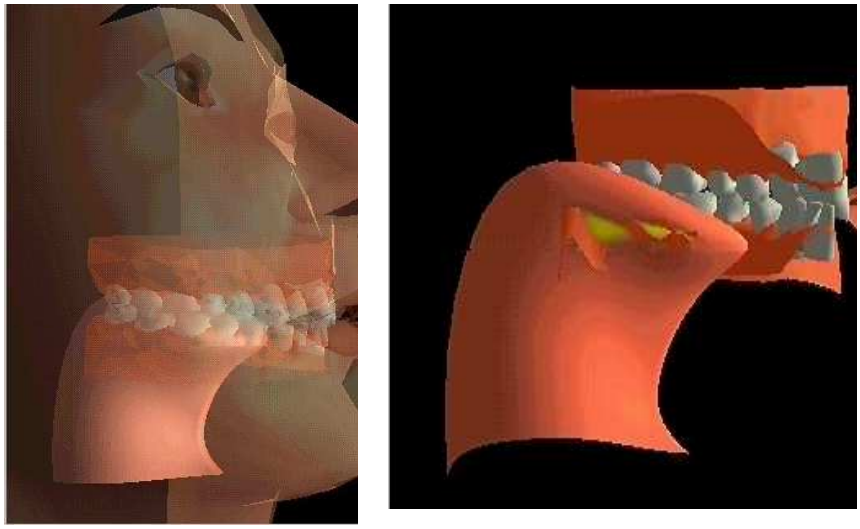
In teaching the participant how to produce the two sounds, four different views were shown: a view from the back of his head looking in, a side view of Baldi's mouth alone (static and dynamic), another side view of Baldi's whole face where his skin was transparent, as well as a frontal view of Baldi's face with transparent skin. As illustrated in Figure 13 for two of the views, each view gave the participant a unique perspective of the activity taking place during production.

We expected these multiple views to facilitate learning and to be more robust to individual differences. The order of presentation of the viewpoints was always the same.

First, Baldi told the participant that they were about to see a back view of his head, and that they should imagine his oral cavity as though it were his/her own. After Baldi produced the word in the pair containing the phoneme /r/ at a reduced speed rate of 63%, he asked the participant the repeat this word back to him after he gave the participant helpful tips about tongue positioning, etc. (e.g. “Remember to point your tongue, raise the sides of your tongue, and round your lips”). A speech recognition module was used to recognize this articulation and feedback was provided about the participants’ production ability via a happy or sad face. The same procedure was carried out for the word in the pair containing the phoneme /l/.

Next, Baldi informed the participant that they were about to see the inside of his mouth from a side profile. Baldi said both words from the pair at a reduced speed of 63%, while a static image of only his oral cavity from a side view was displayed next to him on the screen. Baldi then showed the participant a side profile again with dynamic changes during production. Baldi informed the participant further that they might find it easier to view this side image again if his whole face was displayed. The same reduced speed of 63% was used while Baldi produced the two words from the pair while his skin was transparent. He then asked the participant to try to say the word from the pair including the phoneme /r/ on his/her own. Feedback was given based on the recognition module. The participant was then asked to produce the word containing /l/ and feedback was given again.

Figure 13: *Two of the four presentation conditions giving a side view of Baldi's whole face where his skin was made transparent, and giving a side view of Baldi's tongue, teeth, and palate.*



Baldi told the participant to watch him as he said the two words for the last time. This time, a frontal view was shown. After Baldi produced the two words at a reduced speed of 63%, the tutoring phase ended by Baldi saying “Okay, now let’s see what you’ve learned”.

In the no-articulators condition, Baldi did not show the participant any inside views of the oral cavity. The instructions were simply to listen to the words and watch Baldi during his production. This process was to simulate the abilities of a human tutor. Other than this difference, the training procedure was the same for both groups. That is, the number of training presentations and tests was exactly the same in the two conditions.

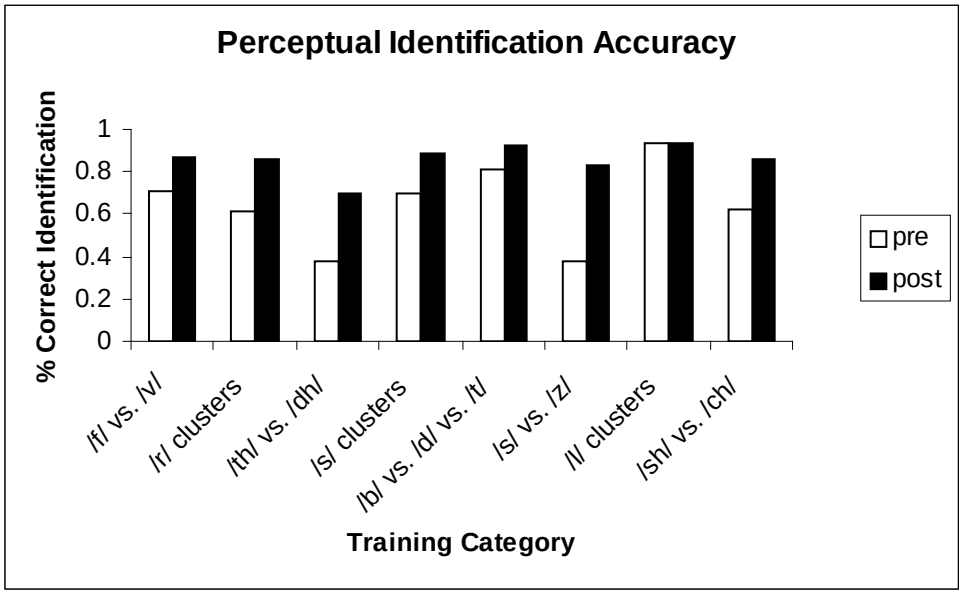
Japanese speakers of English as a second language were trained to identify and produce American English /r/ and /l/ in a pretest posttest design over a three week period. Three minimal word pairs were used in identification and production training (r/light, r/lip and grew/glue).

Results indicated varying difficulty with respect to word pair involved in training (r/light being the easiest to perceive and grew/glue showing the most difficulty). Most importantly, better learning occurred with a view of the internal articulators than with the traditional method of training. Generalization of learning to the production of new words was also revealed.

Speech Tutoring for Hard of Hearing Children

One of the original goals for the application of our technology was to use Baldi as a

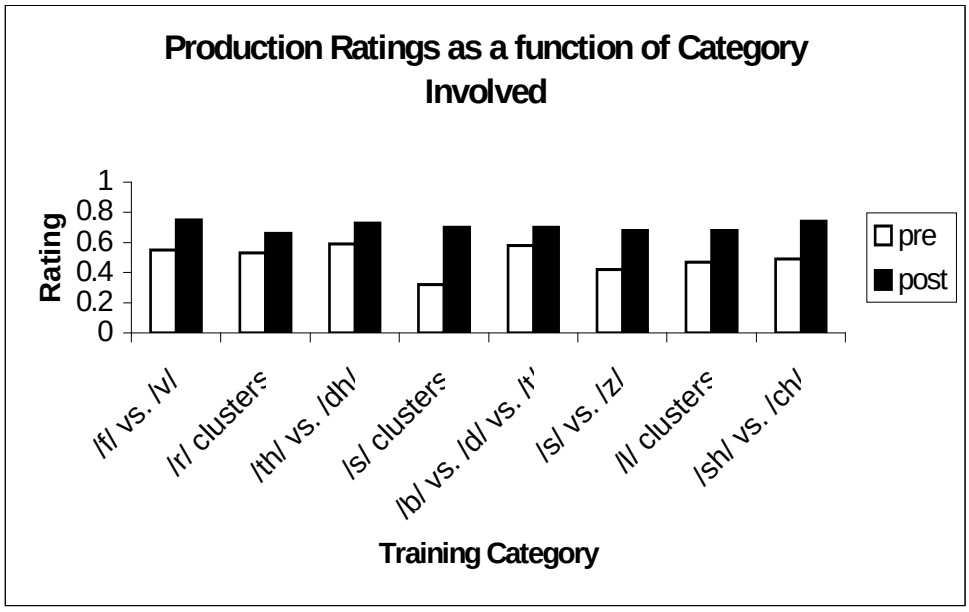
Figure 14. Percentage of correct identifications during pretest and posttest for each of the eight training categories.



language and speech tutor for deaf and hard of hearing children. Baldi's technology seems ideally suited for improving the perception and production of English speech segments. Baldi can speak slowly, illustrate articulation by making the skin transparent to reveal the tongue, teeth, and palate, and show supplementary articulatory features such as vibration of the neck to show voicing and air expulsion to show frication. Massaro and Light (submitted) implemented these features in a set for language exercises. Seven hard of hearing students between the ages of eight and thirteen were trained for six months on eight categories of segments (4 voiced vs. voiceless distinctions, 3 consonant cluster distinctions and 1 fricative vs. affricate distinction). Training included practice at the segment and the word level. Perception and production improved for each of the seven children. Figure 14 shows that perceptual identification accuracy improved for each of the eight types of distinctions.

There was also significant improvement in production of these same segments. The students' productions of words containing these segments were recorded and presented to native English college students. These judges were asked to rate the intelligibility of a word against the target text, which was simultaneously presented on the computer monitor. Intelligibility was rated on a scale from one to five (1:unintelligible, 2:ambiguous, 3:distinguishable, 4:unambiguous, 5:good/clear pronunciation). Figure 15

Figure 15. Intelligibility ratings of the pretest and posttest word productions for each of the eight training categories.



shows the judges' ratings transformed to a scale ranging from 0 to 1. According to these ratings, the children's speech production improved for each of the eight categories of segments. Speech production also generalized to new words not included in our training lessons. Finally, speech production deteriorated somewhat after six weeks without training, indicating that the training method rather than some other experience was responsible for the improvement that was found.

Retrospective

Speech perception has been studied extensively in the last decades, and we have learned that people use many sources of information in perceiving and understanding speech. Utilizing a general framework of pattern recognition, we have described the important contribution of visible information given in the talker's face and how it is combined with auditory speech. Speech perception is usually successful because perceivers optimally integrate several sources of information. In addition, audible and visible speech are complementary in that one source of information is most informative when the other source is not. These properties are well-described by a fuzzy logical model of perception (FLMP), a process model mathematically equivalent to Bayes theorem. The FLMP has also proven to provide a good description of performance in a wide variety of other domains of pattern recognition. For example, it describes how cues from both the face and the voice are evaluated and integrated to perceive emotion. The FLMP is also consistent with findings in neuroscience and provides an algorithmic description of two different neural mechanisms of multisensory processing.

Our empirical and theoretical research has encouraged us to apply our findings to facilitate language learning. Given that speechreading is highly functional in a variety of situations, it follows that the pursuit of visible speech technology could be of great practical value in many spheres of communication. We have developed a synthetic talking face, called Baldi, to achieve control over the visible speech and accompanying facial movements to study those visible aspects that are informative. Our talking head can be heard, communicates paralinguistic as well as linguistic information, and is controlled by a text-to-speech system or can be aligned with natural speech. Baldi, the animated talking agent, has innovative features and testing has proven him to be an effective speech tutor. These features include skin transparency controls that reveal the

vocal cavity, so that the lips, tongue and teeth can show how sounds are formed for better inspection, the head can be rotated at any angle, moved near and far, or displayed in cross section. Finally, the visual enunciation of speech can be paused, slowed, or replayed. In summary, the study of multisensory processing has not only uncovered fundamental facts about how we perceive and act in a world of many sensory inputs, it has led to a pedagogy and technology that is useful for language learning.

Acknowledgements

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), a Cure Autism Now Foundation Innovative Technology Award, and the University of California, Santa Cruz. The author would like to thank an anonymous reviewer for their helpful comments on the chapter.

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.*

Arbib, M. A. (2002). The mirror system, imitation, and the evolution of language. In K. Dautenhahn & C. L. Nehaniv (Eds.) *Imitation in animals and artefacts* (pp. 229-280). Cambridge, MA: MIT Press.

Aslin, R.N., Saffran, J.R., & Newport, E.L. (1998). Computation of conditional probability statistics by 8-month old infants. *Psychological Science,* 9, 321-324.

Best, C. (1993). Emergence of language-specific constraints in perception of non-native speech: A window on early phonological development. In: B. de Boysson-Bardies & S. de Schonen (Eds.). *Developmental neurocognition: Speech and face processing in the first year of life* (p. 289-304). Norwell, MA: Kluwer Academic Publishers.

Best, C.; McRoberts, G.& Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *Journal of the Acoustical Society of America, 109* (2), 775-795.

Campbell, Christopher S.; Schwarzer, Gudrun; Massaro, Dominic W. (2001). Face perception: An information processing perspective. In: Michael J. Wenger, Ed; James T. Townsend, Ed. *Computational, geometric, and process perspectives on facial cognition: Contexts and challenges* (pp. 285-345). Mahwah, NJ: Lawrence Erlbaum Associates.

Chomsky, N. (1980). Rules and representations. Oxford: Blackwell.

Cohen, M. M., Beskow, J., & Massaro, D.W. (1998). Recent developments in facial animation: An inside view. *Proceedings of the International Conference on Auditory-Visual Speech Processing—AVSP'98* (pp. 201-206). Terrigal, Australia.

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

Dennett, D. C. (1991). *Consciousness explained*. Boston: Little, Brown.

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,* 423-422.

Fodor, J. A. (1983). *Modularity of mind*. Cambridge, MA: Bradford Books.

Fodor, J. A. (2000). *The mind doesn't work that way: The scope and limits of computational psychology*. Cambridge, MA: MIT Press.

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

Friedman, D., Massaro, D.W., Kitzis, S.N., & Cohen, M.M., (1995) "A Comparison of Learning Models," *Journal of Mathematical Psychology*, 39, 164-178.

Geraci, K., & Massaro, D. W. (2002). *Teaching Speechreading: Is unimodal or bimodal training more effective?* Unpublished paper.

Gomez, R.L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences, 4*, 178-186.

Jordan, T .& Sergeant, P. (2000). Effects of distance on visual and audiovisual speech recognition. *Language and Speech, 43*, 107-124.

Kitzis, S.N., Kelley, H., Berg, E., Massaro, D.W., & Friedman, D., (1999), "Broadening the tests of learning models," *Journal of Mathematical Psychology*, 42, 327-355.

Liberman, A. M. (1996). *Speech: A special code.* Cambridge, MA: MIT Press.

Liberman, A.M., and Mattingly, I.G. "The motor theory of speech perception revised," *Cognition*, Vol. 21, 1985, pp 1-36.

Lieberman, P. (2000). *Human language and our reptilian brain: The subcortical bases of speech, syntax, and thought*. Cambridge, MA:. Harvard University Press.

Light, J., & Massaro, D.W. (2002). Learning to perceive and produce non-native speech. Unpublished paper.

Ling, D. (1976). *Speech and the hearing-impaired child: Theory and practice*. Washington, DC: Alexander Graham Bell.

Marcus, G. F. (2000). Pabiku and Ga Ti GA: Two mechanisms infants use to learn about the world. *Current Directions in Psychological Science*, 9, 145-147.

Massaro, D.W. (1987a). Categorical partition: A fuzzy logical model of categorization behavior." In S. Harnad (Ed.), *Categorical Perception*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Massaro D.W., (1989). 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

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

Massaro, D. W. (1998a). Categorical Perception: Important Phenomenon or Lasting Myth? Proceedings Paper: *International Congress of Spoken Language Processing* (Sydney, New South Wales, Australia) November 30 to December 6, 1998.

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

Massaro, D.W. (1999). Speechreading: Illusion or window into pattern recognition. *Trends in Cognitive Sciences*, 3, 310-317.

Massaro, D. W. (2002). Multimodal Speech Perception: A Paradigm for Speech Science In B. Granstrom, D. House, & I. Karlsson (Eds.) *Multilmodality in language and speech systems* (pp. 45-71). Kluwer Academic Publishers, Dordrecht, The Netherlands.

Massaro, D.W., Cohen, M. M., Tabain, M., Beskow, J., & Clark, R. (2002). Animated speech: Research progress and applications. In Vatikiotis-Bateson, E., Perrier, P., & Bailly, G. (Ed.). *Advances in audio-visual speech processing*. Cambridge: MIT Press (in preparation).

Massaro, D.W., and Cohen, M.M. (1983). Categorical or continuous speech perception: A new test. *Speech Communication, 2*, 15-35.

Massaro, D.W., & Cohen, M.M. (1993). Perceiving Asynchronous Bimodal Speech in Consonant-Vowel and Vowel Syllables, *Speech Communication, 13*, 127-134.

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

Massaro, D.W., Cohen, M. M., & Beskow, J. (2000). Developing and evaluating conversational agents. In J. Cassell, J. Sullivan, S. Prevost, & E. Churchill (Eds.) Embodied conversational agents. Cambridge, MA: MIT Press.

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

Massaro, D.W., Cohen, M.M., & Gesi, A.T., (1993) "Long-term Training, Transfer, and Retention in Learning to Lipread," *Perception and Psychophysics*, 53(5), 549-562.

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

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

Massaro, D.W., Cohen, M.M., & Smeele, P.M.T. (1996). Perception of Asynchronous and Conflicting Visual and Auditory Speech, *Journal of the Acoustical Society of America, 100*, 1777-1786

Massaro, D.W., Cohen, M.M., & Thompson, L.A. (1988). Visible language in speech perception: Lipreading and reading, *Visible Language, 22*, 8-31.

Massaro, D.W., and Ferguson, E.L. (1993). Cognitive Style and Perception: The Relationship between Category Width and Speech Perception, Categorization and Discrimination, *American Journal of Psychology, 103*, 25-49.

Massaro, D.W., & Light, J. (2002). Using Visible Speech for Training Perception and Production of Speech for Hard of Hearing Individuals. *Journal of Speech, Language, and Hearing Research*, submitted.

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

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

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

Mesulam, M. M., 1998. From sensation to cognition. *Brain, 121*, 1013-1052.

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.

Nearey, T. M. (1992). Context effects in a double-weak theory of speech perception. Special Issue: Festschrift for John J. Ohala. *Language & Speech, 35*, 153-171.

O'Regan, J.K., & Noe, A. (2001). A sesnsorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences, 24*, 939-1031.

Pashler, H. E. (1998). The psychology of attention. Cambridge, MA, US, The MIT Press.

Platt, J.R. (1964). Strong inference. *Science, 146*, 347-353.

Popper, K.R. (1959). *The Logic of Scientific Discovery*, New York: Basic Books.

Pullum, G. K., & Scholz, B. C. (2002). Empirical assessment of stimulus poverty arguments. *The Linguistic Review 19* (special issue, nos. 1-2), 9-50.

Quinn, P. C. (2002). Early categorization: A new synthesis. In U. Goswami (Ed.) *Blackwell handbook of childhood cognitive development* (pp. 84-101). Malden, MA: Blackwell publishing.

Robert-Ribes, J., Schwartz, J.-L., & Escudier, P. (1995a). A comparison of models for fusion of the auditory and visual sensors in speech perception. *Artificial Intelligence Review, 9*, 323-346.

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month old infants. *Science*, 274, 1926-1928.

Saffran, J. R.; Johnson, E. K.; Aslin, R. N.; Newport, E.L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, 70, 27-52.

Savage-Rumbaugh, S.; Shanker, S. G.; Taylor, T. J. (1998). *Apes, language, and the human mind*. New York, NY, Oxford University Press.

Smeele, P.M.T., Massaro, D.W., Cohen, M.M., & Sittig, A.C. (1998). "Laterality in visual speech perception," *Journal of Experimental Psychology: Human Perception and Performance*, 24, 1232-1242.

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. Cambridge, MA: MIT Press.

Trout, J.D. (2001). The biological basis of speech: What to infer from talking to the animals. *Psychological Review*, in press.

Walden, B., Prosek, R., Montgomery, A., Scherr, C. K., & Jones, C. J. (1977) Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research*, 20, 130-145.

Werker, J. & Logan, J. (1985). Cross-language evidence for three factors in speech perception. *Perception and Psychophysics, 37* (1), 35-44.