Bimodal Speech Perception: A Progress Report

Dominic W. Massaro

University of California - Santa Cruz

Abstract. The study of audible and visible speech perception has grown into a cottage industry. For us, it has evolved into our prototypical paradigm for inquiry (Massaro, 1987, 1992, Massaro & Cohen, 1995). I provide a description of our paradigm and the fuzzy logical model of perception (FLMP), give principles for empirical inquiry, and describe the progress we have made. Bimodal speech perception conforms to a prototypical pattern recognition situation, is well-described by the FLMP across a wide range of individual and task variability, and is highly robust across a variety of situational contexts.

Keywords. Human pattern recognition, bimodal integration models, individual variability, task variability, modality independence

1 THEORETICAL FRAMEWORK

1.1 Perception and Pattern Recognition

I use the term pattern recognition to describe what we commonly mean by perception, recognition, identification, and categorization. These four acts appear to entail the same fundamental processes. All of them can be characterized as choice or pattern recognition situations in which a person, given a stimulus event, makes one of a set of alternative choices. Pattern recognition has been found to be fundamental in such different domains as playing chess, examining X-rays, and reading text. Our operating assumption is that it is also central to multimodal speech perception. I now describe our theoretical framework, which is then formalized in the context of a specific model of pattern recognition.

The theoretical framework is the information processing approach, which assumes that there is a sequence of processing stages in spoken language understanding. Auditory speech perception is hierarchical with transduction along the basilar membrane, sensory cues, and perceived attributes. A single cue can influence several perceived attributes. The duration of a vowel provides information about vowel identity, prosodic information such as stress, and the syntactic role of the word in the sentence. Another example is that the pitch of a talker's voice is informative about both the identity of the talker and intonation. The best known example of multiple cues to a single perceived attribute in speech is the case of the many cues for the voicing of a medial stop consonant (Cohen, 1979; Lisker, 1978). These include the duration of the preceding vowel, the onset frequency of the fundamental, the voice onset time, and the silent closure interval. Visible speech also provides multiple cues and one of the focal points of this conference is that both sound and the visible mouth movements of the talker influence perception of speech segments. In addition, both sets of cues can be used in speech recognition by machine.

Three basic principles are compatible with our current understanding of perception. These principles are a) perception is a process of inference, b) perceptual inference is not deductively valid, and c) perceptual inferences are biased (Massaro & Cowan, 1993). The first two assumptions go back to at least Helmholtz whereas the third simply means that a given perceptual system prefers some interpretations relative to others. In speech perception, we are biased to perceive the speech input in terms of the segments of our language and to perceive the segments as comprising a meaningful communication. These three principles describe the perceiver's solution to the inverse mapping problem: the perceiver's goal is to solve the problem of what environmental situation exists given the current conflux of sensory cues. To structure our information processing analysis of spoken language understanding, we use a specific theoretical model that has received substantial support from a variety of experiments in speech perception.

1.2 Fuzzy Logical Model of Perception (FLMP)

The central assumptions of FLMP are that a) there are multiple sources of information supporting speech perception and b) the perceiver evaluates and integrates all of these sources to achieve perceptual recognition. Well-learned patterns are recognized in accordance with a general algorithm, regardless of the modality or particular nature of the patterns (Massaro, 1987). As shown in Figure 1, pattern recognition consists of three operations: feature evaluation, feature integration, and decision. Continuously-valued features are evaluated, integrated, and matched against prototype descriptions in memory, and an identification decision is made on the basis of the relative goodness of match of the stimulus information with the relevant prototype descriptions.

The schematic form of this model has been modified from previous representations to clarify the assumptions of the model and to account for an important finding in bimodal speech perception. To emphasize that evaluation of the auditory features and visual features occurs independently, we show the respective evaluations as occurring in separate boxes. That is, the degree of visible mouth opening at the onset of the syllable can be evaluated independently of whether or not there is also auditory information and the nature of the auditory information if present. Given that there is some evidence that perceivers can use the onset differences between the visual and auditory (or tactile) sources as a cue to voicing of stop consonants (Breeuwer & Plomp, 1986), we show this cue as a separate source of information that is evaluated independently of the auditory and visual sources. Thus, there is cross-modal information available about voicing even though the evaluation of place features, for example, can occur independently along the auditory and visual modalities. This cross-modal information would be integrated with the modality-specific sources of information in the same

multiplicative manner as assumed by the FLMP.

An expanded factorial design is usually used to test this model and to contrast its predictions against those of other models. In the study of speech perception by eye and ear, this design includes the unimodal conditions as well as all factorial combinations of the bimodal conditions. In one typical experiment, five levels of audible speech varying between /ba/ and /da/ are crossed with five levels of visible speech varying between the same alternatives. The audible and visible speech are also presented alone giving a total of 25 + 5 + 5 = 35 independent stimulus conditions. Subjects are instructed to listen and to watch the talker, and to identify the syllable as either /ba/ or /da/.

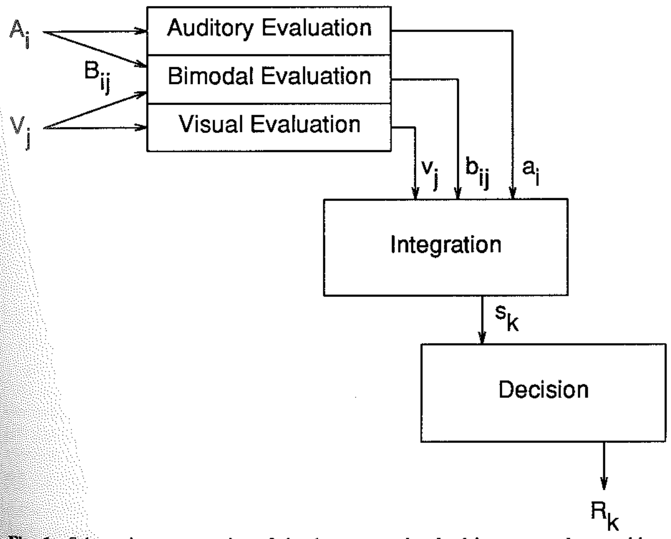

Fig. 1. Schematic representation of the three stages involved in perceptual recognition. The three stages are shown to proceed left to right in time to illustrate their necessarily successive but overlapping processing. The sources of information are represented by uppercase letters. Auditory information is represented by A_i and visual information by V_j. In addition, information about the temporal asynchrony between the auditory and visual information is indicated by B_{ij}. The evaluation process transforms these sources of information into psychological values (indicated by lowercase letters a_i, v_j, and b_{ij}) These sources are then integrated to give an overall degree of support for a given alternative s_k. The decision operation maps this value into some response, R_k, such as a discrete decision or a rating.

The points in Figure 2 give the mean proportion of identifications for a prototypical subject. The identification judgments changed systematically with changes in the audible and visible sources of information. The likelihood of a /da/

identification increased as the auditory speech changes from /ba/ to /da/, and analogously for the visible speech. Each source had a similar effect in the bimodal conditions, relative to the corresponding unimodal condition. In addition, the influence of a one source of information was greatest when the other source is neutral or ambiguous. (Although the curves in the middle panel appear to be parallel to one another, this perception is illusory. The spread between the curves is about four times greater in the middle than at the end of the auditory continuum.)

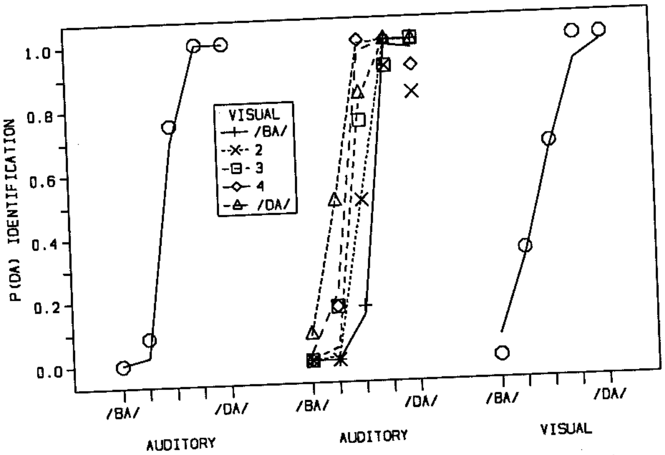

Fig. 2. The points give the observed proportion of /da/ identifications for a typical observer in the auditory-alone (left panel), the factorial auditory-visual (center panel) and the visual-alone(right panel) conditions as a function of the five levels of the synthetic auditory and visual speech varying between /ba/ and /da/. The lines give the predictions for the FLMP.

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 a given participant in the task. Three points are circled in the figure to highlight the conditions in which the third level of auditory information is paired with the first level of visual information. When presented alone, $P(/ba/|A_3)$ is about .2 whereas $P(/ba/|V_1)$ is about .8 When these two stimuli occur together, $P(/ba/|A_3V_1)$ is about .5 Both the FLMP and a simple averaging integration can predict this result.

Other observations, however, allow us to reject the averaging alternative. Figure 4 gives the results for another participant in the task. Three points are circled in the figure to highlight the conditions in which the third level of auditory information is paired with the first level of visual information. When presented alone, $P(/ba/|A_2)$ is about .8 and $P(/ba/|V_2)$ is about .8 When these two stimuli occur together, $P(/ba/|A_2V_2)$ is about 1. This so-called superadditive result (the bimodal is more extreme than either unimodal response proportion) allows us to reject the averaging alternative in favor of the FLMP.

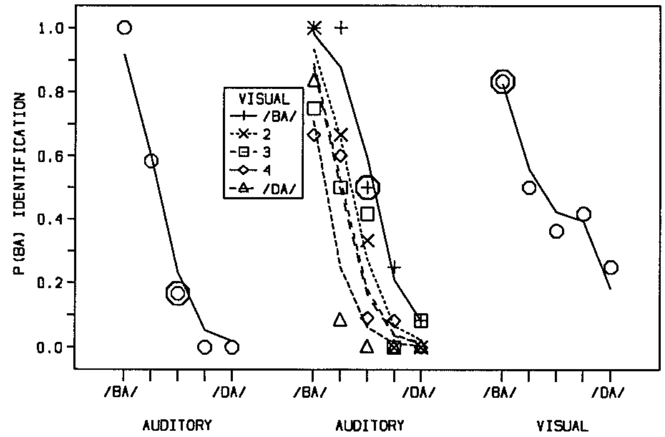

Fig. 3. The points give the observed proportion of /ba/ identifications for an observer in the auditory-alone (left panel), the factorial auditory-visual (center panel) and the visual-alone(right panel) conditions as a function of the five levels of the synthetic auditory and visual speech varying between /ba/ and /da/. The three circled points give two unimodal conditions and the corresponding bimodal condition. The relationship among the three points can be explained by either an averaging or multiplicative integration.

The lines in Figures 2, 3, and 4 give the predictions of FLMP. Applying the FLMP to all of the results, both sources are assumed to provide continuous and independent evidence for the alternatives /ba/ and /da/. Given a prototype's independent specifications for the auditory and visual sources, the value of one source cannot change the value of the other source at the prototype matching stage. The decision operation determines the relative merit of the /ba/ and /da/ alternatives leading to the prediction that

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

where $P(/da/|A_iV_j)$ is the probability of a /da/ response to a particular A_iV_j combination and a_i and v_j are the auditory and visual support for the alternative /da/. The predicted probability of a /ba/ identification in a two-alternative response task is

$$P(/ba/|A_iV_j) = \frac{(1-a_i)(1-v_j)}{a_i v_j + (1-a_i)(1-v_j)} = 1 - P(/da/|A_iV_j). \tag{2}$$

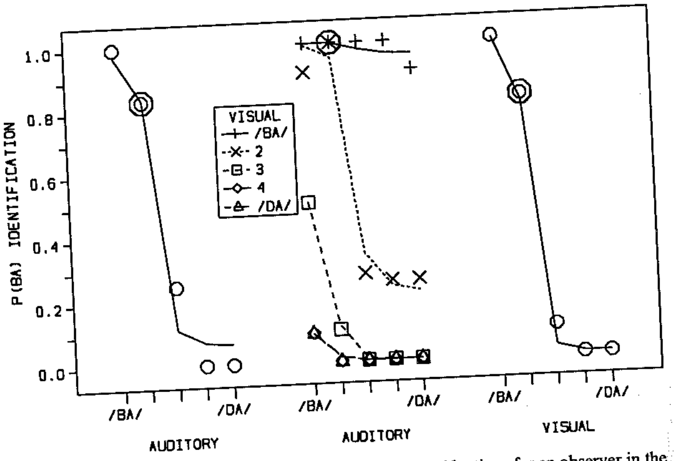

Fig. 4. The points give the observed proportion of /ba/ identifications for an observer in the auditory-alone (left panel), the factorial auditory-visual (center panel) and the visual-alone(right panel) conditions as a function of the five levels of the synthetic auditory and visual speech varying between /ba/ and /da/. The three circled points give two unimodal conditions and the corresponding bimodal condition. The relationship among the points cannot be explained by an averaging integration, but can be described by a multiplicative one.

The model requires 5 parameters for the visual feature values and 5 parameters for the auditory feature values.

The FLMP is fit to the individual results of each subject, by finding the parameters that maximize the goodness of fit. The goodness-of-fit of the model is given by the root mean square deviation (RMSD)—the square root of the average squared deviation between the predicted and observed values. For all three participants, the FLMP provides a good description of the identifications of both the unimodal and bimodal syllables. Figure 5 shows the best fitting parameters for each subject. As can be seen in the figure, the parameter values differ for the different subjects but, for each subject, they change in a fairly systematic fashion across the five levels of the audible and visible synthetic speech.

1.3 Testing among Alternative Models

Given the value of falsification and strong inference (Massaro, 1987), it is essential to contrast one model with other models that make alternative assumptions. Success in this enterprise will require quantitative predictions and a fine-grained analysis of the results. Within the domain of speech perception, one alternative to the FLMP is the theory of categorical perception, which Massaro (1987) formalized as the categorical model of perception (CMP). The CMP was fit to the individual results in the same manner as in the fit of the FLMP. The CMP gave a poor description of the observed results. The CMP is mathematically equivalent to both a single channel model in which only a single source of information is used at any one time and an averaging model in which the sources of information are averaged to obtain the overall support. Thus, its falsification also provides evidence against these hypotheses. Given that the CMP represents three different models, its falsification is a particularly informative outcome.

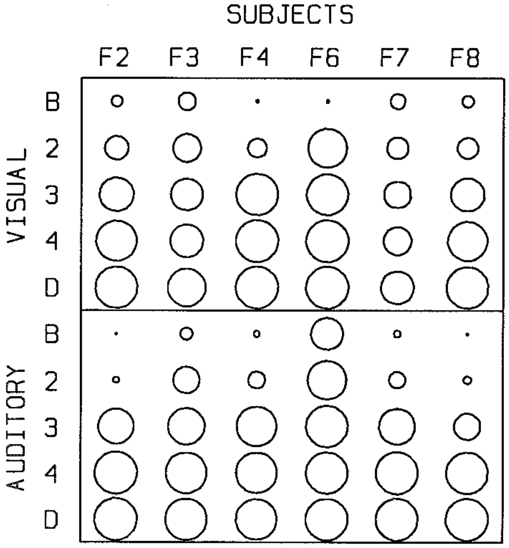

Fig. 5. Parameter values for subjects shown in Figures 2,3,4,6, and 7 for the support of the response alternative /da/ as a function of the five levels along the visual and auditory speech continua. The parameter value is given by the area of each circle.

1.4 Normative Models

I have argued for process models of our experimental tasks and the testing among alternative models (Massaro, 1987). In addition, the process or descriptive models under consideration should be evaluated for their optimality properties or to what extent predictive behavior is assumed to be as good as possible. Massaro (1987) and Massaro and Friedman (1990) illustrated how the FLMP could be formulated as mathematically equivalent to Bayes' Theorem—an optimal model for combining multiple sources of information. "If perceptual systems evolved to integrate multiple sources of information, we might expect that the integration would be highly efficient and productive (i. e., optimal)" (Massaro, 1987).

1.5 Proximate versus Ultimate Causation

An important distinction made in evolutionary biology is between proximate and ultimate causes of (or influences on) behavior (Alcock, 1993). Proximate causes are immediate or close in time and describe psychological processes that influence behavior. For example, we might ask what environmental information the gannet (a large seabird) uses to signal closing its wings when landing on water. Ultimate causes might concern why the gannet closes its wings when landing—what evolutionary significance it might have. The psychologist's and engineer's concern with proximate causes might make the framework of evolution less applicable to psychological inquiry. To the extent proximate and ultimate causes are related, however, an evolutionary framework may be productive in psychological inquiry. It would be very helpful to know about the role that audible and visible speech played during the evolution of spoken language. We may never know the answer to this question, however.

1.6 Falsifiability of Models

A potentially devastating charge is that the FLMP is not falsifiable (Cutting, Bruno, Brady, & Moore, 1992). It has been claimed that somehow the FLMP seems capable of predicting a plethora of functions and also has the magic power to absorb random noise. We have criticized their demonstrations, logic, and interpretations in Massaro and Cohen (1993). We provided an empirical test of whether the FLMP absorbs random variability by evaluating the goodness of fit as a function of the number of observations per data point. Group results are not necessarily less variable than the results of individual subjects. The total number of observations per data point is important in both cases. The goodness of fit of the FLMP actually increases with increases in the number of observations, as it should. Massaro and Cohen (1993) also illustrated the dangers of averaging across subjects, and propose that model tests should be carried out on the results of individual data. We also showed that the FLMP is falsifiable: it cannot predict a plethora of performance functions and it does not have the magic power to absorb random noise.

2. INDIVIDUAL VARIABILITY

Experience in the methodological and theoretical intricacies of bimodal speech perception has convinced me of productive strategies for psychological inquiry. I have formulated these in various set of prescriptions. These prescriptions, I believe, are a productive route to the formulation of a set of general principles describing speech perception by eye and ear. One set of prescriptions involves analyzing and manipulating additional well-known variables having to do with individual variability. These variables differ from the typical independent variables that are aimed at influencing psychological processes, and are usually considered to be orthogonal to the questions of interest. Individual variability plays a central role in evolutionary theory and inquiry, and psychology—the study of behavior that has necessarily evolved—should be no different. The proposed variables having to do with individual variability are 1) individuals, 2) development and aging, 3) languages, 4) sensory-impaired individuals, 5) patients with brain trauma, 6) personality, and 7) experience and learning.

In the next sections, the processing of bimodal speech is assessed in the context of different sources of individual variability. We will see that individual variability modulates the phenomena of interest. However, these modulations are well-described within our theoretical framework. Thus, the variability we see actually highlights what is consistent in the information processing of speech by ear and eye.

2.1 Individuals

This prescription has to do with looking at individual differences and similarities. It is well known that individual differences exist and usually our experimental investigations are aimed at reducing them as much as possible. This procedure may preclude discovery of important properties of the processes of interest. Individual differences can be meaningless or misleading, however, unless the investigator has available a good process model of the task. We all know individuals differ, but we want to know how they differ; Individuals might simply differ with respect to the information they have or they might differ in how they process the information. The meaningfulness of the parameter values justify an important distinction between information and information processing. The parameter values represent how informative each source of information is. The integration and decision algorithms specify how this information is processed.

This distinction plays an important role in locating several sources of variability in our inquiry. The variability in the information is analogous to the variability in predicting the weather. There are just too many previous contributions and influences to allow quantitative prediction. In addition, small early influences can lead to dramatic consequences at a later time (the butterfly effect in chaos theory). However, once this variability is accounted for (by estimating free parameters in the fit of the model, for example), 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 level of audible or visible speech is for a given individual, we can predict how the two sources of information are integrated. In addition, the model does take a stand on the evaluation process in the sense that it is assumed that the sources of information are evaluated independently of one another. In previous papers, I have shown that individual differences could be attributed solely to information differences and not to information processing differences (Massaro, 1992).

To illustrate the huge variability that exists in the information and the constancy of the information processing, Figures 6 and 7 give the results of two subjects tested in this task. Figure 5 gives the parameter values determined in the fit of the FLMP to their results. As can be seen in Figure 6, this participant was primarily influenced by the visual information whereas the opposite was the case for the

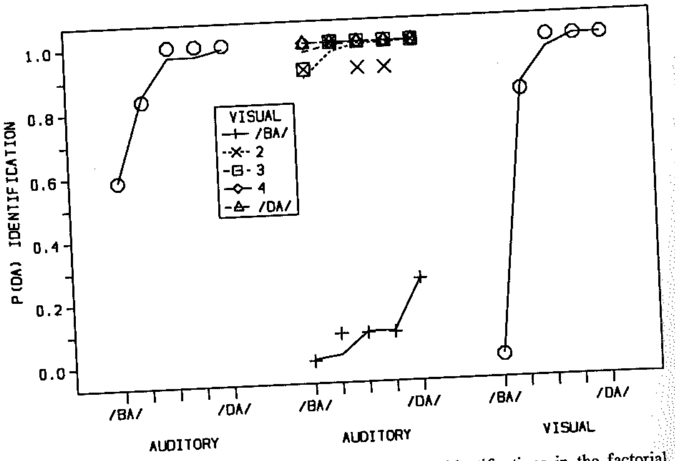

Fig. 6. The points give the observed proportion of /da/ identifications in the factorial auditory-visual conditions as a function of the five levels of the synthetic auditory and visual speech varying between /ba/ and /da/. The left panel plots the auditory variable on the x-axis and the visual variable as the parameter of the graph. The right panel reverses the plot of these two variables. Results for a participant primarily influenced by visible speech. The lines give the predictions for the FLMP.

participant in Figure 7. Given these large differences, one might expect that the information processing would also be very different for the two participants. This was not the case, however, because the FLMP gave a good description of both sets of results. It is impressive that the FLMP is able to give a good account of both observers with simply a change in parameter values reflecting the information value of each source (see Figure 5). These differences should correct the common belief that a good theory should necessarily be parameter free or have a fixed set of parameter values. Results of this type inform us about what we can expect to predict (information processing) and what cannot be specified in advance (information).

2.2 Development and Aging

This prescription is related to the study of individual differences and experience because it addresses developmental and aging effects. Analogous to individual differences and similarities, a process model is valuable for the evaluation of developmental and aging differences. Lifespan differences and similarities are observed in bimodal speech perception. There are significant overall differences between the perception of bimodal speech across the life span. However, when the categorization performance is accounted for by a process model of the task, the differences are accounted for completely in terms of the the information from the auditory and visual sources, rather than differences in how that information is combined.

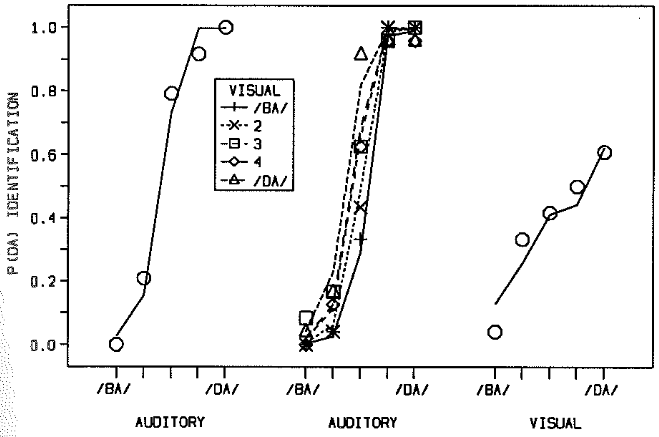

Fig. 7. The points give the observed proportion of /da/ identifications in the factorial auditory-visual conditions as a function of the five levels of the synthetic auditory and visual speech varying between /ba/ and /da/. The left panel plots the auditory variable on the x-axis and the visual variable as the parameter of the graph. The right panel reverses the plot of these two variables. Results for a participant primarily influenced by auditory speech. The lines give the predictions for the FLMP.

To illustrate a lifespan contrast, senior citizens were tested in the same expanded factorial design described in Section 1.2, except that there were 8 possible response alternatives. Thirteen senior citizens were instructed to listen and to watch the talker, and to identify the syllable as one of eight alternatives: /ba/, /da/, /ga/, /va/, /ða/, /bda/, /dba/, and other. Their results were similar to those found with college students, except for one significant difference. Most senior citizens do not respond with the response cluster /bda/, whereas most college students do. However, there were a few college students that resembled senior citizens and a few senior citizens that resembled college students. More importantly, the FLMP gave a good description of all participants. The FLMP has had similar successes with preschool children. The fundamental processes involved in pattern recognition as described by the FLMP appear to exist at age 3 and to remain constant for the next seven decades.

Although differences in the processes underlying speech perception do not appear to change across the lifespan, there are significant differences in performance. A distinction between information and information processing is central to understanding these changes across the lifespan. There were significant differences in the information value of audible and visible speech as a function of age, but no differences in information processing. With respect to the information value of auditory and visual speech, there are substantial changes with age. These changes are readily explained by increased experience with age and changes in the sensory systems with aging. Preschool children are still acquiring speech-perception skills. They do not lipread as well as adults and they less capable of discriminating differences in auditory speech. The aging differences in performance are accurately described in terms of the feature evaluation stage of the FLMP. A given source of information is less informative for preschool children than for adults. This is not surprising given that it is experience with speech that permits speech data to be treated as information. We can expect that the prototype descriptions of the distinguishing characteristics of speech will increase in resolution with experience.

Aging, on the other hand, can decrease the resolution of the sensory systems resulting in less accurate speech perception. Luckily, the availability of multiple sources of information usually precludes a catastrophic breakdown even with a fairly severe loss of a given source. For example, visible speech from the talker's lips appears to compensate for hearing loss with age. Some older adults report that they hear the TV better with their glasses on. The value of the FLMP is that it not only describes how speech perception might be accomplished, it provides a framework for understanding how it changes with development and aging. These findings have developmental implications both within and outside the field of speech perception (Massaro, 1992, Massaro & Burke, 1991).

2.3 Languages

Related to the work on developmental differences, the next prescription has to do with cross-linguistic differences. Following the approach that we have taken previously, we can look for differences and similarities as a function of language. The FLMP allows one to make an important distinction between *information* and *information processing* (Massaro, 1987, 1992). One component of information corresponds to the outcome of evaluation: how much a particular stimulus presented to a given input channel supports the various alternatives. One component of information processing corresponds to the process of integration: how the various sources of information are combined. Perceivers of different linguistic groups might differ with respect to either or both of these characteristics. Consider the second level along a synthetic auditory speech continuum between /ba/ and /da/. This stimulus might support the alternative /ba/ for one language significantly more than for another language. In experimental studies, we cannot hope to equate the amount of support for a given category across different linguistic groups. We simply synthesize the same range of speech stimuli for the different languages and have the subjects categorize these stimuli in their native language.

Our paradigm addresses issues of both information and information processing. Although it has been claimed that the Japanese are less influenced by visible speech, results from our standard task falsifies this claim (Massaro et al., 1993). Figure 8 gives the results of a typical Japanese observer. As can be seen in the figure, these results are similar to those from Americans and are also well-described by the FLMP. Given the unique phoneme inventories and phonologies of the different languages, however, we will probably observe different response patterns from the different linguistic groups. The FLMP makes a very strong prediction concerning information processing. Regardless of the amount of /ba/-ness from a given source of information, it will be combined with other sources of information, as prescribed by the integration and decision operations. With respect to the integration of audible and visible speech, the information value of a given modality might differ but it will be combined with the other modality in the same manner for all languages. The model allows for linguistic differences in the truth values or degrees of support assigned at the level of evaluation, but not in the processes of integration and decision. Thus, testing the FLMP against the results also tests whether linguistic differences can be located entirely at the evaluation stage of processing.

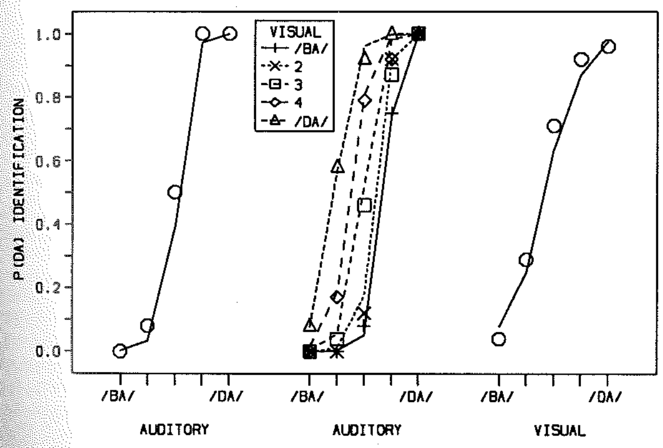

Fig. 8. The points give the observed proportion of /da/ identifications for a typical Japanese observer in the auditory-alone (left panel), the factorial auditory-visual (center panel) and the visual-alone(right panel) conditions as a function of the five levels of the synthetic auditory and visual speech varying between /ba/ and /da/. The lines give the predictions for the FLMP.

The expanded factorial design helps illustrate the cross-linguistic predictions given by the FLMP. For ease of exposition, consider a task with the two alternatives /ba/ and /da/. If a Japanese talker identifies some auditory syllable as /ba/ 70% of the time and some visible syllable 80%, then the bimodal syllable composed of these two auditory syllables should be identified as /ba/ about 90% of the time. This same prediction holds for a talker of English or a talker of any other language. Cross-linguistic differences in information will more or less guarantee that the unimodal syllables will be identified differently by talkers of different languages. The FLMP simply predicts the nature of integration and decision, not the evaluation of the unimodal syllables. These evaluations require the free parameters in the model because we cannot predict beforehand how much a given source of information will support a given alternative.

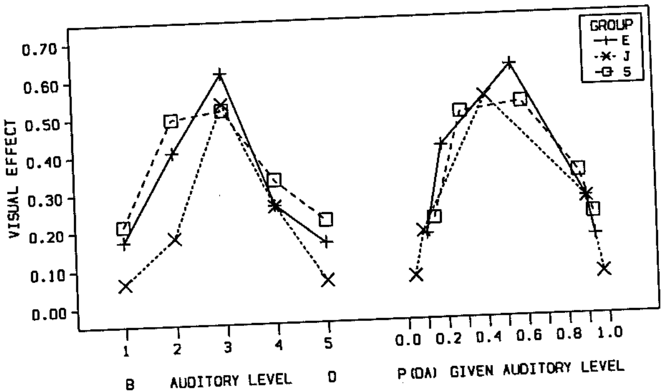

Fig. 9. Visual effect as a function of the auditory level (left plot) and as a function of the probability of a /da/ response given the auditory level (right plot) over participants in the bimodal condition for American English (A), Japanese (J) and Spanish (S) talkers.

The methodology of the Massaro et al. experiments allows us to separate information differences from information processing differences. The experiments with native-English Americans, Spanish Japanese, and Dutch talkers (Massaro & Cohen, 1990; Massaro et al., 1993, 1995) indicated important contributions of auditory and visual speech in bimodal speech perception. Most importantly, the experiments revealed both differences and similarities in performance across the different languages. The English talkers gave mostly /ba/ and /da/, /bda/, /ða/, and /va/ responses. Visible speech had a strong influence on the perceptual judgments of the English talkers. Visible articulations on the /ba/ end of the continuum increased the number of /ba/ judgments. The number of /bda/ judgments increased when a visible /ba/ was paired with an auditory syllable from the /da/ end of the continuum. Visible /da/ articulations increased the likelihood of /da/, /ða/, and /va/ responses. Although the Japanese talkers were also highly influenced by visible speech, they gave a different set of responses. These differences between Japanese and English talkers reflects the differences in the phonemic repertoires, phonetic realizations of the syllables, and phonotactic constraints in the two languages.

Analogous results were found with Spanish and Dutch talkers. Figure 9 plots the visual influence on performance as a function of the auditory information for three different languages. These results show that talkers of different languages are similarly influenced by visible speech. In addition, the contribution of one source is largest to the extent the other source is ambiguous. The details of these judgments are nicely captured in the predictions of the FLMP, which gives a significantly better fit than other extant models. The experiments substantiate the distinction made between information and information processing. The information made available by evaluation naturally differs for different languages. However, the information processing involved in integration and decision is identical across languages. Thus, these results provide some of the first findings that the FLMP provides a good account of bimodal speech perception in languages other than English.

2.4 Sensory-Impairment

It is important to determine to what extent the processing of information changes across cases of sensory impairment. The value of the present perspective is apparent in providing supplement sources of information for the disabled individual. One such source of information that has been used for profoundly deafened individuals is cochlear prosthesis. This involves electrical stimulation of residual auditory nerve fibers using intracochlear electrodes (Shannon, 1983; Simmons, 1985). Usually, some parameters of the speech signal are derived and used to drive the location, amplitude, and rate of electrode stimulation. Although this information is not usually sufficient for complete communication, remarkably good performance can be obtained when it is combined with lipreading. In one study, a patient with a multiple-channel cochlear implant was tested with just electrical stimulation, just lipreading, or both of these sources of information (Dowell, Martin, Tong, Clark, Seligman, & Patrick, 1982). Twelve consonants were presented in a VCV context with the vowel /a/. Twenty observations were made on each consonant spoken by a female talker in one test and a male talker in the other. The results are of the form of a 12 by 12 confusion matrix under each of the three presentation conditions. The FLMP was applied to the results of the female and male talkers separately and gave a very good description of the results. Thus, the model is capable of describing the integration of lipread information with electrical stimulation to the cochlear in the same manner as with normal hearing. There is a promising potential for utilizing the present approach as a framework for the assessment and rehabilitation of communication disorders.

2.5 Brain Trauma

Using a videotape of our standard task (Massaro & Cohen, 1990), Ruth Campbell tested an integrative visual agnostic HJA in bimodal speech perception. Seven years after his stroke, HJA still cannot recognize the faces of his closest relatives by sight. Even so, his speech perception by ear and eye appears to be representative of normal adults at his age. Figure 10 gives the observed and predicted proportion of identifications for this participant as a function of the five levels of the synthetic auditory and visual speech varying between /ba/ and /da/. HJA was influenced by visible speech, as well as audible speech, and his results were well-described by the FLMP. These results match those of other participants of the same age. This result indicates very little interaction between face recognition and speech perception. This dissociation should not be surprising because different sources of information are probably used in the two domains.

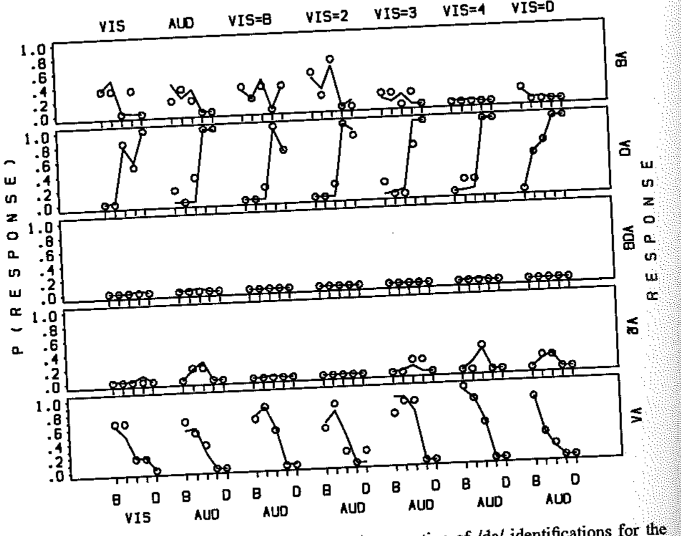

Fig. 10. Observed (points) and predicted (lines) proportion of /da/ identifications for the visual-alone (left panel), auditory-alone (second from left panel). and the factorial auditory-visual (other 5 panels) and conditions as a function of the five levels of the synthetic auditory and visual speech varying between /ba/ and /da/. Results for prosopagnosic participant HJA. The lines give the predictions for the FLMP.

2.6 Personality

Massaro and Ferguson (1993) questioned whether speech categorization and discrimination are influenced by differences in personality. Category width is a cognitive variable that purportedly reveals individual differences in categorization strategy. Broad and narrow subjects differ in terms of category width—a measure of the extent they will accept exemplars as good instances of a category. Our goal was to test the hypothesis that category width would be related to how subjects behave in different speech perception tasks. Analyses of the results in terms of a) discrimination and feature evaluation of auditory and visual information in speech events, b) integration of these sources of information, c) the process of decision, and d) subjective preference for a two-choice versus a nine-choice response method, revealed no effects attributable to category width. The results from both male and female and broad and narrow subjects supported the predictions made by the FLMP. Given the common processes involved in speech and other pattern perceptual recognition tasks, it appears that fundamental processes involved in pattern recognition are unlikely to vary with personality measures, such as category width.

2.7 Experience and Learning

Given the importance of the visual modality for spoken language understanding, a significant question is to what extent skill in lipreading can be learned. In addition, it is important to determine whether the FLMP can describe speech perception at several levels of skill. A long-term training paradigm in lipreading was used to test the FLMP across changes in experience and learning (Massaro, Cohen, & Gesi, 1993). The experiment also extended tests of the model to include the prediction of confusion matrices, as well as performance at several different levels of skill. The predictions of the FLMP were contrasted with the predictions of a Pre-Labeling Integration Model (PRLM) developed by Braida (1991). Subjects were taught to lipread 22 initial consonants in three different vowel contexts. Training involved a variety of discrimination and identification lessons with the consonant-vowel syllables. Repeated testing was given on syllables, words, and sentences. The test items were presented visually, auditorily, and bimodally, and presented at normal rate or three times normal rate. Subjects improved in their lipreading ability across all three types of test items. Replicating previous results, the present study illustrates that substantial gains in lipreading performance are possible. Relative to the PRLM, the FLMP gave a better description of the confusion matrices at both the beginning and end of practice. The FLMP was able to account for the gains in bimodal speech perception as the subjects improved their lipreading and listening abilities.

3. TASK VARIABILITY

Another set of prescriptions involves analyzing and manipulating additional well-known variables having to do with the tasks. Generally, we need to know to what extent the processes uncovered in the task of interest generalize across 1) modalities, 2) domains, 3) items, 4) responses, 5) instructions, 6) and tasks. The processes involved in bimodal language processing, for example, might be revealed more readily by addressing these variables in addition to those traditionally manipulated. The hope is that the interactions with these variables will inform and constrain the kinds of processing mechanisms used to explain the basic observations. Differences and similarities with respect to these variables should be informative. Investigators should be concerned with whether their findings generalize across different aspects of the experimental task. Evaluating behavior changes as a function of variations within tasks and across different tasks provides valuable information on several fronts. The study of performance in various tasks improves the chances of gaining insights into underlying mechanisms. We should not expect superficial results, such as whether or not an interaction between two independent variables was significant, to generalize across tasks. However, we do expect our theories to account for performance across variability in the task domain. Our understanding of psychological mechanisms is good to the extent we can predict across different tasks.

One of the most engaging issues of the last decade has been Modularity of Mind (Fodor, 1983). This thesis makes the very strong prediction that mechanisms uncovered in one domain will not be adequate to describe performance in a different domain. Of course, this thesis is most directly tested by studying behavior as a function of modality and domain variability.

3.1 Modalities

With respect to speech perception by eye and ear, the question is to what extent similar processes occur in speech perception via other modalities. There is substantial evidence that the processes found in speech perception by ear and eye generalize to electrical stimulation of cochlear implants and tactile stimulation on the skin.

There are several powerful modes of communication that function very much like bimodal speech perception. In addition to the widespread use of sign language, there are other forms of communication that supplement rather than replace speech. Cued speech (Mohay, 1983), for example, supplements lipread information with manual hand movements for communicating to the hearing-impaired. For individuals without sight and sound, the Tadoma method involves the receiver placing his or her hands on the face and neck of the talker (Norton, Schultz, Reed, Braida, Durlach, Rabinowitz, & Chomsky, 1977).

Given that manual gestures appear to have properties and functions that are strikingly similar to speech (McNeill, 1985), we have extended our framework to study the integration of a pointing gesture with audible speech (Thompson & Massaro, 1986, 1994). Following the strategy of an expanded factorial design, subjects were presented with gesture, speech, and both sources of information. The task was to indicate whether a ball or a doll was intended by the talker. An auditory continuum of five levels was made between the words ball and doll. The gestural information was also varied by pointing to either the ball or doll objects. The results for both preschoolers and college students were essentially identical to those found in audible and visible speech. There were significant effects of both sources of information and the judgments followed the predictions of the FLMP.

3.2 Domains

Modularity of mind has been the center of much controversy, and the issue of the modularity of speech perception is equally relevant Mattingly & Studdert-Kennedy, 1991). Do the processes uncovered in speech perception occur in other domains? Table 1 lists the different domains that have supported the processes assumed by the FLMP. In addition to speech perception, the FLMP has given a good description in a variety of domains such as letter and word recognition in reading, object identification, sentence interpretation, the perception of depth, memory retrieval, reasoning, and the recognition of affect.

Table 1.0. Domains of Evidence for FLMP.

SPEECH PERCEPTION	Acoustic Features Phonological Constraints Lexical Constraints Syntactic Constraints Semantic Constraints Semantic & Syntactic Information Audible & Visible Speech Speech & Gesture
READING	Letter Features Orthographic Constraints Spelling-to-Speech Constraints Lexical Constraints
CATEGORIZATION	Cups & Bowls
VISUAL PERCEPTION	Cues to Exocentric Distance
MEMORY RETRIEVAL	Letters & Semantic Cues Implicit Memory Explicit Memory
SOCIAL EVENTS	Person Impression
REASONING	Conjunction Fallacy
EMOTION	Facial Cues to Affect Facial and Vocal Cues to Affect

We have learned that the FLMP generalizes very nicely across different communication modalities. It is also of interest to determine to what extent the theoretical framework generalizes across widely disparate performance domains. Table 1 lists the domains in which the FLMP has been tested. Support for the FLMP has been found in speech perception, reading, object recognition, sentence interpretation, the recognition of affect, memory, and decision making.

3.3 Items

It is important to know to what extent some observed phenomena generalize across all items, or whether they are limited to the few items tested in most experiments. Thus, it is of interest to determine to what extent speech perception by ear and eye hold for different segments of the language, words, sentences, and conversation. Item analyses can illuminate to what extent the contribution of visible speech varies as a function of different speech contrasts. There has been a tradition of comparing the perception of vowels and consonants, and it is of interest to question whether these two classes of segments behave differently in bimodal speech perception. In one study, bimodal speech perception of two consonants /ba/ and /da/ was compared to two vowels /i/ and /u/. The results indicated that visible speech had a larger impact in the perception of consonants than vowels (Massaro & Cohen, 1993; Cohen & Massaro, 1995). However, the FLMP gave an adequate description of performance with both sets of stimulus items.

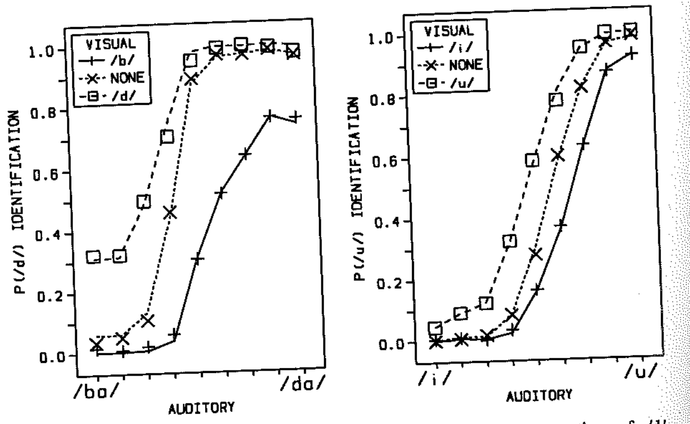

Fig. 11. Observed (points) and the FLMP's predicted (lines) proportion of /d/ identifications as a function of the auditory and visual levels of the speech event for a stop consonant continuum between /ba/ and /da/ and a vowel continuum between /i/ and /u/.

3.4 Responses

Given an interest in how the input is processed, it is easy for the investigator to neglect the role of output processes. It is important, however, to assess to what extent models generalize across different types of responses. Most speech perception work uses a traditional psychophysical measure of response probability, and our work has continued this tradition. In addition, we have also used rating responses and response accuracy to test the FLMP under different response measures.

Most experiments involve discrete response alternatives, but it is important to assess other behavioral measures. According to the FLMP, people have information about the degree to which a given alternative is present rather than just information about which alternative is present. In one study, participants were asked to either identify the test item as /ba/ or /da/ or rate the degree to which the test item corresponded to /ba/ versus /da/ on a nine-point scale. The results of both types of tasks are shown in Figure 12. As can seen in the figure, the results are very similar for both types of dependent measures. The FLMP predicts that the output R_k is equivalent to a rating response normalized to be in the interval zero to one. Not surprisingly, the FLMP accurately predicts the finding that rating judgments follow the same form as the identification judgments (Massaro & Ferguson (1993).

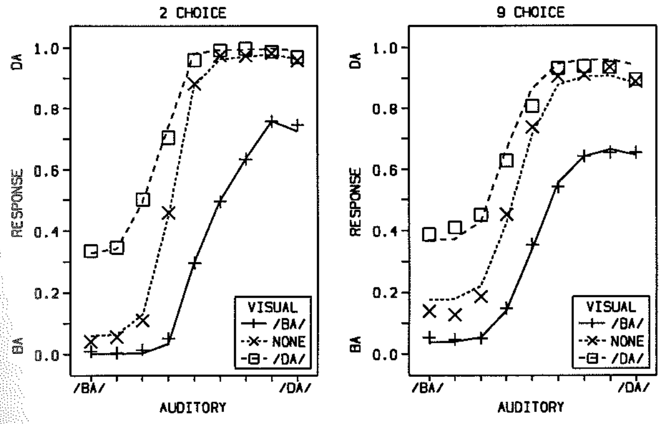

Fig. 12. Proportion of /da/ identifications in the 2 choice (/ba/ versus /da/) task (left panel) and average rating between /ba/ and /da/ alternatives (on a 9-point scale linearly normalized to be between the values zero and one).

The FLMP also makes strong predictions about the accuracy of performance in that it predicts that two sources of information can be more informative than just one. Thus accuracy should be greater given bimodal speech than given the audible or visible speech presented unimodally. We have verified that word recognition can be more informative given two sources of information relative to just one. The stimuli were 420 one syllable English words on a laser videodisk recorded by Bernstein and Eberhardt (1986). They included words from the Modified Rhyme Test (Kreul, Nixon, Kryter, Bell, Lang & Schubert, 1968) as well as additional words. Given that identification of the auditory words is close to perfect, we also presented the words at one-third of their original duration. For this presentation rate, only every third visual and/or auditory frame from the videodisk was presented. At this speeded presentation rate, two sources of information was clearly superior to just one. Subjects averaged 55% correct given the auditory words, 4% correct given the visual words, and 73% correct given the bimodal presentation.

Finally, reaction times can be an informative independent variable. The RTs provided converging evidence for the process of integrating auditory and visual speech. The RTs increase with increases in the ambiguity of the speech event. RTs increase in the middle of a unimodal continuum in which there is maximum ambiguity. Similarly, RTs are very long when the two sources of information support different alternatives, creating an ambiguous event.

3.5 Instructions

This prescription assesses to what extent observed behavior can be changed by instructions. The value of instructional manipulations is apparent in our study of bimodal speech perception. Although subjects might use only the auditory or only the visual, both sources are integrated and influence perceptual judgments. A strong test of integration is to assess to what extent subjects can bypass integration. Can subjects voluntarily identify the speech event on the basis of just a single source of information? If subjects cannot selectively attend to a single source, we have strong evidence for integration. Subjects have been tested in the bimodal speech recognition task under two selective attention conditions: the subjects identified only what they saw or only what they heard. Even with these instructions, subjects were influenced by the modality they were trying to ignore. Integration appears to be a natural form of processing bimodal speech.

3.6 Tasks

It is important to know whether our view of bimodal speech perception holds up under a wide variety of tasks. Roberts and Summerfield (1981) used selective adaptation to obtain a somewhat indirect measure of the nature of the influence of visible speech. In selective adaptation, listeners are exposed to a number of repetitions of an "adapting" syllable and then asked to identify syllables from a speech continuum between two speech categories. Relative to the baseline condition of no adaptation, the identification judgments of syllables along the speech continuum are pushed in the opposite direction of the adapting syllable (a contrast effect). Roberts and Summerfield (1981) employed different adaptors to evaluate auditory adaptation along a /be/ to /de/ continuum. After adaptation, subjects identified syllables from an auditory continuum between /be/ and /de/. Roberts and Summerfield found no evidence for cross-modal adaptation. The visual adaptors presented alone produced no adaptation along the auditory continuum. Similarly, equivalent levels of adaptation were found for an auditory adaptor and a bimodal adaptor with the same phonetic information. The most impressive result, however, was the adaptation obtained with the conflicting bimodal adaptor. The auditory /be/ paired with visual /ge/ adaptor produced adaptation equivalent to the auditory adaptor /be/. This result occurred even though the subjects usually experienced the bimodal adaptor as /de/ (unfortunately, the authors did not provide an exact measure of the subject's identification of the adaptors). Thus, adaptation follows the auditory information and is not influenced by the visual information or the phenomenal experience of the bimodal syllable.

The adaptation results provide strong support for the FLMP assumption that the auditory and visual sources are evaluated independently of one another. The same results provide strong evidence against competing accounts of bimodal speech perception. Robert-Ribes (this conference) proposes a motor space recoding theory in which the auditory and visual inputs are projected onto a motor representation space. It follows from this model that a bimodal syllable composed of an auditory /ba/ and a visual /ge/ should produce a different type of adaptation than an auditory /be/ adaptor. A similar outcome should occur according to the

interactive activation model (McClelland & Elman, 1986). This models predicts that the bottom-up activation of the phoneme /d/ would provide top-down activation of the features representing that phoneme. It follows that subjects should not have adapted to the bimodal syllable experienced as /de/ in the same manner as their adaptation to an auditory syllable experienced as /be/.

We have also demonstrated that observers have access to modality-specific information at evaluation even after integration has occurred (Massaro & Cohen, 1995). Participants performed both an identification and a discrimination task. Participants found it easy to discriminate two syllables that they identified as the same syllable. This result is similar to the finding that observers can report the degree to which a syllable was presented even though they categorically label it as one syllable or another. The FLMP assumes that continuous auditory and visual information is maintained in unaltered form at the evaluation stage even after the two sources have been combined for identification at the integration stage. A system is robust when it has multiple representations of the events in progress, and can draw on the different representations when necessary.

Robustness of FLMP Integration

Rather than summarize our progress report, I close by mentioning experimental and theoretical findings that substantiate the robustness of the FLMP integration of audible and visible speech. By robustness is meant that integration occurs across a broad range of stimulus conditions, and that the FLMP gives a good account of this integration.

The ecological correspondence between audible and visible speech is not necessary for integration. Speech can be created both naturally by human talkers and synthetically by machines. We have studied all 4 possible pairings between these two sources of information. Visible speech from a synthetic talking head is integrated with natural auditory speech, and so on across all 4 possible combinations. The FLMP gives a good account for all of these cases.

Integration, by which we mean FLMP integration, is also robust across relatively large temporal asynchronies between the audible and visible speech. The FLMP gives a good account of integration out to roughly one-quarter of a second (Massaro et al., unpublished). FLMP integration is also robust across relatively large differences in the spatial location of the face and the voice (Massaro, 1992).

Finally, people also appear to integrate information from an inverted face in the same manner as they integrate information from a normal upright face (Massaro & Cohen, unpublished).

Acknowledgment

The research reported in this paper and the writing of the paper were supported, in part, by grants from the Public Health Service (PHS R01 NS 20314), the National Science Foundation (BNS 8812728), and the graduate division of the University of California, Santa Cruz.