THE DEVELOPMENT OF INTERSENSORY PERCEPTION:
Comparative Perspectives

Edited by

David J. Lewkowicz
New York State Institute for Basic Research in Developmental Disabilities

Robert Lickliter
Virginia Polytechnic Institute and State University

LEA LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
1994 Hillsdale, New Jersey Hove, UK

CHAPTER FIFTEEN

Bimodal Speech Perception Across the Life Span

Dominic W. Massaro
University of California, Santa Cruz

The theme of this book involves the development of intersensory perception and how it occurs across different species. The present chapter is concerned with the understanding of spoken language, a skill so far unique to humans. Although spoken language is usually thought of as unimodal, there is now good evidence for its multimodal nature. Hence many speech scientists are turning from the study of auditory speech perception to spoken language understanding. Within the domain of spoken language, the scientist views understanding as being influenced by many sources of information. In addition to auditory information and situational context, two important sources in face-to-face communication are gestures and facial movements of the lips and tongue.

The multimodal nature of speech perception should come as no surprise once it is acknowledged that speech perception is simply a form of pattern recognition. More generally, pattern recognition consists of evaluating and integrating several sources of influence or sources of information. Consider the recognition of the word *performance* in the sentence "The actress was praised for her outstanding performance." The recognition of the critical word is achieved by integrating a variety of sources of information. In information-processing terms, bottom-up sources include the auditory and visual stimulus properties. The evidence for multiple auditory sources of information in speech perception comes from the discovery of many different cues or features that contribute to the discriminable contrasts found in speech. The phonetic difference between voiced and voiceless stop consonants in medial position appears to have up to 18 acoustic characteristics that could function as acoustic

features. For example, the perceived distinction between /aga/ and /aka/ can be influenced by the preceding vowel duration, the silent closure interval, the voice-onset time, and the onset frequency of the fundamental.

In addition to these bottom-up sources, phonological, lexical, and sentential constraints function as top-down sources of information. Massaro and Cohen (1983) asked subjects to identify a liquid consonant in different phonological contexts (Massaro, 1989b). Each speech sound was a consonant cluster syllable beginning with one of the four consonants /p/, /t/, /s/, or /v/ followed by a liquid consonant ranging from /l/ to /r/, followed by the vowel /i/. Both the bottom-up source (the acoustic properties of the liquid consonant) and the top-down source (the initial consonant) had a strong influence on performance. The probability of an /r/ response increased systematically as the test phoneme varied from /l/ and /r/. Subjects responded /r/ more often given the context /t/ than given the context /p/. Similarly, there were fewer /r/ responses given the context /s/ than given the context /p/. An important result was that the phonological context effect was greatest when the information about the liquid was ambiguous. More generally, given multiple sensory sources of information, we can expect the least ambiguous source to have the greatest influence.

Recent research has also found that visible speech also contributes to perception. One is more likely to visually attend to a person speaking if the environment is noisy and distracting or if one has a hearing loss. There is evidence that the deaf gain phonological information primarily from a code based on how words look when they are said. Similar phonological information is obtained by both deaf and hearing children, suggesting that information can be derived from visual as well as auditory speech (O'Connor & Hermelin, 1981). Although these demonstrations of a significant contribution of visible speech involve the case in which the auditory signal is either absent or degraded by white noise, or is presented to hearing-impaired listeners, there seems to be a positive influence even when the auditory speech is perfectly intelligible. Perhaps the most convincing demonstration of the intersensory property of speech is the McGurk effect (McGurk & MacDonald, 1976). Viewing a videotape of a speaker uttering the syllable /ga/, which has been dubbed with the auditory syllable /ba/, the viewer reports "perceiving" or "hearing" the syllable /va/, /da/, or /ðə/. This demonstration and subsequent sophisticated experimental studies have shown that the perception of face-to-face spoken language is truly intersensory—it is influenced by both sight and sound.

The availability of auditory and visual information in face-to-face speech perception illustrates the multiplicity of cues in a natural situation. One might expect the validity and reliability of the cues to be perfectly correlated. However, variation and noise in the environment and in the sensory systems involved could alter these two sources of information differentially because

they are somewhat independent. For example, the perceiver might have a varying view of the speaker's face, and extraneous background noise might vary randomly over time. Thus, for any particular speech event, its auditory quality is not necessarily correlated with its visual quality. In experiments, one can accordingly manipulate the auditory and visual sources independently of one another. Given that the two cues are not perfectly correlated in the natural situation, the factorial manipulation of the cues will not be completely new to the perceiver. Central to our experimental inquiry are expanded factorial designs that independently manipulate multiple aspects of speech input jointly and in isolation.

Our research follows the logic of falsification and strong inference. We formulate a set of alternative hypotheses about speech perception by eye and ear. To pursue this goal, we build on the developments in information-integration theory (Anderson, 1981, 1982) and mathematical model testing (Townsend & Ashby, 1983). The alternative hypotheses—called binary oppositions—are considered to be central to a complete description of the phenomenon. A hierarchical set of binary oppositions concerning speech processing is illustrated in Fig. 15.1.

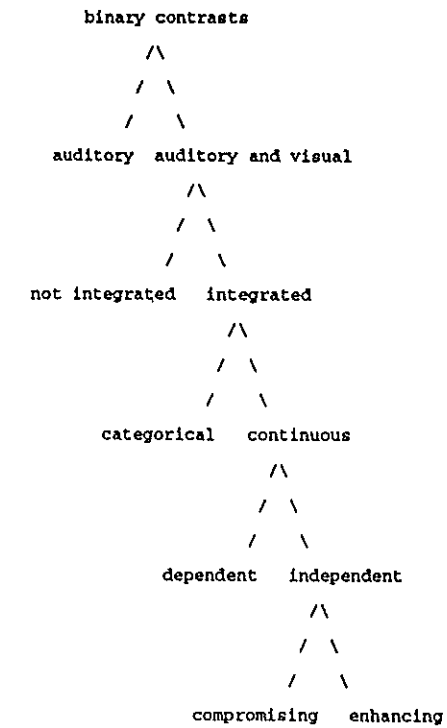

FIG. 15.1. Tree of wisdom illustrating binary contrasts central to the domain of speech perception by eye and ear.

The top contrast addresses the question of whether speech is unimodal or bimodal. As noted in the preceding paragraph, there is good evidence that perceivers can be influenced by visible speech. The next contrast asks the important question of whether the two modalities are actually combined (integrated) in speech perception. It is possible that only a given modality has an influence at any given time. Our research has shown, however, that perceivers naturally integrate the two modalities when presented with bimodal speech (Massaro, 1987, 1989a). The next question has a long history in the study of speech perception and asks whether perceivers are limited to categorical information. Although categorical perception has been the accepted dogma for several decades, there is now overwhelming evidence that perceivers of speech have continuous information (Massaro, 1992). The fourth contrast addresses the degree to which the auditory and visual sources of information are evaluated independently of one another. Our research has shown that the two sources are evaluated independently of one another, and then integrated. The last contrast concerns the nature of the integration process. The tests of mathematical models of performance has supported the idea of an enhancing integration process in which several ambiguous sources of information can create a less ambiguous event for the perceiver (what we call an enhancing integration; Massaro, 1987, 1989a).

In some cases, the contrast at one level is dependent on the answers to the contrasts at higher levels. As an example, the issue of whether or not multiple sources of information are integrated (combined) in perception requires that multiple sources, rather than just a single source, be available to the perceiver. Similarly, questions about the nature of integration are meaningful only if integration occurs. On the other hand, whether sources of data provide continuous or categorical information can be assessed regardless of whether or not they are processed independently.

As illustrated in Fig. 15.1, a falsification and strong-inference strategy of inquiry guides the present research. Results are informative only to the degree that they distinguish among alternative theories. Thus, the experimental task, data analysis, and model testing are devised specifically to reject some theoretical alternatives. Following the research strategy of strong inference (Platt, 1964), a fuzzy logical model of perception (FLMP), an auditory dominance model (ADM), and a categorical model of speech perception (CMP) are formalized and tested against the results. Our experience has convinced us of the superiority of the FLMP, and we begin with the description of this model.

FUZZY LOGICAL MODEL OF PERCEPTION

We believe that human recognition of speech is robust because there are usually multiple sources of information that the perceiver evaluates and integrates to achieve perceptual recognition. The results from a wide variety

of experiments can be described within a framework of a fuzzy logical model of perception (FLMP). The assumptions central to the model are:

1. Each source of information is evaluated to give the degree to which that source specifies various alternatives.
2. The sources of information are evaluated independently of one another.
3. The sources are integrated to provide an overall degree of support for each alternative.
4. Perceptual identification follows the relative degree of support among the alternatives.

Support for these assumptions can be found in Massaro (1987, 1989a, 1989b, 1990).

According to the FLMP, well-learned patterns are recognized in accordance with a general algorithm, regardless of the modality or particular nature of the patterns. Three operations assumed by the model are illustrated in Fig. 15.2. The three stages are drawn as overlapping in Fig. 15.2 in order to illustrate that the stages can be overlapping in time. In the first operation, continuously valued features are evaluated to give some degree of support for each of the relevant alternatives. The relevant alternatives consist of the

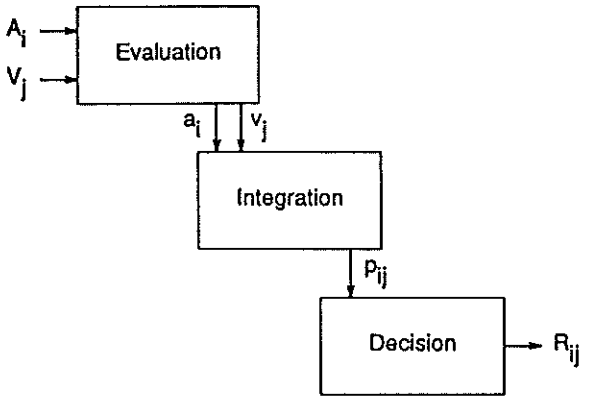

FIG. 15.2. Schematic representation of the three stages involved in perceptual recognition. The three stages are drawn as overlapping in order to illustrate that the stages can be overlapping in time. The evaluation of the auditory source of information A_i produces a truth value a_i, indicating the degree of support for alternative R. An analogous evaluation occurs for the visual source V_j. Integration of the truth values gives an overall goodness of match p_{ij}. The response R_{ij} is equal to the value p_{ij} relative to the goodness of match of all response alternatives.

fundamental speech categories in the perceiver's language. We use syllables as test alternatives because of the important role of syllables in language perception (Massaro, 1975). There is evidence that the evaluation of the auditory sources occurs independently of the evaluation of the visual source. The next operation combines or integrates these degrees of support made available by the evaluation process. The outcome of integration provides an overall goodness-of-match with each of the relevant alternatives. The decision operation uses the outputs of the integration operation to make a discrete decision about which alternative was presented. The choice of a given alternative is based on the relative goodness-of-match with the relevant alternatives.

In a typical experiment, synthetic auditory and visual speech are manipulated in an expanded factorial design. This design is shown in Fig. 15.3. Given an expanded factorial design, we must describe how the identification of each bimodal syllable occurs as a function of the processing of the unimodal syllables that compose it. This design is more powerful than simple factorial in differentiating among different models of categorization behavior (Massaro & Friedman, 1990). The onsets of the second and third formants were varied to give an auditory continuum between the syllables /ba/ and /da/. In analogous fashion, we systematically varied parameters of an animated face 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.

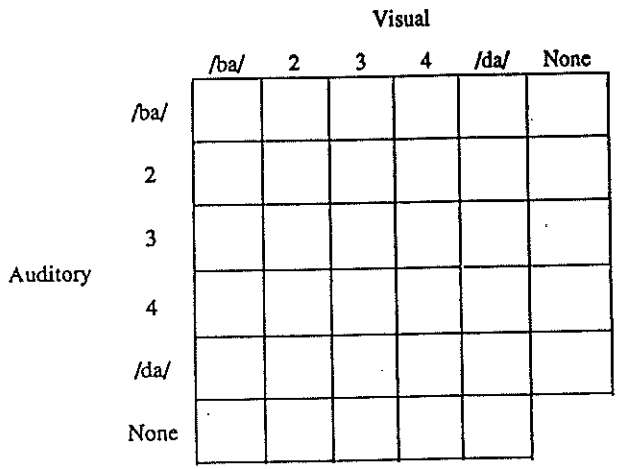

FIG. 15.3. Expansion of a typical factorial design to include auditory and visual conditions presented alone. The five levels along the auditory and visible continua represent auditory and visible speech syllables varying in equal physical steps between /ba/ and /da/.

Applying the FLMP to this task, both sources are assumed to provide continuous and independent evidence for each of the response alternatives. Response alternatives simply correspond to alternatives in the perceiver's language that match the auditory and visual sources of information. Defining the onsets of the second (F2) and third (F3) formants as the important auditory feature and the degree of initial opening of the lips as the important visual feature, the prototype for /da/ might be something like:

/da/: Slightly falling F2-F3 & Open lips.

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

/ba/: Rising F2-F3 & Closed lips.

A prototype would exist for each of the potential response alternatives.

Given a prototype's independent specifications for the auditory and visual sources, the value of one source cannot change the value of the other source. The integration of the features defining each prototype is evaluated according to the product of the feature values. If a_{Di} represents the degree to which the auditory stimulus A_i supports the alternative /da/, that is, has Slightly falling F2-F3; and v_{Dj} represents the degree to which the visual stimulus V_j supports the alternative /da/, that is, has Open lips, then the outcome of prototype matching for /da/ would be:

/da/: $a_{Di}\, v_{Dj}$

where the subscripts i and j index the levels of the auditory and visual modalities, respectively. Analogously, if a_{Bi} represents the degree to which the auditory stimulus A_i has Rising F2-F3 and v_{Bj} represents the degree to which the visual stimulus V_j has Closed lips, the outcome of prototype matching for /ba/ would be:

/ba/: $a_{Bi}\, v_{Bj}$

In general, the support for a given syllable alternative is equal to the auditory support for that alternative times the visual support for that alternative.

The decision operation computes the support for one alternative relative to the sum of the support for all possible response alternatives. This is called a relative goodness rule (RGR) by Massaro and Friedman (1990). With only a single source of information, such as the auditory one A_i, the probability of a /da/ response, $P(/da/)$, is predicted to be the auditory support for /da/ divided by the sum of the auditory support for all alternatives:

$$P(/da/ \mid A_i) = \frac{a_{Di}}{\sum_k a_{ki}} \qquad (1)$$

In Equation 1, the numerator is the support for the alternative /da/ and the denominator is the sum of the merits of all k response alternatives.

Given two sources of information A_i and V_j, P(/da/) is predicted to be the auditory support for /da/ times the visual support for /da/, divided by the sum of the combined auditory/visual support for all alternatives:

$$P(/\text{da}/ \mid A_i \text{ and } V_j) = \frac{a_{Di} \times v_{Dj}}{\sum_k a_{ki} \times v_{kj}} \qquad (2)$$

In Equation 2, the numerator is the bimodal support for the alternative /da/ and the denominator is the sum of the bimodal merits of all k response alternatives.

In general, the probability of response r, P(r), is predicted to be:

$$P(r \mid A_i \text{ and } V_j) = \frac{a_{ri} \times v_{rj}}{\sum_k a_{ki} \times v_{kj}} \qquad (3)$$

One important assumption of the FLMP is that the auditory source supports each alternative to some degree and analogously for the visual source. Each alternative is defined by ideal values of the auditory and visual information. Each level of a source supports each alternative to a differing degree represented by feature values. Because we cannot predict the degree to which a particular auditory or visible syllable supports a response alternative, a free parameter is necessary for each unique syllable and each unique response. An auditory parameter is forced to remain invariant across variation in the different visual conditions, and analogously for a visual parameter.

In our previous research, it has been important to distinguish between information and information processing. *Information* refers to just the output of the evaluation operation in the FLMP (see Fig. 15.2). *Information processing* refers to how this information is processed. That is, information processing corresponds to the nature of the evaluation, integration, and decision operations. Our latest work primarily addresses differences in information processing across aging. We did not attempt to control for either hearing impairment or visual impairment across the two age groups for several reasons. First, it is well known that hearing impairment does not predict everyday communication difficulties (Working Group on Speech Understanding and Aging, 1988). Thus, controlling for hearing impairment would not have necessarily equated the two groups in communication difficulty. Second, we accept the fact that aging will be correlated with decreases in information because of hearing loss (presbycusis) and visual loss. Our goal is to determine whether aging is correlated with changes in information processing—the mental operations in integrating the auditory and visual speech and making a categorization decision. Our experimental

task enables us to make this assessment without having an exact contr over the amount of information in the two groups. The expanded factori design presents unimodal stimuli that provide a direct index of th information available from the auditory and visual speech, respectively. Th FLMP and the other models of speech perception being tested have fre parameters that index the amount of available information. The experimer provides a test of the nature of the information processing, without confounding of potential differences in information.

Although perceivers of different ages might process bimodal speech in th same manner, a given level of auditory or visual information will nc necessarily have equivalent effects across different ages. In fact, given th sensory changes across aging, it is unlikely that a given speech stimulus wi be identified equivalently by two different subjects or by two subjects c different ages. For example, some hearing loss occurs with aging (Corsc 1959, 1963; van Rooij, Plomp, & Orlebeke, 1989; Working Group on Speec Understanding and Aging, 1988). There is also some evidence that the hearin loss occurs earlier for men than for women (Corso, 1963). There is also som evidence for aging differences in the use of visual information. Van Rooij e al. (1989) found that elderly adults were slower than young adults in simpl and choice reaction time tasks. This result may be of particular interest give the fact that the visual evoked response is positively correlated with lipread ing skill (Samar & Sims, 1983; Shepard, 1982). Farrimond (1959) founc changes in lipreading skill across aging, with a decline after middle age. The hypothesis of no differences in information processing only predicts that the two sources of information will be processed in the same manner acros different ages. The alternative hypothesis predicts that the FLMP will fail anc will not give a good description of the results for different age groups. Fo example, because of experience with a hearing loss, senior citizens might be overly influenced by visible speech in bimodal speech perception. If this hypothesis is correct, then the FLMP should fail. Both young adults and senio citizens should be as good in identifying visible speech when presented alone, but senior citizens should be more influenced by this information in the identification of bimodal speech.

AUDITORY DOMINANCE MODEL

A second potential explanation is that an effect of visible speech occurs only when the auditory speech is not completely intelligible (Sekiyama & Tohkura, 1991). Sekiyama and Tohkura tested four labial and six nonlabial consonants in the context /a/, under auditory and auditory–visual conditions. The auditory speech was presented either in quiet or in noise. As expected, identification of the auditory speech was very good in quiet and poor in

noise. The influence of visible speech in the bimodal condition depended on the quality of the auditory speech. There was very little influence with good-quality auditory stimuli and substantial visible influence with poor-quality auditory speech. In many cases, visible speech had an influence for only those auditory stimuli that were not perfectly identified in the auditory condition. However, there were exceptions to this general trend. The auditory syllable /ma/ was perfectly identified in the auditory condition, but was identified as nonlabial about 6% of the time when it was paired with a nonlabial visible articulation. The hypothesis that auditory intelligibility determines whether or not visible speech will have an effect is difficult to test, primarily because intelligibility is not easily defined. Furthermore, the hypothesis rests, at least implicitly, on the assumption that intelligibility is an all-or-none property rather than a continuous property. However, any reasonable definition of intelligibility requires some continuous measure. Intelligibility cannot be equated with 100% correct identification, because 100% in one condition might not give 100% in another.

Even given these limitations in the measure of intelligibility, we can formulate a testable model, called the *auditory dominance model* (ADM), that can predict the influence of visible speech solely as a function of whether or not the auditory speech is identified correctly. This model is related to a single-channel model (Thompson & Massaro, 1989) in which only a single source of information is used on each trial.

On auditory alone trials, the predicted probability of a /da/ response is predicted to be the likelihood of identifying the auditory stimulus as /da/ plus those trials on which the auditory source is not identified as any given alternative, times the bias of identifying the auditory source as /da/ when the auditory source has not been identified.

$$P(/\text{da}/|A_i) = a_{Di} + N \times w_D \tag{4}$$

In Equation 4, a_{Di} is the probability of identifying the ith level of the auditory source as /da/, N is the probability of not identifying the auditory source as any specific alternative, and w_D is the bias of identifying the auditory source as /da/ when the auditory source has not been identified.

For visual alone trials, the visual speech is identified as /da/ or some other alternative. Therefore, the predicted probability of a /da/ response on visual alone trials is equal to

$$P(/\text{da}/|V_j) = v_{Dj} \tag{5}$$

where v_{Dj} is the probability of identifying the jth level of the visual source as response /da/.

On bimodal trials, the auditory speech is identified or it is not. When the subject identifies the auditory stimulus as /da/, she responds with that alternative. In the case that no identification is made the subject responds according to the visual information as described earlier. Therefore, the predicted probability of a /da/ response on bimodal trials is equal to

$$P(/\text{da}/|A \text{ and } V) = a_{Di} + N v_{Dj} \tag{6}$$

Equation 4 states that either the auditory stimulus is identified or else the subject bases his or her decision on the visual information. In Equation 6, a_{Di} is the probability of identifying the ith level of the auditory source as /da/, N is the probability of not identifying the auditory source as any specific alternative, and v_{Dj} is the probability of identifying the visual source as /da/.

The predictions for the other response alternatives are exactly analogous to those given in Equation 4–6.

CATEGORICAL MODEL OF PERCEPTION

The central assumption of the categorical model of perception (CMP) is that only categorical information is available from the auditory source and from the visual source. The $i$the level of the auditory source elicits a /da/ categorization with the probability a_{Di}, and so on for all possible categorizations of the visual source. Similarly, the jth level of the visual source elicits a /da/ categorization with probability v_{Dj}, and so on for all possible categorizations of the visual source.

For unimodal trials, the probability of response /da/ is simply equal to a_{Di} for the auditory trials and v_{Dj} for the visual trials. For the bimodal trials, the identification judgment is based on the separate categorizations of the auditory and visual sources. If the two categorizations to a given speech event agree, the identification response can follow either source. When the two categorizations disagree, it is assumed that the subject will respond with the categorization to the auditory source on some proportion p of the trials, and with the categorization to the visual source on the remainder $(1-p)$ of the trials. The weight p reflects the bias to follow the auditory source rather than the visual.

Given these assumptions, the probability of a /da/ identification response, P(/da/), given a particular bimodal speech event is predicted to be

$$P(/\text{da}/ \mid A_i \text{ and } V_j) = (p)(a_{Di}) + (1-p)(1-v_{Dj}) \tag{7}$$

where i and j index the levels of the auditory and visual modalities, respectively. The a_{Di} value represents the probability of a /da/ categorization

given the auditory level i, and v_{Dj} is the probability of a /da/ categorization given the visual level j. The predictions for the other alternatives are exactly analogous to those given in Equation 7 (see Massaro, 1987, chap. 5).

TESTS OF THE MODELS ACROSS THE LIFE SPAN

There is now a good body of evidence supporting the FLMP over the ADM and CMP (Massaro, 1987, 1992). The bulk of this research has been carried out with young adults of college age. However, any theory of language processing must eventually consider the acquisition and maintenance of the processes involved in this skill. It is surprising how little research on speech perception across development has been carried out relative to the large number of studies of infants and young adults. The focus of this chapter is to evaluate these models across the life span. To illustrate the value of the falsification and strong inference framework in the study of developmental changes, consider a subset of an experimental study reported in Massaro (1987). An expanded factorial design was used: Preschool and college subjects identified auditory, visual, and bimodal speech. The three tasks taken in combination assessed the role of auditory and visual information in speech perception. One question addressed how bimodal speech perception results from some combination of the auditory and visual sources of information.

Three experimental conditions were tested between blocks of trials: bimodal, visual, and auditory. During each trial of the bimodal condition, one of the five auditory stimuli on the continuum from /ba/ to /da/ was paired with one of the two visual stimuli, a /ba/ or a /da/ articulation. Trials in the visual condition used the same videotape but without the speech sounds. In the auditory condition, the TV screen was covered so that only the auditory information was presented. In the bimodal condition, subjects were instructed to watch and to listen to the "man on the TV" and to tell the experimenter whether the man said /ba/ or /da/. Before the visual condition, each child watched the experimenter's mouth as she demonstrated silent articulations of the two alternatives. In this condition, children were instructed to report whether the speaker's mouth made /ba/ or /da/. The college students were told simply to lip-read. In the auditory condition, the subjects were instructed to listen to each test sound to indicate whether the man said /ba/ or /da/. Because the task required a two-alternative forced-choice judgment, a single dependent variable, the proportion of /da/ responses, provides all the information about choice performance. Figure 15.4 shows the results from this experiment. For both groups, the auditory and visual variables produced significant differences, as did the interaction between these two variables in the bimodal condition.

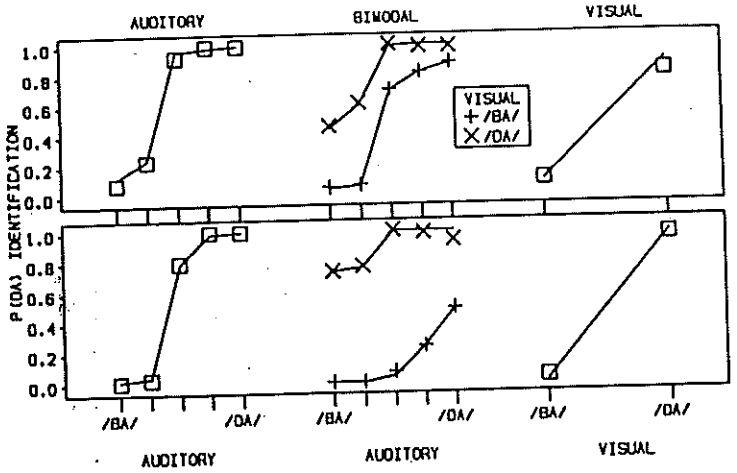

FIG. 15.4. Observed (points) and predicted (lines) proportion of /da/ identifications for auditory (left panel), bimodal (center panel), and visual (right panel) trials as a function of the auditory and visual levels of the speech event for experienced preschool children (top panel) and college students (bottom panel). The predictions are for the FLMP.

The results were tested against the FLMP and CMP. The FLMP gave a much better description of performance than did the CMP for both groups of subjects. This advantage of the FLMP held both for the results of individual subjects and for the results of the average subject. The fit of the FLMP to the average subject was about three or four times better than was the description given by the CMP. The much better description given by the FLMP is especially impressive given that this model required one fewer free parameter than did the CMP. The good fit of the FLMP suggests that preschool children integrate independent and continuous sources of information in the same manner as adults. Using the FLMP, we ask whether the information values for the auditory and visual sources change with age and whether the processes involved in the perceptual recognition of speech differ with age.

A reasonable measure of information value is the degree to which a subject discriminated the different levels of the speech dimensions. An index of discrimination can be determined by taking the difference in response probabilities given two different levels of a speech dimension. In this case, visual discrimination is given by the degree of /da/-ness of a visual /da/ minus the degree of /da/-ness of a visual /ba/. Visual discrimination for the preschoolers was .885 minus .087 or .792. The auditory discrimination is given by the degree of /da/-ness of the most /da/-like auditory syllable minus the degree of /da/-ness of the most /ba/-like auditory syllable (that is, the two endpoints along the auditory continuum). This auditory discrimination value was .992 minus .114 or .878 for the preschoolers.

Corresponding values for the college students were .962 for visual discrimination and .975 for auditory discrimination. Thus, the significantly larger discrimination values for the college students indicate that they had significantly more information about both auditory and visual speech than did the preschool children. This result is consistent with McGurk and MacDonald's (1976) original finding of a larger influence of visible speech on adults than on young children (age 3–8).

In summary, these results, and the good description by the FLMP of both groups, suggest that the developmental differences are due only to information differences. The better fit of the FLMP than the CMP for both age groups argues against a developmental change from one type of process to another in perceptual recognition of speech. At every age, performance is more appropriately described as following the operations of the FLMP, which accordingly provides a framework for assessing life-span differences in information value. The processing of the information appears to follow the operations of the FLMP for both groups.

This developmental study of speech perception by ear and eye revealed some remarkable similarities and differences. With respect to the binary oppositions described in Fig. 15.1, the outcomes of the tests did not change across development. The fundamental processes involved in pattern recognition appear to be the same for preschool children and adults. Even though the processes underlying speech perception were the same, there were significant differences in performance. A distinction between information and information processing is central to understanding these differences. There were significant differences in the informational value of audible and visible speech as a function of age. These differences are easily seen in the unimodal conditions of Fig. 15.4. The adults gave better discrimination of the /ba/ and /da/ alternatives than did the preschool children. These changes are readily explained in terms of increasing experience with development. Preschool children are still acquiring speech-perception skills. They do not lip-read as well as adults and are less able to discriminate changes along an auditory speech continuum. The 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.

One question for the present chapter is whether bimodal speech is processed in the same manner across the life span. The study of bimodal speech perception has been primarily limited to the study of young adults (Massaro & Cohen, 1990; Summerfield, 1979, 1983). Bimodal speech perception offers a valuable domain for the study of aging differences and

similarities. It is important to know to what extent the results to date are dependent on the subject population being used. In addition, aging differences offer a powerful paradigm for broadening the domain for inquiry (Massaro, 1992). Our empirical findings, theories, and models tend to be limited to highly specific situations. Developmental and aging studies allow us to assess the degree to which we can generalize our conclusions.

It might be expected that a loss of hearing with aging might lead to an enhanced ability to perceive the lip and facial movements of the speaker, called *speechreading*. It has been found that visual information from the speaker's face contributes more to speech perception with decreases in the auditory signal-to-noise ratio (Sumby & Pollack, 1954). If perceivers somehow attend more to the visual information in situations with less auditory information, then the aging person with some hearing impairment might develop increased speechreading ability. However, Farrimond (1959) found a decrease of about 8% per decade for the speechreading of men after age 30–39. Similarly, Shoop and Binnie (1979) found a decline of the visual perception of speech across the adult life span. Finally, Ewertsen and Nielsen (1971) found a decline from 20 to 50 to 70 years of age in auditory, visual, and auditory–visual speech perception. It appears that there is a slight decline in speechreading ability with increases in age.

To assess these contrasting models across the life span, we tested two populations of subjects. In the first group, 13 subjects in the age range of 53–81 with a median age of 69 participated. The other population of subjects consisted of students with an age range of 17–46 and a median age of 19.

As in our previous work, the stimuli consisted of synthetic auditory and visible speech. Using an auditory speech synthesizer, we created a continuum of sounds that varied between a good /ba/ and a good /da/. The first sound was a good /ba/. The last sound was a good /da/. The middle sound was halfway between /ba/ and /da/. The second sound was somewhat more /ba/-like and the fourth sound was somewhat more /da/-like. In an exactly analogous manner using computer animation, we synthesized a face saying /ba/ and /da/ and also saying three syllables intermediate between them. Thus, a five-step continuum going from /ba/ to /da/ was created. An expanded factorial design was used, as illustrated in Fig. 15.3. There were 5 auditory and 5 visual syllables, and 25 auditory–visual syllables created by crossing the two continua.

In order to create the synthetic auditory speech, tokens of the first author's /ba/ and /da/ were analyzed using linear prediction to derive a set of parameters for driving a software formant serial resonator speech synthesizer (Klatt, 1980). By altering the parametric information specifying the first 80 msec of the consonant-vowel syllable, a set of five 400-msec syllables covering the range from /ba/ to /da/ was created. The center and lower panels of Fig. 15.5 show how some of the acoustic synthesis parameters

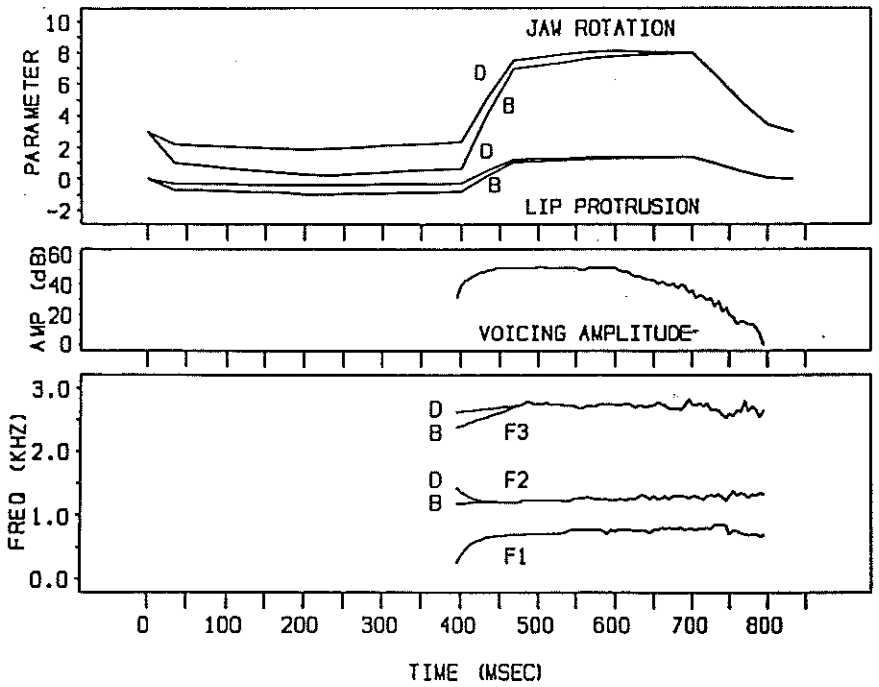

FIG. 15.5. Visual and auditory parameter values over time for visual /ba/ and /da/ stimuli and auditory /ba/ and /da/ stimuli. Bottom panel shows formants F1, F2, and F3, middle panel shows voicing amplitude, and top panel shows jaw rotation and lip protrusion. See text for details.

changed over time for the most /ba/-like and /da/-like of the 5 auditory syllables. During the first 80 msec, the first formant (F1) went from 250 to 700 Hz following a negatively accelerated path. The F2 followed a negatively accelerated path to 1199 Hz, beginning with one of five values equally spaced between 1187 and 1437 Hz from most /ba/-like to most /da/-like, respectively. The F3 followed a linear transition to 2729 Hz from one of five values equally spaced between 2387 and 2637 Hz. All other stimulus characteristics were identical for the 5 auditory syllables.

To create the synthetic visible speech, we used a parametrically controlled polygon topology to generate a fairly realistic animation facial display (Cohen & Massaro, 1990; Parke, 1974). The animation display was created by modeling the facial surface as a polyhedral object composed of about 900 small surfaces arranged in three dimensions, joined together at the edges (Parke, 1974, 1975, 1982). The left panel of Fig. 15.6 shows a framework rendering of this model. To achieve a natural appearance, the surface was smooth shaded using Gouraud's (1971) method (shown in the right panel of Fig. 15.6). The face was animated by altering the location of various

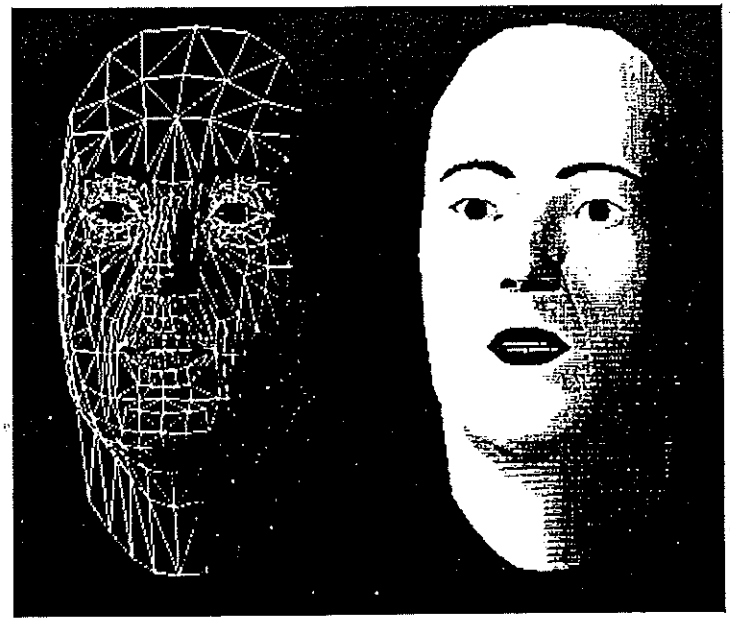

FIG. 15.6. Framework (left) and Gouraud shaded (right) renderings of polygon facial model.

points in the grid under the control of 50 parameters, 11 of which were used for speech animation. Control parameters used for several demonstration sentences were selected and refined by the investigator by studying his own articulation frame by frame and estimating the control parameter values (Parke, 1974). Recently, this software and facial topology has been translated from the original JOVIAL language to C and given new speech- and expression-control routines (Pearce, Wyvill, Wyvill, & Hill, 1986). In this system, a user can type a string of phonemes, which are then converted to control parameters, which are changed over time to produce the desired animation sequence. Each phoneme is defined in a table according to target values for segment duration, segment type (stop, vowel, liquid, etc.), and 11 control parameters. The parameters that are used are jaw rotation, mouth x scale, mouth z offset, lip corner x width, mouth corner z offset, mouth corner x offset, mouth corner y offset, lower lip "f" tuck, upper lip raise, and x and z teeth offset.

The revised software of Pearce et al. (1986) was implemented by us on a Silicon Graphics Inc. IRIS 3030 computer. We adapted the software to allow new intermediate test phonemes and wrote several output processors (pipes) for rendering the ploygonal image information in different ways. One pipe produces wireframe images, a second produces Gouraud shaded images with a diffuse illumination model, a third also includes specular illumination (white

highlights), and a fourth pipe uses tesselation (recursive polygon subdivision) for improved skin texture appearance as well as randomly determined hair. The diffuse pipe used in the present experiment now takes about 1 min to render and record each frame, whereas the diffuse plus specular rendering takes 3 min. To create an animation sequence, each frame was recorded using a broadcast quality BETACAM video recorder under control of the IRIS.

Figure 15.7 gives pictures of the facial model at the time of maximum stop closure for each of the five levels between /ba/ and /da/. The top panel of Fig. 15.5 shows how the visual synthesis parameters changed over time for the first (/ba/) and last (/da/) visual levels. For clarity, only two of the visual parameters are shown—jaw rotation (larger parameter means more open), and lip protrusion (smaller number means more protrusion). Not shown in the figure, the face with the default parameter values was recorded for 2,000 msec preceding and 2,000 msec following the time shown for a total visual stimulus of 4,866 msec. A dark screen of the same total duration was presented for the auditory alone trials.

Following the synthesis a BETACAM tape was dubbed to ¾-in. U-MATIC tape for editing. Only the final 4,766 msec of each video sequence was used for each trial. A tone marker was dubbed onto the audio channel of the tape at the start of each syllable to allow the playing of the 400-msec auditory speech stimulus just following consonant release of the visual stimulus. The marker tone on the video tape was sensed by a Schmidt trigger on a PDP-11/34A computer, which presented the auditory stimuli from digitized representations on the computer's disk. Figure 15.5 shows the temporal relationship between the auditory and visual parts of the stimulus. As can be seen in the figure, the parameter transitions specifying the consonantal release occurred at about the same time for both modalities.

In this experiment, synthetic auditory and visual speech were manipulated in an expanded factorial design previously illustrated in Fig. 15.3. The onsets of the second and third formants were varied to give an auditory continuum between the syllables /ba/ and /da/. In analogous fashion, we systematically varied parameters of the facial model 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. These trials were recorded on videotape for use in the experiments. Six random sequences were determined by sampling the 35 conditions without replacement.

Subjects were instructed to listen and to watch the speaker, and to identify the syllable as /ba/, /da/, /bda/, /dba/, /ðа/, /va/, /ga/, or "other." These response alternatives were determined from pilot studies in which the

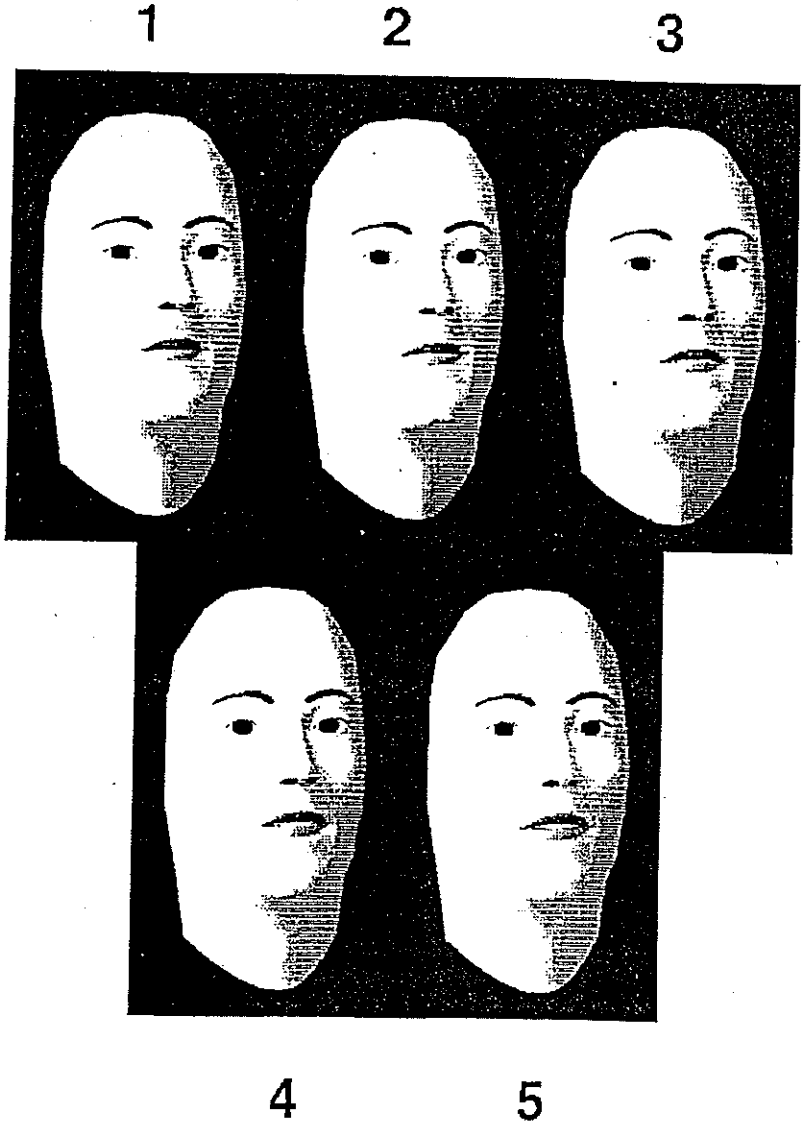

FIG. 15.7. The facial model at the time of maximum stop closure for each of the five levels of visible speech between /ba/ and /da/.

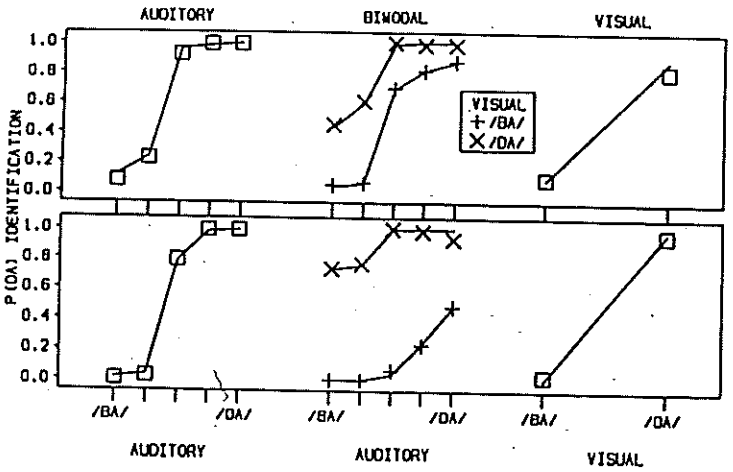

FIG. 15.8. Observed (points) and predicted (lines) proportion of /da/ identifications for auditory (left panel), bimodal (center panel), and visual (right panel) trials as a function of the auditory and visual levels of the speech event for experienced preschool children (top panel) and college students (bottom panel). The predictions are for the FLMP.

responses were not constrained. Each of the 35 possible stimuli were presented a total of 12 times during two sessions and the subject identified each stimulus during a 3-sec response interval. Prior to the experimental stimuli, subjects were given 15 practice trials to familiarize them with the task. Subjects were given a short break, approximately 5 min, after completing the tape of 210 trials. Unknown to the subjects, the tape was rewound and played again, repeating the 210 trials.

The mean observation proportion of identifications was computed for each subject for each of the unimodal and bimodal conditions. Separate analyses of variance were carried out on the auditory, visual, and bimodal conditions. The factors in the design were auditory speech, visible speech, and response. In all cases, there were significant effects (at $p < .001$). Both the auditory and the visual sources of information had a strong impact on the identification judgments. As illustrated in Fig. 15.9 and 15.10, the proportion of responses changed systematically across the visual continuum, both for the unimodal and the bimodal conditions. Similarly, the pattern of responses changed in an orderly fashion across the auditory continuum, for both the unimodal and bimodal conditions. Finally, the auditory and visual effects were not additive, as demonstrated by the significant auditory–visual interaction on response probability in the bimodal condition.

The two groups did not show any difference with respect to identification of the syllables as a function of the auditory continuum alone, the visual continuum alone, or either the auditory or visual levels in the bimodal

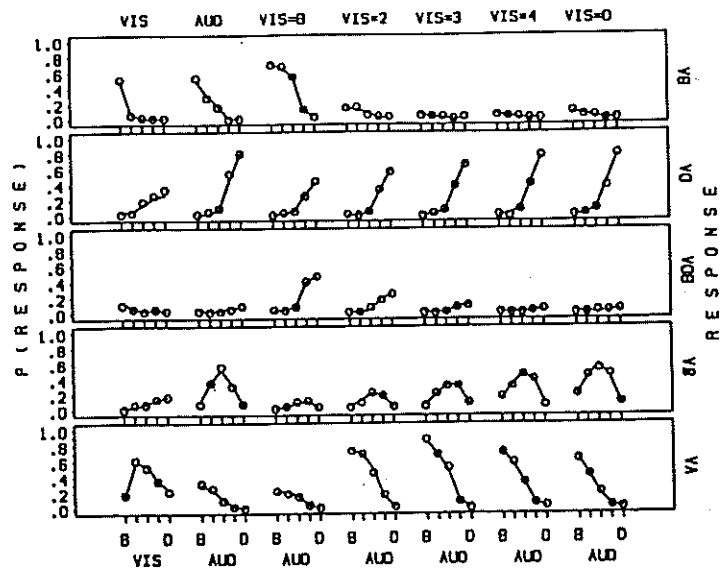

FIG. 15.9. Observed (points) and predicted (lines) proportion of /ba/, /da/, /bda/, /ðа/, and /va/ identifications for the visual alone (left panel), auditory alone (second panel) and bimodal (remaining panels) conditions as a function of the five levels of the synthetic auditory (AUD) and visual (VIS) speech varying between /ba/ (B) and /da/ (D). The observations are from 13 college students. The lines give the predictions for the FLMP.

condition. There was, however, a significant complex interaction of visual and auditory levels by group for the bimodal stimuli. This interac reflected the greater number of /bda/ judgments for the college stuc when visual /ba/ is paired with an auditory /da/ (see Fig. 15.9 and 15

RELATIVE INFLUENCE OF VISIBLE AND AUDIBLE SPEECH

One question of interest is the relative contribution of visible and au speech in the bimodal condition. An index of the magnitude of the e of one modality can be described by the difference in response probab to the two endpoint stimuli from that modality. This difference was comp for each subject for both audible and visible sources of information. F example, an overall .90 probability of /da/ given the /da/ endpoint stir and an overall .2 probability of /da/ given the /ba/ endpoint stimulus w give an effect of .7 Analyses of variance were carried out on these sc The magnitude of the visual effect did not differ across the two groups

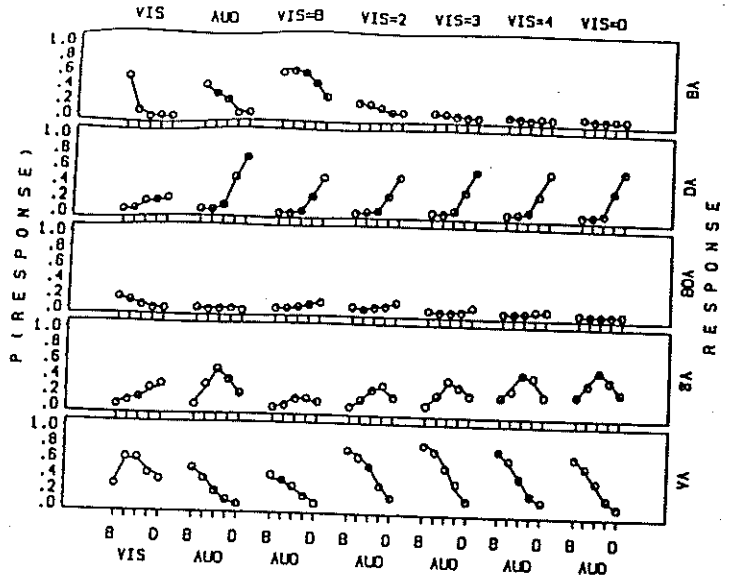

FIG. 15.10. Observed (points) and predicted (lines) proportion of /ba/, /da/, /bda/, /ða/, and /va/ identifications for the visual alone (left panel), auditory alone (second panel) and bimodal (remaining panels) conditions as a function of the five levels of the synthetic auditory (AUD) and visual (VIS) speech varying between /ba/ (B) and /da/ (D). The observations are from 13 senior citizens. The lines give the predictions for the FLMP.

the average effect was .101. The magnitude of the auditory effect was marginally significant across the two groups. The average effect was .144 for the college students and .123 for the senior citizens, consistent with the general finding of a decrease in auditory sensitivity with aging (van Rooij et al., 1989; Working Group on Speech Understanding and Aging, 1988).

TESTS OF THE MODELS

The FLMP, ADM, and CMP were fit to the individual results of each of the 26 subjects in the two groups. The quantitative predictions of the model are determined by using the program STEPIT (Chandler, 1969). A model is represented to the program in terms of a set of prediction equations and a set of unknown parameters. By iteratively adjusting the parameters of the model, the program minimizes the squared deviations between the observed and predicted points. The outcome of the program STEPIT is a set of parameter values that, when put into the model, come closest to predicting the observed results. Thus, STEPIT maximizes the accuracy of the description

of a given model. Of greatest interest here is the goodness-of-fit of a model indexed by the root mean square deviation (RMSD)—the square root of the average squared deviation between the predicted and observed values.

Figures 15.9 and 15.10 show that the FLMP provided a good description of the identifications of both the unimodal and bimodal syllables for both age groups. The average RMSD was .0525 and .0468 for the college students and senior citizens, respectively. These averages were computed from the fits of the 13 individual subjects in each group.

Figures 15.11 and 15.12 give the average observed results and the average predicted results of the ADM. The RMSD was .0958 and .0732 for the young and old adults, respectively. An analysis of variance on the RMSD values showed that the FLMP gave a significantly better description of the results than did the ADM.

Figures 15.13 and 15.14 give the average observed results and the average predicted results of the CMP. As can be seen in the figures, the CMP gave a poor description of the observed results. The RMSD was .1113 for the college students and .0878 for the senior citizens. An analysis of variance on the RMSD values showed that the FLMP gave a significantly better description of the results than did the CMP.

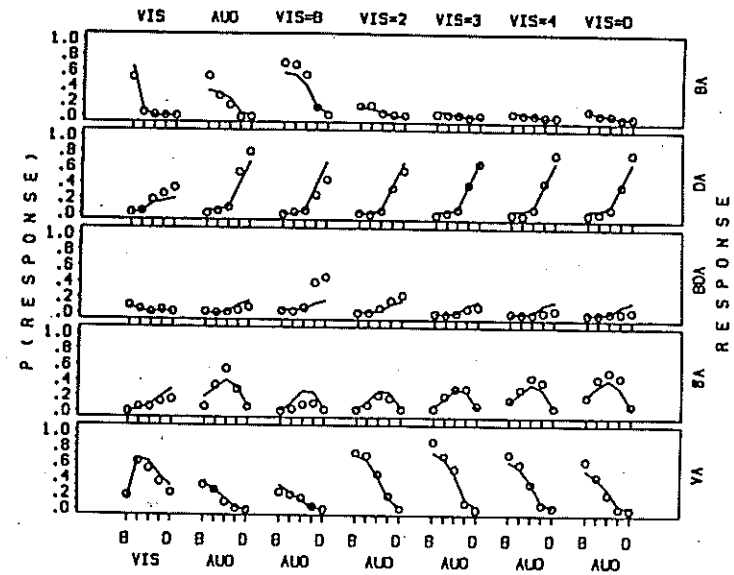

FIG. 15.11. Observed (points) and predicted (lines) proportion of /ba/, /da/, /bda/, /ða/, and /va/ identifications for the visual alone (left panel), auditory alone (second panel) and bimodal (remaining panels) conditions as a function of the five levels of the synthetic auditory (AUD) and visual (VIS) speech varying between /ba/ (B) and /da/ (D). The observations are from 13 college students. The lines give the predictions for the ADM.

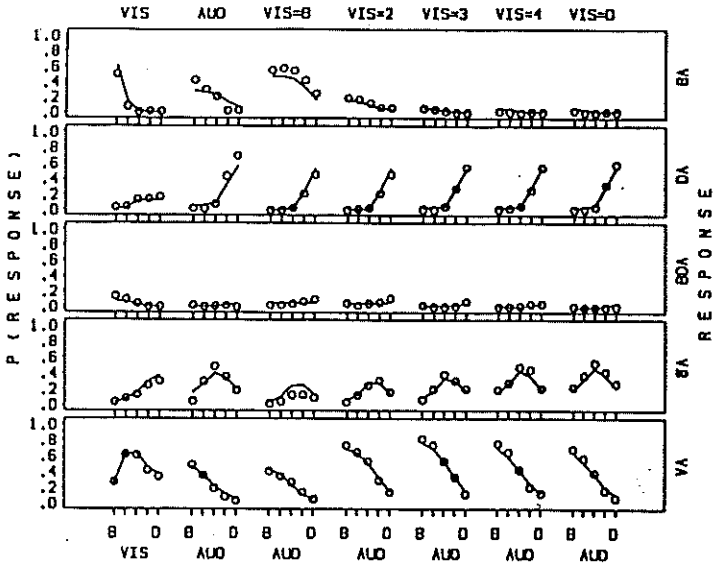

FIG. 15.12. Observed (points) and predicted (lines) proportion of /ba/, /da/, /bda/, /ða/, and /va/ identifications for the visual alone (left panel), auditory alone (second panel) and bimodal (remaining panels) conditions as a function of the five levels of the synthetic auditory (AUD) and visual (VIS) speech varying between /ba/ (B) and /da/ (D). The observations are from 13 senior citizens. The lines give the predictions for the ADM.

The CMP is mathematically identical to weighted adding or a weighted averaging model (Massaro, 1987). Thus, a test of the CMP also allows a test of whether the inputs are added or combined in a nonadditive manner. The good fit of the FLMP relative to the CMP is evidence against additive integration. The integration of the multiple sources appears to result in the least ambiguous sources having the most impact on processing. Given that the FLMP is mathematically equivalent to Bayes' theorem—an optimal algorithm for integrating multiple sources of information—the good fit of the model to the present results is evidence for optimal integration of auditory and visual information in speech perception.

NUMBER OF /bda/ JUDGMENTS

A frequent response is the consonant cluster /bda/ when the stimulus is an auditory /da/ paired with a visible /ba/. This perceptual judgment is reasonable because visible /bda/ is almost identical to visible /ba/, and audible /da/ is very similar to audible /bda/. The alternative /dba/ is not reasonable because of the huge mismatch of visible /dba/ with visible /ba/. The /bda/

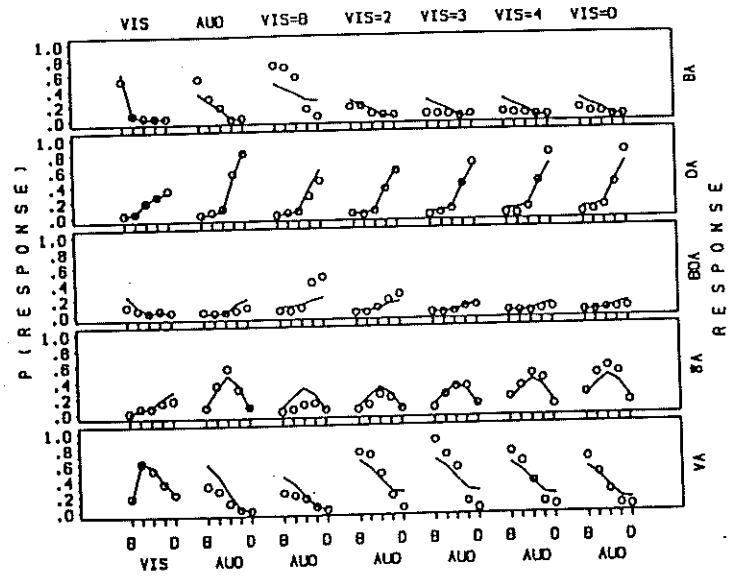

FIG. 15.13. Observed (points) and predicted (lines) proportion of /ba/, /da/, /bda/, /ða/, and /va/ identifications for the visual alone (left panel), auditory alone (second panel) and bimodal (remaining panels) conditions as a function of the five levels of the synthetic auditory (AUD) and visual (VIS) speech varying between /ba/ (B) and /da/ (D). The observations are from 13 college students. The lines give the predictions for the CMP.

cluster is also a reasonable response alternative in English because this cluster occurs in compound words and across word boundaries. In English, we can have compound words like crabdish and word boundaries separating /b/ and /d/, as in curb dog or tab down or lab dirt, and so on. Similarly, one can have rap time or nap time or crap time. It is of interest whether older adults give as many cluster responses as younger adults.

As can be seen in Fig. 15.9 and 15.10, senior citizens give fewer /bda/ judgments than college students. It turns out that a given subject either did or did not give predominantly /bda/ responses. Only 3 of the 13 old adults gave /bda/ responses, but they gave them about as frequently as the typical young adult. In like fashion, 4 of the 13 college students did not respond /bda/. Thus, in terms of whether /bda/ judgments occur, there were a few college students that resembled senior citizens and a few senior citizens that resembled college students. Most importantly, however, the FLMP gave a good description of the individual subjects independently of whether they tended to give /bda/ responses.

There are probably various tenable explanations of aging effects on the number of /bda/ judgments. Generally, there may be age differences in the

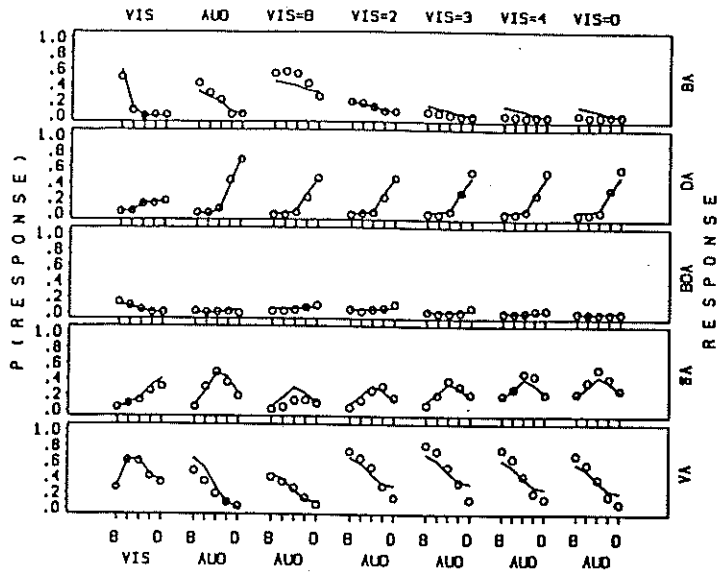

FIG. 15.14. Observed (points) and predicted (lines) proportion of /ba/, /da/, /bda/, /ða/, and /va/ identifications for the visual alone (left panel), auditory alone (second panel) and bimodal (remaining panels) conditions as a function of the five levels of the synthetic auditory (AUD) and visual (VIS) speech varying between /ba/ (B) and /da/ (D). The observations are from 13 senior citizens. The lines give the predictions for the CMP.

ality of the memory trace for the consonant cluster /bda/. On the other hand, ler adults may have simply been less willing to give cluster responses in the k. There is also some literature supporting the distinction between fluid and stallized processing. Older adults tend to be more crystallized than younger ults, and this might account for fewer adults giving /bda/ judgments.

One question of interest is whether the fundamental processes underlying rformance differ as a function of the occurrence of /bda/ responses. One y to test this question is to determine if the accuracy of the FLMP scription differs for the different response protocols. In all cases, the FLMP ve a good description of the results. Although we are not sure why a ung adult relative to an elderly adult is more likely to respond with /bda/ en a visible /ba/ and an auditory /da/, the fundamental processes seem work equivalently across these different response patterns.

There is a body of research supporting the idea of a mental slowing for : elderly. We take this slowing as representative of a loss in information rived from the senses and memory. In our view, mental slowing does not lect differences in information processing—by which we mean differences the mental processes supporting speech perception.

SUMMARY

We have been concerned with the development and maintenance of intersensory speech perception. Understanding spoken language is a form of pattern recognition involving the evaluation and integration of multiple sources of information. Although the perception of spoken language is usually thought of as unimodal, we observed good evidence for its multimodal nature. Our experimental studies using expanded factorial designs and tests of quantitative models have shown that the perception of face-to-face spoken language is truly intersensory—it is influenced by both sight and sound.

A new experiment was successful in comparing bimodal speech perception across aging. The results indicated that the processes engaged by bimodal speech are fundamentally equivalent across the life span. The only observed difference indicated that a smaller percentage of older adults gave /bda/ judgments relative to the percentage of younger adults. Although one could conceive of a variety of explanations for this finding, we give little significance to the finding itself. We distinguish between information and information processing. Information refers to just the output of the evaluation operation in the FLMP (see Fig. 15.2). Information processing refers to how this information is processed. That is, information processing corresponds to the nature of the evaluation, integration, and decision operations. The tests among the three quantitative models address the issue of differences in information processing across aging. The outcome of these tests revealed that the FLMP gave a significantly better description of performance for both young and older adults. Thus, the process assumed by the FLMP account for both groups and we can conclude that the processing of bimodal speech appears to be constant across aging.

ACKNOWLEDGMENTS

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), the graduate division of the University of California, Santa Cruz. The author thanks Michael M. Cohen for help in all aspects and the research and Antoinette Gesi for testing subjects.

REFERENCES

Anderson, N. H. (1981). *Foundations of information integration theory.* New York: Academic Press.

Anderson, N. H. (1982). *Methods of information integration theory.* New York: Academic Press.

Chandler, J. P. (1969). Subroutine STEPIT—Finds local minima of a smooth function of several parameters. *Behavioral Science, 14,* 81–82.

Cohen, M. M., & Massaro, D. W. (1990). Synthesis of visible speech. *Behavioral Research Methods and Instrumentation, 22,* 260–263.

Corso, J. (1959). Age and sex differences in pure-tone thresholds. *Journal of the Acoustical Society of America, 31,* 498–507.

Corso, J. (1963). Age and sex differences in pure-tone thresholds. *Archives in Otolaryngology, 77,* 385–405.

Ewertsen, H. W., & Nielsen, H. B. (1971). A comparative analysis of the audiovisual, auditive and visual perception of speech. *Acta Otolaryngologica, 72,* 201–205.

Farrimond, T. (1959). Age differences in the ability to use visual cues in auditory communication. *Language and Speech, 2,* 179–192.

Gouraud, H. (1971). Continuous shading of curved surfaces. *IEEE Transactions on Computers,* C-20(6), 623–628.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America, 67,* 971–995.

Massaro, D. W. (Ed.). (1975). *Understanding language: An information processing analysis of speech perception, reading, and psycholinguistics.* New York: Academic Press.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Massaro, D. W. (1989a). Multiple book review of *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry. Behavioral and Brain Sciences, 12,* 741–794.

Massaro, D. W. (1989b). Testing between the TRACE model and the Fuzzy Logical Model of Perception. *Cognitive Psychology, 21,* 398–421.

Massaro, D. W. (1990). A fuzzy logical model of speech perception. Proceedings of the XXIV International Congress of Psychology. In D. Vickers & P. L. Smith (Eds.), *Human information processing: Measures, mechanisms, and models* (pp. 367–379). Amsterdam: North Holland.

Massaro, D. W. (1992). Broadening the domain of the fuzzy logical model of perception. In H. L. Pick, Jr., P. Van den Broek, & D. C. Knill (Eds.), *Cognition, conceptual, and methodological issues* (pp. 51–84). Washington, DC: American Psychological Association.

Massaro, D. W., & Cohen, M. M. (1983). Phonological context in speech perception. *Perception & Psychophysics, 34,* 338–348.

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

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

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

O'Connor, N., & Hermelin, B. (1981). Coding strategies of normal and handicapped children. In R. D. Walk & H. L. Pick (Eds.), *Intersensory perception and sensory integration* (pp. 315–343). New York: Plenum.

Parke, F. (1974). *A parametric model for human faces* (Tech. Rep. UTEC-CSc-75-047). Salt Lake City: University of Utah.

Parke, F. (1975). A model for human faces that allows speech synchronized animation. *Computers and Graphics Journal, 1*(1), 1–4.

Parke, F. (1982). Parameterized models for facial animation, *IEEE Computer Graphics, 2*(9), 61–68.

Pearce, A., Wyvill, B., Wyvill, G., & Hill, D. (1986). Speech and expression: A computer solution to face animation. *Graphics Interface '86.*

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

Samar, V. J., & Sims, D. G. (1983). Visual evoked-response correlates of speechreading performance in normal-hearing adults: A replication and factor analytic extension. *Journal of Speech and Hearing Research, 26,* 2–9.

Sekiyama, K., & Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of America, 90,* 1797–1805.

Shepherd, D. C. (1982). Visual-neural correlate of speechreading ability in normal-hearing adults. *Journal of Speech and Hearing Research, 25,* 521–527.

Shoop, C., & Binnie, C. A. (1979). The effect of age upon the visual perception of speech. *Scandinavian Audiology,* 8, 3–8.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26,* 212–215.

Summerfield, A. Q. (1979). Use of visual information in phonetic perception. *Phonetica, 36,* 314–331.

Summerfield, A. Q. (1983). Audio-visual speech perception. In M. E. Lutman & M. P. Haggard (Eds.), *Hearing science and hearing disorders.* London: Academic.

Thompson, L. A., & Massaro, D. W. (1989). Before you see it, you see its parts: Evidence for feature encoding and integration in preschool children and adults. *Cognitive Psychology, 21,* 334–362.

Townsend, J. T., & Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes.* London: Cambridge University Press.

van Rooij, J. C. G. M., Plomp, R., & Orlebeke, J. F. (1989). Auditive and cognitive factors in speech perception by elderly listeners. I: Development of test battery. *Journal of the Acoustical Society of America, 86,* 1294–1309.

Working Group on Speech Understanding and Aging. (1988). Speech understanding and aging. *Journal of the Acoustical Society of America, 83,* 859–893.