

2

Speech Perception by Ear and Eye

Dominic W. Massaro

Program in Experimental Psychology, University of California, Santa Cruz, and Zentrum für Interdisziplinäre Forschung der Universität Bielefeld

The information provided by both audible and visible speech represents another instance of multiple sources of information supporting pattern recognition and interpretation. The present framework, derived from research in a wide variety of other domains, acknowledges the presence of multiple, independent and continuous sources of information in language processing. The research strategy utilizes factorial designs, functional measurement, testing of mathematical models and strong inference in studying hearing by ear and eye. The chapter reviews research addressing a number of fundamental issues confronting any theoretical account of the phenomenon. Experimental and theoretical tests have converged on the following understanding of bimodal speech perception. The audible and visible sources are integrated in the sense that both sources simultaneously contribute to perception. The two sources provide continuous rather than discrete information at integration, and the information provided by one source can be considered to be independent of the other source. The nature of the integration process can be considered to be an enhancing operation, rather than a simple compromising one. Extending these questions to the developmental domain, the research has demonstrated that young children behave identically to adults, even though visible speech is less informative because of less lip-reading skill. These observations provide major constraints for any potential account of the phenomena. A fuzzy logical model of perception developed in other domains provides a good description of exactly these phenomena. Therefore, one can reject the idea that bimodal speech perception is outside the domain of prototypical pattern recognition.

INTRODUCTION

In 1976, McGurk and MacDonald reported that hearing “ba” at the same time as “ga” is seen to be spoken can give rise to an illusory heard “da”.

Further studies (see Summerfield, Chapter 1) have confirmed and extended this finding: both fusions (where the illusory percept, while sharing phonological features of the auditory and visual syllable, has no consonantal phoneme in common with either) and blends (where the percept is of at least one of the presented phonemes in conjunction with another) have been reported. So, for example, "bda" or "bga" are often perceived when a seen "ba" is co-incident with a heard "ga" (see e.g. MacDonald & McGurk, 1978).

When I first heard about the auditory-visual blend illusion, I was informed but not surprised. It was difficult to understand why the phenomenon was being touted as an embarrassment to theories of speech perception. It is true that both theoretical work and empirical studies (e.g. Fodor, Bever & Garrett, 1974; Massaro, 1975b) had not addressed the contribution of lip-read information and that it would now be necessary to do so. On the other hand, it remained to be determined to what extent various frameworks for the study of speech perception had to be revised given this new demonstration of a visual source of information.

The general framework developed in our research seemed ideally suited for accounting for the contribution of lip-read information. We had already begun to account for the integration of the wide variety of acoustic characteristics contributing to segmental differences and the contribution of contextual constraints to language understanding (Massaro, 1975a, 1975b; Massaro & Cohen, 1976; Massaro & Oden, 1980; Oden & Massaro, 1978). The research enterprise assumed the presence of multiple, independent and continuous sources of information in language with the perceiver's task one of evaluation and integration of these sources. Within this framework, lip-read information simply assumes the status of another source of information that should be treated equivalently to the plethora of other bottom-up and top-down sources.

With respect to the auditory-visual blend illusion, it should not be viewed as an illusion at all. Although the sight of the speaker's lips modifies what the perceiver hears, the outcome only reflects the natural process of integrating other sources of information with those given acoustically. The situation is equivalent, in principle, to the contribution of phonological context, lexical status or sentential constraints to perceptual recognition (Ganong, 1980; Isenberg, Walker & Ryder, 1980; Marslen-Wilson & Welsh, 1978; Massaro & Cohen, 1983c; Tyler & Wessels, 1983). Subjects tend to identify an initial voiced stop consonant as "d" when it occurs with the following context *ash* and as "t" with the following context *ask* (Ganong, 1980). The perceptual recognition of the initial consonant is biased in the direction of the lexical status given by the context, in the same way that the perceptual recognition of an auditory stop consonant is biased in the direction of the speaker's lip movements (McGurk & MacDonald, 1976). The only difference is that the

lexical status might be considered to be a top-down source, whereas lip-read information is probably more appropriately described as a bottom-up source.

We were attracted to the study of the contribution of lip-read information in speech perception because it seemed to provide a natural, yet fruitful, domain to study speech perception given multiple sources of information. We had previously objected to traditional studies that manipulated only a single aspect of the speech signal (Massaro, 1979; Massaro & Cohen, 1976). The realization of the functional role of lip-read information made possible experiments in which both the auditory and the visual characteristics of the speech signal could be manipulated simultaneously. This more complicated paradigm makes possible a variety of questions that cannot be addressed in the more common experiment manipulating only a single variable.

The goal of Chapter 2 is to develop and answer questions generated by the finding that lip-read information is functional in speech perception. That is, we are faced with lip-read information (no pun intended) as an additional source of information in speech, and the questions are aimed at providing an adequate understanding of its functional role. These questions are similar to those that would be generated in any domain of pattern recognition, and they are considered to be essential to providing a complete psychological level of description. The answers to the questions impose important constraints on any theories offered as an account of speech perception by ear and by eye.

Our research strategy follows the tenets of falsification and strong inference (Platt, 1964; Popper, 1959) in that binary oppositions are constructed and tested. For each opposition, multiple tests are implemented so that no conclusion rests on just one or two observations. Given the limited scope of the current chapter, however, I will be able to present only a single research finding for each opposition. Each finding serves as an illustration of how the question is answered and is consistent with considerable evidence on the question. The reader is referred to other papers for additional evidence (Cohen, 1984; Massaro, 1984; Massaro, in press; Massaro & Cohen, 1983b; Massaro, Thompson, Barron & Laren, 1986). The belief is that the dissection of this phenomenon within the framework of binary oppositions, combined with the tools of information integration (Anderson, 1981, 1982) and mathematical-model testing, illuminates not only the phenomenon itself but also more general problems of perception and pattern recognition.

The binary oppositions to be considered are arranged hierarchically in Fig. 2.1. In some cases, the question at one level is dependent on the answers to the questions at higher levels. As an example, the issue of whether or not audible and visible sources of information are integrated (combined) in perception requires that both sources rather than just a single source be functional for the perceiver.

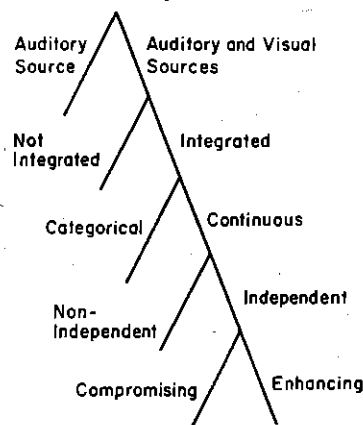


FIG. 2.1 Tree of wisdom illustrating a set of binary oppositions central to the domain of speech perception by eye and ear.

AUDIBLE VERSUS AUDIBLE AND VISIBLE SOURCES

The question of whether visual information, in addition to auditory information, contributes to understanding speech is redundant in the context of this book. However, unlike auditory speech, visual speech distinguishes among only a subset of speech contrasts (Walden, Prosek, Montgomery, Scherr & Jones, 1977). Even so, visual speech appears to be utilized by the hearing perceiver, the auditory-visual blend illusion providing the most direct evidence (McGurk & MacDonald, 1976; MacDonald & McGurk, 1978). What we see clearly influences what we perceive in speech perception. For example, pairing the sound "ba" with a seen "ga", articulation is usually recognized perceptually as "da". As we will see, however, visible speech does not completely determine place of articulation, as originally suggested by the discoverers of the illusion.

INTEGRATION VERSUS NON-INTEGRATION

The outcome of our first contrast indicates that both auditory and visual sources of information are utilized in speech perception. The answer to the next question might seem obvious since it seems only natural that two functional sources should be integrated. Integration of two sources of information refers to some process of combining or utilizing both sources to make a perceptual judgement. However, demonstrating that two sources are

integrated in perceptual recognition is no easy matter (Massaro, in press). A perceiver might utilize the auditory source on some trials and the visual source on others, giving the overall impression of integration.

It might not be possible to demonstrate integration if subjects are tested only with a factorial combination of the two sources. By including judgements of the single modalities, however, the question of integration might be tested. Consider the perception of bimodal speech events created by the combination of synthetic speech sounds along an auditory "ba"-to-"da" continuum paired with "ba" or "da" visual articulations. By adding the single auditory and single visual cue conditions to the factorial design as illustrated in Fig. 2.2, it is at least logically possible to reject the possibility of a subject using only one source on each trial. What is necessary is to find judgements of certain bimodal speech events that cannot be accounted for by judgements of the visual or auditory dimensions presented alone.

In our experiments (Massaro, in press), subjects are usually asked to identify bimodal speech events, auditory alone, and visual alone trials as illustrated in Fig. 2.2. For the bimodal trials, an auditory synthetic syllable along a nine-step "ba"-to-"da" continuum is dubbed onto a videotape of the speaker saying "ba" or "da". In addition, the auditory speech stimuli are presented alone with no lip movements on some trials, and the "ba" and "da" articulations are presented without sound on other trials. The subjects are permitted an open-ended set of response alternatives.

For the question of integration, we limit our analysis to the occurrence of "bda" judgements shown in Fig. 2.3. The pattern of occurrences provides strong evidence for a true integration of the auditory and visual sources. The critical finding is the large proportion of "bda" judgements given a visual "ba" and an auditory "da" when this same judgement is seldom given to either the visual or the auditory modalities presented alone. We find over five times as many "bda" judgements given to the bimodal events relative to the visual-only condition, and the auditory-only condition almost never pro-

	AUDITORY				
	BA			DA	NONE
	BA				
	DA				
VISUAL	NONE				

FIG. 2.2 Expansion of a typical factorial design to include auditory-alone and visual-alone conditions. The five levels along the auditory continuum represent speech sounds varying in equal steps between "ba" and "da".

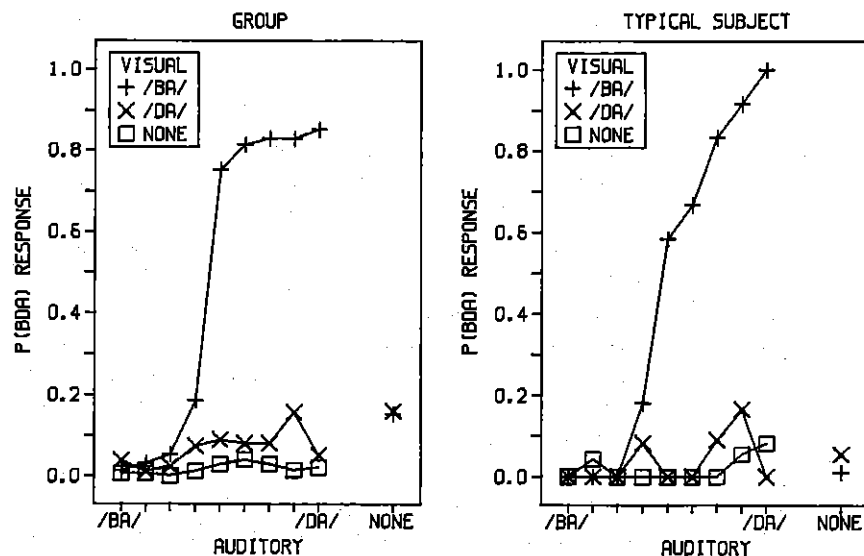


FIG. 2.3 Probability of "bda" judgements to bimodal, auditory and visual speech events. The left panel gives the group results, and the right panel gives the results for a typical subject. The nine levels along the auditory continuum represent speech sounds varying in equal steps between "ba" and "da".

duces "bda" judgements. It follows that the "bda" judgements observed on bimodal trials could not have resulted from identification of just one of the two sources on a trial. The result represents the outcome of the integration of both auditory and visual sources of information, in that both contributed to a single perceptual report.

Integration seems to be an efficient system for perceiving speech. Given multiple sources of information that are susceptible to random variability, the optimal strategy would be to evaluate and integrate all of the sources even though they might be ambiguous. One cost to a system designed in this way would be the relative inability to process selectively a single dimension of the speech event. Thus, subjects find it difficult to attend to the auditory speech while looking at the speaker's lips, and vice versa (Massaro, in press). That is, we find it difficult to process selectively one dimension of the speech event independently of other dimensions.

CATEGORICAL VERSUS CONTINUOUS INFORMATION

The question of the categorical or continuous nature of speech will be the most familiar to students of speech perception. Categorical information implies that only discrete (phonetic) information is available about the speech event. Continuous information implies that useful information need

not be discrete. Although these two hypotheses might seem to be easily distinguished, in reality they are not. Both the categorical and the continuous hypothesis can predict a continuous change in identification responses or rating responses with continuous changes along a speech continuum (Massaro & Cohen, 1983a). A discriminating test between the hypotheses requires an analysis of the distribution of rating responses to repeated presentations of a speech event.

Consider the bimodal speech events illustrated in Fig. 2.2 if the task of the subject is to rate each event along a nine point "ba"-to-"da" continuum. Categorical information predicts that the ratings to repeated presentations of a single event will come from two kinds of trials: those trials on which the event was identified as one alternative, "ba", and those on which the event was identified as the other alternative, "da". Thus, categorical perception predicts that the distribution of ratings to a given stimulus is a result of two different phonetic categorizations or a mixture of "ba" identification and "da" identification trials. On the other hand, continuous perception predicts that the rating is based on continuous information by which the perceiver can reliably rate the degree to which the syllable is "ba"-like or "da"-like. Hence, the distribution of ratings to a given speech event will result from a single kind of trial on which the perceiver has continuous information about the speech event.

As noted by Massaro and Cohen (1983a), analysing the distribution of ratings can test between categorical and continuous models of speech perception. Figure 2.4 gives the distribution of ratings for a typical subject in

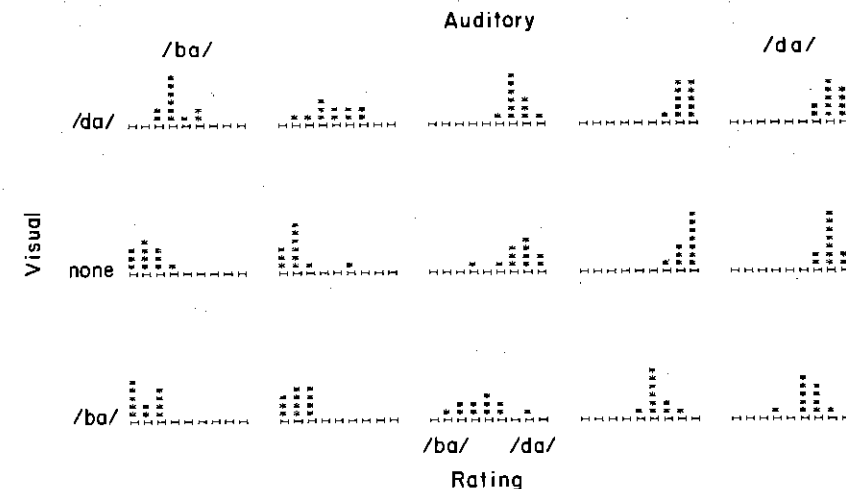


FIG. 2.4 Frequency distributions of ratings to bimodal speech events for a typical subject. The categorical model predicts that repeated ratings to a given speech event result from two distributions, whereas the continuous model predicts that the ratings result from a single distribution.

an experiment in which subjects were required to rate the "ba"-ness to "da"-ness along a nine-point scale. As can be seen in the figure, it is very difficult to see how these ratings could have resulted from a mixture of two different distributions. The categorical and continuous models were quantified to predict the distribution of ratings under the various experimental conditions. For each subject, the continuous model gave a much better description of the results than did the categorical model. That is, the evidence from the distribution of ratings supports the use of continuous rather than categorical information in bimodal speech perception.

INDEPENDENT VERSUS NON-INDEPENDENT EVALUATION OF SOURCES

The next branch of the binary-opposition tree involves the issue of whether the two sources of information are non-independent or independent. Independent sources of information imply that the information value determined in the evaluation of one source remains independent of the information value of the other. Non-independent sources implies a violation of this principle. Cohen (1984) used reaction times to completely unambiguous single-dimension and bimodal events to determine whether the two dimensions show some form of non-independence. If they do, then it should not be possible to account for the reaction times to a bimodal "ba" in terms of simply the reaction times to a visual "ba" and to an auditory "ba". If the two dimensions are independent, we might expect reaction times to the bimodal event to be somewhat faster than those to the single-dimension events, but the advantage should be completely accounted for by statistical facilitation (Gielen, Schmidt & Van Den Heuvel, 1983; Raab, 1962).

The reader might have realized that the predictions of independence appear to contradict those of integration. That is, the independence prediction implies that subjects respond to the auditory or visual source that is processed first, without waiting to integrate the two sources. Although this implication is correct, it does not necessarily mean that the two sources are not integrated, only that a response can be initiated before integration occurs. In the present task, the auditory and visual sources are completely unambiguous and they always agree with one another in the bimodal condition. Subjects are also instructed to respond as quickly as possible. Thus, although integration may still have occurred, it might not be observed in the reaction times since subjects could initiate a response based on just a single dimension, whether or not integration of the two dimensions was complete.

Figure 2.5 shows the distribution of reaction times of two subjects to the single modality and bimodal conditions. As can be seen in the figure, the

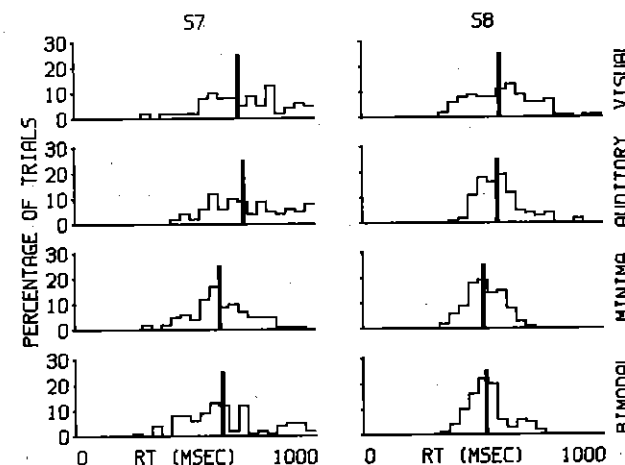


FIG. 2.5 Distribution of reaction times for two subjects to visual-alone, auditory-alone, and bimodal speech events. The minima distribution is that predicted for bimodal trials based on independent processing of the auditory-alone and visual-alone trials. The solid bar gives the mean reaction time.

subject is somewhat faster to the bimodal speech event but no faster than expected if the subject simply begins to initiate a response when either the auditory or visual dimension is identified. The advantage of bimodal trials can be accounted for simply in terms of the variability of the processing times along each dimension, allowing the average reaction time to bimodal events to be shorter than to either dimension presented alone. These results support independence of auditory and visual information without necessarily contradicting the earlier conclusion that the two sources of information are integrated for perceptual recognition.

COMPRISING OR ENHANCING INTEGRATION

The final branch to be discussed in this chapter involves the combination rule used to integrate the two sources. The exact integration rule that is used is not easy to determine since the question is much more involved than those that we have discussed previously. For the present opposition, we distinguish between two general possibilities of the nature of integration. Both possibilities are defined in terms of the outcome of the integration process relative to the information specified by each of the two sources. We define a compromising integration as one always involving a compromise between the audible and visible sources. An averaging rule developed by Anderson and his colleagues (Anderson, 1981, 1982) is an example of a compromising integration. We define an enhancing integration as one that can produce a more

extreme decision than that warranted by either of the two sources when considered alone. The multiplicative rule of the fuzzy logical model of perception is an example of an enhancing integration process (Massaro & Oden, 1980). The Enhancing/Compromising distinction is not testable given conflicting auditory and visible sources, since both operations would predict a compromise in this situation.

As in some of our other contrasts, one test of the nature of the integration process can be performed on judgements of the single modality and bimodal speech syllables. Most tests of the integration rule are based only on bimodal trials, and extending the test to include the single-modality conditions permits a more powerful test. Consider judgements of a particular auditory syllable and a particular visual syllable and the relationship of these two judgements to the judgement of the bimodal speech syllable composed of the same auditory and visual levels. A compromising integration of the information given by the component dimensions predicts that the judgement of the bimodal syllable must be no more extreme than the judgements of either unimodal syllable. On the other hand, an enhancing integration of the information given by the component modalities predicts that the identification judgement of a bimodal syllable can be more extreme than either judgement given the separate modalities. This test appears to be scale-free, since it does not seem to be necessary to assume that the response scale is an interval or linear one.

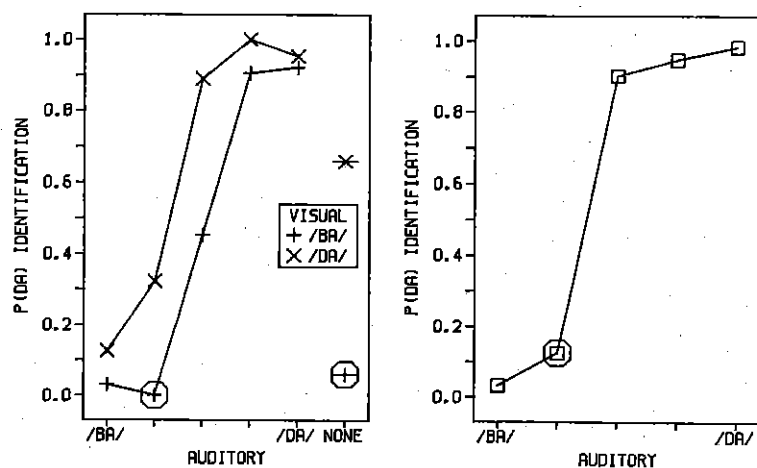


FIG. 2.6 Proportion of "da" identifications as a function of the auditory and visual levels of the speech event for the bimodal and visual alone conditions (left panel) and auditory alone condition (right panel). The circled points provide a test between compromising and enhancing integration rules.

Subjects identified as "ba" or "da" auditory, visual, and bimodal speech syllable is more extreme than both the 8% "da" judgements to the visual and the visual articulation "ba" and "da". Figure 2.6 gives the results for fourth-grade subjects. For the test of the nature of the integration rule, we focus on the three conditions involving the second auditory level and the "ba" articulation. These three conditions are circled in the graph. As noted previously, the compromising integration rule predicts that the judgement of the bimodal syllable cannot be more extreme than either of the single modality conditions. However, the 1% of "da" judgements to the bimodal syllable is more extreme than both the 8% "da" judgements to the visual dimension and the 13% "da" judgements to the auditory dimension when these dimensions are presented alone. This result provides strong evidence against a compromising integration rule and provides some support for the enhancing integration rule. The information from the two sources is integrated to produce an optimal decision rather than a simple compromise between the two dimensions.

CONCLUSIONS

The binary contrasts have been successful in eliminating plausible and intuitive interpretations of speech perception by ear and eye. The results also suggested that perceivers appear to evaluate evidence from both the audible and visible domains in bimodal speech perception. The evaluation of one source seems to occur independently of the properties of the other source. This evaluation process makes available continuous information indicating the degree to which relevant alternatives are supported. The sources of information are integrated to provide an overall degree of support for each alternative. The integration process is an enhancing one and not simply a compromising operation, for mild support from each of the two sources can be integrated to produce strong support for a given alternative.

Scientific progress might be considered in terms of the elimination of viable explanations. One question concerns to what extent bimodal speech perception is a unique human endeavour, or whether it represents one of a number of situations in which multiple sources of information support pattern recognition. At the current time, there is no reason to accept the former alternative over the latter. In fact, an immediate challenge to the uniqueness idea is to explain other contributions such as higher-order context to speech perception. Lexical and sentential context influence what is heard at the segmental level, and the question is why these contributions should be explained any differently than the contribution of visible speech.

DEVELOPMENT ISSUES

The framework guiding our research is also ideal for the study of developmental changes in processing audible and visible speech. Any theory of language processing must eventually confront the acquisition of the processes involved in this skill. Thus, developmental studies are central to evaluating theories of the processes responsible for observed differences and similarities in language processing with age. A developmental study implies that all of the questions that can be tested within the framework can also be tested in terms of interactions with development. Thus the binary oppositions generated in the present paradigm can now be asked as a function of developmental level. Each of the oppositions will be reviewed with special emphasis on possible interactions with development. I will then present our experimental studies that address these interactions.

SINGLE VERSUS MULTIPLE SOURCES ACROSS DEVELOPMENT

With respect to our first contrasts, we can inquire whether children also have both visible and audible sources of information available in their spoken discourse. Although a positive answer is likely, the development of auditory and visual processing might not coincide exactly. In addition, children might not have the same access to the sources as do adults. For example, children are short and adults are tall, and this difference in height might be expected to limit visible relative to audible speech for children. For older adults, high-frequency hearing loss with age might degrade the audible information relative to any loss of visual acuity with age. These observations lead to the impression that while multiple sources of information are available regardless of age, the relative quality of the various sources might vary systematically across the life span.

INTEGRATION VERSUS NON-INTEGRATION ACROSS DEVELOPMENT

There is a tradition in developmental theory, best represented by Piaget, that postulates an early stage of development in which integration does not occur. In the classic failure of conservation of mass, the young child attends only to the height and not the width of the glass of liquid (Anderson & Cuneo, 1978; Piaget & Inhelder, 1967). If perceiving speech is viewed as an analogous function, we might expect relatively young children to utilize just one of the available sources in perceptual judgement. McGurk and MacDonald (1976)

found fairly convincing support for integration for children as young as three years. Subjects were presented with auditory and bimodal speech events and asked to report what they heard the speaker saying. The bimodal speech events were dubbed articulations with two sources in conflict. Performance was highly accurate for auditory speech, but not for bimodal speech, in the sense that the judgements to bimodal speech did not correspond to the auditory domain. The authors observed combination errors representing responses including the components from both modalities, such as the response "bda" given an auditory "da" and a visual "ba". This result is difficult to explain in terms of the children utilizing just a single dimension in bimodal speech.

CATEGORICAL VERSUS CONTINUOUS INFORMATION ACROSS DEVELOPMENT

Much of the speech research with infants has been interpreted as supporting the categorical perception of certain phonetic contrasts (Eimas, Siqueland, Jusczyk and Vigorito, 1971; Gleitman and Wanner, 1982). In contrast, recent research with adult subjects has demonstrated that listeners have available continuous information corresponding to the degree to which a speech event represents a given perceptual category. For example, Massaro and Cohen (1983a, 1983b) demonstrated that the auditory "ba"-"da" continuum was perceived continuously rather than categorically. Yet this same dimension is supposedly perceived categorically by infants (Eimas, 1974). Thus, it is possible that young children may produce categorical results, whereas adults would not, given a "ba"-"da" auditory continuum paired with visible speech.

INDEPENDENT VERSUS NON-INDEPENDENT SOURCES ACROSS DEVELOPMENT

The next question concerns the independence of the two sources of information. According to the independence view, the auditory and visual inputs provide independent sources of information about the speech event. A contrasting assumption claims that the visual and auditory sources are not evaluated independently, but that the stimulus event is perceived holistically. Shepp (1978) and Smith and Kemler (1977, 1978) have proposed that there is a general developmental trend from holistic processing to dimensional processing. Preschool children supposedly process some stimuli holistically, whereas adults do not. If this hypothesis is correct for visual and auditory speech events, then holistic (non-independent) processing should be found for preschool children but not for adult subjects.

INTEGRATION RULE ACROSS DEVELOPMENT

There is now a body of work in information integration theory that supports less-than-optimal decision rules for young children as contrasted with adults. The judgement of the area of rectangles has been shown to follow a height \times width rule for children older than 7 and to follow either a height + width rule or a maximum linear extent rule for children between 3 and 5 years old (Anderson & Cuneo, 1978; Leon, 1982). It is conceivable, given the developmental differences in the perceptual domains of area and numerosity, that younger children will utilize something more akin to a compromising than to an enhancing rule.

DEVELOPMENTAL STUDIES

To address the binary issues developed in this paper, preschool and adult subjects were tested and compared (Massaro, 1984). They identified speech events consisting of synthetic speech syllables ranging from "ba" to "da", combined with a videotaped "ba" or "da" or no articulation. They were asked to view a speaker on a TV monitor and to indicate whether the speech event was "ba" or "da". Five levels of auditory information going from "ba" to "da" were factorially combined with three levels of visual information: "ba", no articulation and "da".

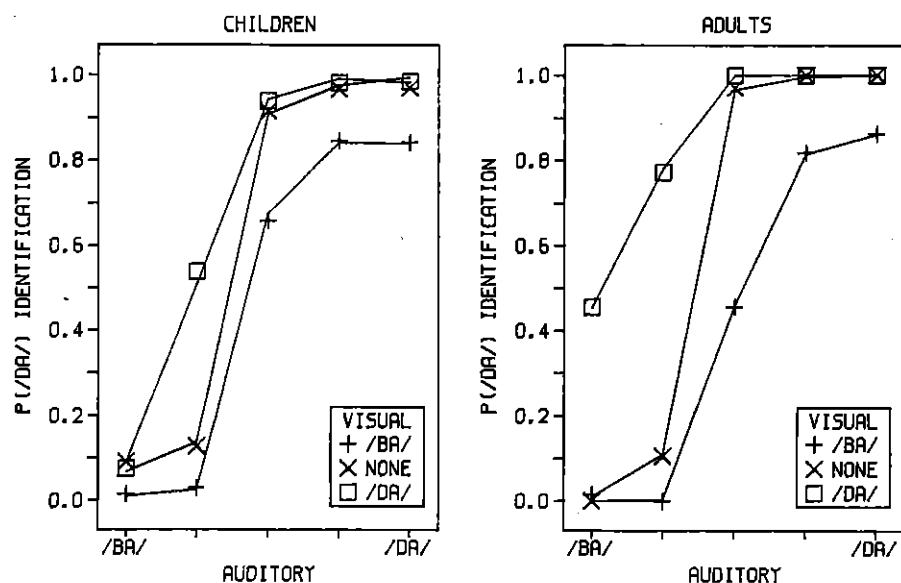


FIG. 2.7 Observed (points) and predicted (lines) proportion of "da" identifications as a function of the auditory and visual levels of the speech event. The predictions are for the fuzzy logical model of perception (after Massaro, 1984).

The left and right panels of Fig. 2.7 give the results for the children and adult subjects, respectively. The proportion of "da" responses as a function of the five levels along the auditory speech continuum is shown for the visual "ba", "da", and no articulation conditions. The average proportion of "da" responses increased significantly as a function of the level of the auditory stimulus. There was also a large effect on the proportion of "da" responses as a function of the visual stimulus. The interaction of these two variables was also significant, since the effect of the visual variable was smaller at the less ambiguous regions of the auditory dimension.

The fuzzy logical model of perception described in the next section gave equally good descriptions of the individual performance of children and adult subjects. This result provides evidence that the same fundamental processes are utilized by both children and adults in the perception of speech by ear and eye. The outcome is also consistent with the hypothesis of continuous and independent featural information for both children and adults. In addition, children seem to use the same type of enhancing integration rule used by adults. With respect to the processes involved in the evaluation and integration of audible and visible information in speech perception, preschool children behave similarly to adults.

Although it is always common to find large differences as a function of development, the failure to find any difference whatsoever would be disappointing. As illustrated in the binary tree in Fig. 2.8, however, the information values representing the two sources might differ for children and adults, even though the fundamental processes do not.

Differences in information value would be reflected in quantitative differences between the children and adult subjects. There was no group main

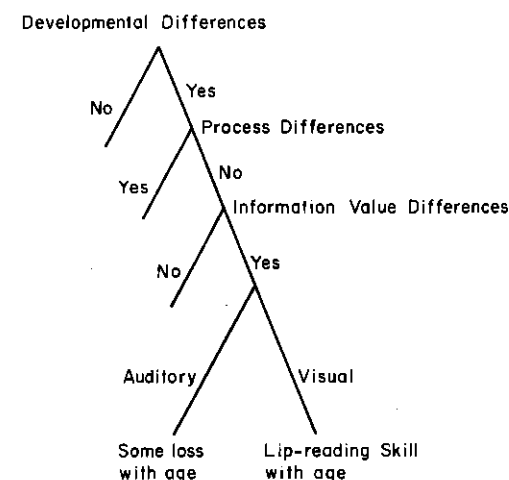


FIG. 2.8 Binary contrasts in terms of developmental differences in processes and/or information value.

effect, and group did not interact with the auditory variable. That is, the auditory source was as effective for children as for adult subjects. This result may be relatively peculiar to this age group and the contrast under study, since other results have revealed less sensitivity for children relative to adults. Zlatin and Koeningsknecht (1975) found increasing sensitivity to voice onset time differences with age, and Krause (1982) reported similar results for vowel duration differences.

The children were influenced by the visual variable only about half as much as were the adults. Figure 2.9 gives the effect of the visual variable for each subject in the two groups. Larger changes in the proportion of "da" identifications across the three visual conditions indicate a larger influence of the visual variable. The smaller influence of the visual variable for the children was highly consistent: 8 of the 11 children showed a smaller effect of the visual variable than 10 of the 11 adults.

To pursue the differences in the visual information for children and adults, Massaro et al. (1986) tested the idea that children are less sophisticated lip-readers than adults. The cues children use to distinguish a visual "ba" from a "da" may be less complete. If the visual variable is less informative for young children than for adults, its influence in bimodal speech perception will be smaller. More generally, we might expect a positive correlation between lip-

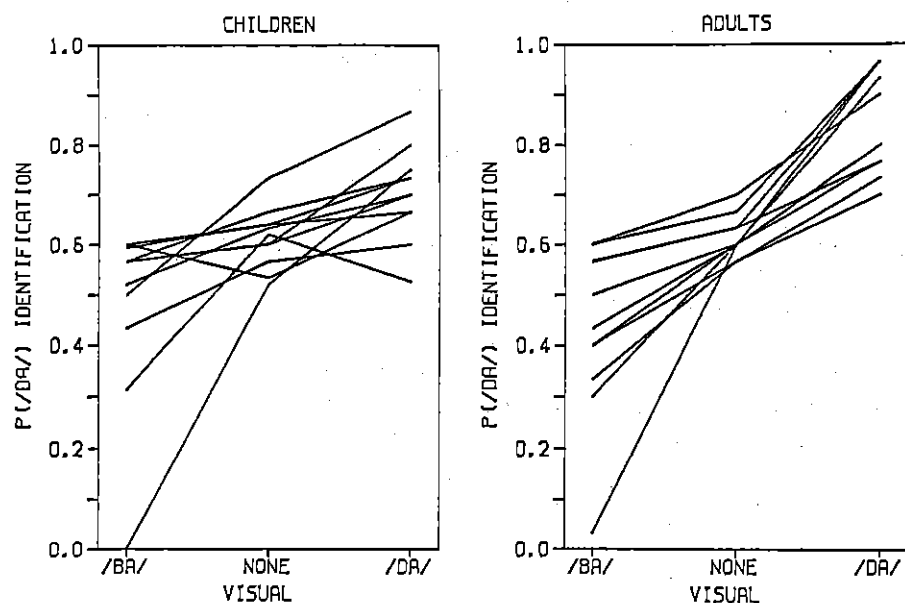


FIG. 2.9 The proportion of "da" identifications for individual subjects as a function of the visual level of the speech event (after Massaro, 1984).

reading ability and the visual influence. Thus, children should be poorer at lip-reading than are adults.

Adults and children were tested in a bimodal condition and a visual-only condition (Massaro et al., 1986). This provided a comparison between the size of the visual effect in the presence of auditory speech and the subject's ability to identify accurately the two types of visual articulation when no sound is present. As expected, the visual variable had a much larger influence on the adults' than on the children's judgements of bimodal speech. The new result showed adult subjects to be much better at lip-reading than children. A positive correlation between subjects' proportion correct in the visual-only condition and the size of their visual effect in the bimodal condition was significant for each subject group and for the combined results of both groups. The developmental difference in the size of the visual influence on bimodal speech may therefore be due to children's poorer ability to lip-read.

In summary, the developmental studies reveal that the answers to the binary oppositions discussed in the first section of this chapter do not change from preschool to adulthood. However, the information value of visible speech is somewhat less for young children than for adults. The acquisition of lip-reading seems to lag behind the utilization of auditory information in speech perception. Pursuing the study of developmental changes appears to offer a productive approach to the study of perceiving bimodal speech.

A FUZZY LOGICAL MODEL OF PERCEPTION

The constraints that we have uncovered in our experimental investigation of the perception of bimodal speech eliminate a fairly large number of potential candidates for models of the phenomena. Even explanations developed explicitly to describe the original auditory-visual blend illusions are contradicted by the results. As an example, the idea that bimodal perception results from place of articulation by eye and from manner and voicing by ear is clearly false (see Summerfield, Chapter 1, this volume). Our experiments establish that both audible and visible speech can simultaneously contribute to the perception of *place* of consonants. On the other hand the rapidly accumulating body of results is consistent with a more general model of perceptual recognition.

According to the model, many different objects and events are recognized in accordance with a general pattern recognition algorithm (Massaro, 1979; Oden & Massaro, 1978). Three operations are viewed as necessary for any complete account of pattern recognition: evaluation, integration and classification. Continuously valued sources of information are evaluated independently of one another and integrated with respect to prototype representations in memory. The likelihood of classification is equal to the goodness of

match of the stimulus information with the relevant prototype relative to the total of the goodness of match values of all contending prototypes. The model is called a fuzzy logical model of perception, and it is helpful to discuss the concept of fuzzy logic in this particular domain and how it is used in the model.

Fuzzy logic is used in the model to provide a common metric to relate the wide variety of sources of information to one another. The use of some abstract common metric is necessary in any theory. Summerfield (Chapter 1, this volume), for example, suggests a variety of ways in which to relate the audible and visible speech components to a single common metric. Different metrics are described. Our aim here, however, is to suggest a more general and powerful perceptual principle. Our use of fuzzy logic is primarily one of convenience, since it provides what we believe are the essential properties of the common metric needed to capture recognition of patterns defined by multiple sources of information. The central properties are some form of representation of the continuous degree to which an alternative is supported by a source of information and the operations for relating the representations from different sources to one another.

Fuzzy logic represents propositions as neither entirely true nor false, but rather with continuous truth values. Within this framework, it is possible to say that a segment is relatively "ba"-like and not very "da"-like. Ordinary logical quantification would require that the alternative be "ba" or some other discrete alternative. Within fuzzy-logic theory (Zadeh, 1965; Goguen, 1969), the truth of a proposition is represented by values between 0 and 1 corresponding to the range between completely false and completely true. It should be noted that the concept of fuzzy truth values is different from probability. If we say that a segment is "ba" to degree 0.2, it does not mean that there is a 0.2 probability that the segment is "ba". Rather, it is true that the segment represents "ba" to degree 0.2. The standard logical operations of negation, conjunction, and disjunction of truth values, $t(x)$, can be described within the theory. Generalizing the definition for negation in standard logic, the additive complement can be defined as

$$t(\sim x) = 1 - t(x) \quad (1)$$

where $t(\sim x)$ is the truth of not x . Thus given a 0.2 truth value for "ba", the truth value of not-"ba" would be 0.8.

Extending the formalization of standard set theory, Zadeh (1965) derived the minimization rule for the conjunction (\wedge) of the truth values of two events a and b :

$$t(a \wedge b) = \min[t(a), t(b)] \quad (2)$$

Goguen (1969), on the other hand, suggested the possibility of a multiplicative rule to overcome certain logical limitations in the minimization rule

$$t(a \wedge b) = t(a) \times t(b) \quad (3)$$

The psychological literature has primarily considered an averaging definition for the conjunction of two events

$$t(a \wedge b) = [t(a) + t(b)]/2 \quad (4)$$

Disjunction rules can be derived from these conjunction rules using DeMorgan's law, but the nature of the disjunction process will not be pursued here.

Research in a number of domains has supported the psychological reality of the multiplicative rule over the other two forms of conjunction. Massaro and Cohen (1976) studied the conjunction of voice-onset time and fundamental frequency as perceptual cues to the "si"-"zi" distinction. A multiplicative combination of the cues values described the results about four times more accurately than did an averaging combination. Oden (1977) investigated which set of definitions of fuzzy logical conjunction best fit judgements about logical combinations of pairs of statements about class membership functions (e.g., a bat is a bird, and a refrigerator is furniture). The data from the experiment were better explained by the multiplication rule than by the minimization or averaging rules. Thus, the multiplicative rule is assumed in the present application of the model.

The fuzzy logical model of perception specifies three operations between stimulus and categorization (Massaro and Oden, 1980; Oden and Massaro, 1978). The first operation, featural evaluation, assesses the independent sources of input transduced by the sensory systems and assigns truth values according to the degree to which each alternative is supported by each source. The formant transitions of a stop consonant, for example, would be evaluated with respect to the degree to which each alternative is supported.

The second operation involves the integration of the truth values with respect to the prototype representations of the alternatives. For bimodal speech, a syllable prototype would represent the conjunction of both auditory and visual properties defining the syllable. The integration operation would consist of replacing the respective properties of each prototype with the corresponding truth values of the relevant speech event. The conjunction of these truth values determines to what degree each prototype is realized in the pattern. To distinguish between "ba" and "da", the truth values corresponding to the evaluation of the auditory formant transitions and the evaluation of the visual lip movements would be integrated and matched against the prototypes for the relevant alternatives "ba" and "da".

The third operation of recognition processing is pattern classification. During this stage, the merit of each relevant prototype is evaluated relative to the summed merits of the other relevant prototypes. The relative goodness of the prototype gives the proportion of times it would be selected as a response or its judged magnitude. This is similar to Luce's (1959) choice rule, which is

based on the relative strengths of the alternatives in the candidate set. In pandemonium-like terms (Selfridge, 1959), we might say that it is not how loud some demon is shouting but rather the relative loudness of that demon in the crowd of relevant demons. The likelihood of a "da" identification would be equal to the goodness-of-match value to the alternative "da" relative to the sum of the goodness-of-match values for all of the relevant alternatives.

In an experiment manipulating the auditory and visual properties, every relevant alternative is represented by a prototype description of the appropriate auditory and visual information. Consider the experiment described in the section *Integration versus Non-integration*. Subjects are asked to identify bimodal speech events, auditory alone, the visual alone trials as illustrated in Fig. 2.2. For the bimodal trials, an auditory synthetic syllable along a nine-step "ba"-to-"da" continuum is dubbed onto a videotape of the speaker saying "ba" or "da". In addition, the auditory speech stimuli are presented alone with no lip movements on some trials, and the "ba" and "da" articulations are presented without sound on other trials. The subjects are permitted eight alternatives determined from a pilot study based on an open-ended set of response alternatives.

Defining the important auditory information as the onsets of the second and third formants ($F2-F3$) and the important visual cue as lip closure, the prototypes for "da" and "ba" can be described by

"da": Slightly falling $F2-F3$ & Open lips

"ba": Rising $F2-F3$ & Closed lips

Given a prototype's *independent* specifications for the auditory and visual sources, the value of one source cannot change the value of the other source at the prototype matching stage. The other alternatives would be defined analogously in terms of the audible and visible information. For example, the alternatives "tha" and "va" might be defined as

"tha": Nearly level $F2-F3$ & Nearly open lips

"va": Slightly rising $F2-F3$ & Nearly closed lips

For predicting the responses, consider the probability of a "da" response given a bimodal speech syllable A_iV_j consisting of the i th level of the auditory continuum and the j th level of the visual continuum. The truth value a_i represents the degree to which the auditory source supports "da" and analogously for v_j representing support from the visual source. The multiplicative conjunction of the two sources of information gives the value a_iv_j for the goodness of match of the prototype "da" with the syllable A_iV_j . The probability of a "da" judgement is equal to this goodness of match value divided by the total of all the relevant goodness of match values

$$P(\text{"da"}:A_iV_j) = \frac{a_iv_j}{\text{total}} \quad (5)$$

An analogous prediction is given for each response alternative. The numerator gives the goodness of match value of the corresponding prototype with the test syllable and the denominator gives the same total of all the relevant goodness of match values.

A critical assumption of the model is that the featural value given a particular level of one source is identical in the unimodal and bimodal conditions. That is, the degree of "da"-ness given by a visual "da" is identical when the visual information is presented alone for lip-reading and when it is combined with auditory speech. The response probability can be predicted by the same equation given for the bimodal conditions. The missing source of information in the unimodal conditions simply would be assigned the completely ambiguous truth value 0.5.

The expanded factorial design provides a challenging test of the model, since a larger number of observations is predicted with the same number of parameters relative to just the bimodal speech task. Given 2 visual levels crossed with 9 auditory levels, we have 18 stimulus conditions with just bimodal speech and 29 stimulus conditions when the auditory-alone and visual-alone conditions are included. The predictions of the fuzzy logical model require free parameters corresponding to the 9 levels along the auditory continuum and the 2 levels along the visual continuum in the present task. What is most relevant is that the same truth value is given a particular level of a dimension at the feature evaluation stage in both the single dimension and bimodal condition. Thus, the model predicts the 9 auditory-alone conditions, the 2 visual-alone conditions, and the 18 bimodal conditions for a total of 29 independent observations with 11 free parameters for each response alternative.

There were 8 valid response alternatives: "ba", "da", "bda", "dba", "va", "tha", "ga", and "other". Thus we have 29 times 8 minus 29, or 203, independent observations predicted by 11 times 8, or 88, parameters. The reason that the number of independent observations is reduced by 29 is that the response probabilities must add to one. Given seven response probabilities to a particular stimulus, the eighth is determined.

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

that is used is the root mean squared deviation, which is the square root of the average squared deviation between the predicted and observed points.

Figure 2.10 gives the observed results averaged across subjects. For the predictions of the model, each of the seven phonetic prototypes and the alternative "other" was permitted a unique truth value representing the auditory support for that alternative and a unique truth value representing the visual support for that alternative. The model was fit separately to the results of each of the eight individual subjects. Figure 2.10 also gives the predicted results averaged across subjects. As can be seen in the figure, the model provides a good description of the results. The root mean squared deviation between the predicted and observed values varied between 0.022 and 0.041 across the eight subjects, with an average value of 0.030. This result is very impressive since it is predicting an essentially open-ended set of response alternatives with identical information for the unimodal and bimodal speech stimuli.

Given the good description of the model, the parameter values shown in Table 2.1 should be psychologically meaningful. The parameter values represent the degree to which a source of information supports a particular alternative. As can be seen in the table, the parameter values reflect the observed results and are consistent with the properties of the audible and

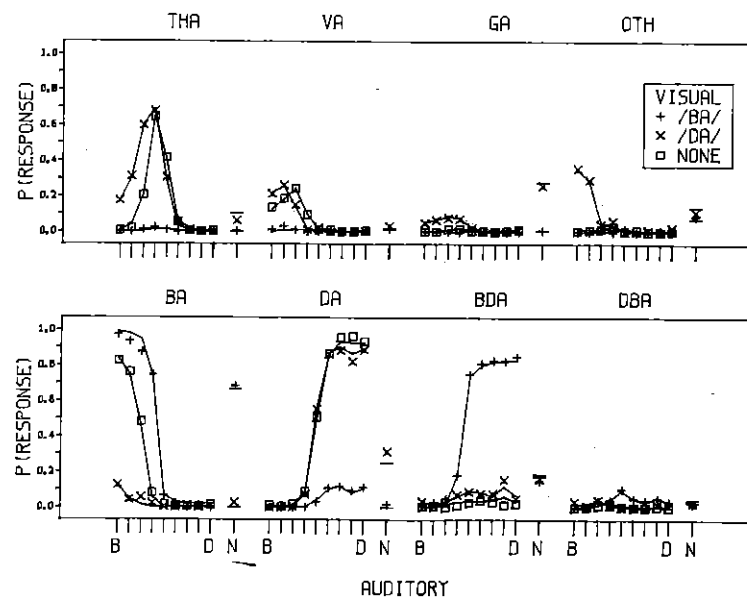


FIG. 2.10 Proportion of identification for the eight alternatives as a function of the auditory, visual and bimodal conditions. Observations given by the points and predictions of the fuzzy logical model of perception by the lines.

TABLE 2.1
Average parameter values for the description of the fuzzy logical model
of perception of an experiment

Source	Response alternatives							
	"ba"	"da"	"bda"	"dba"	"tha"	"va"	"ga"	other
visual "ba"	0.945	0.001	0.300	0.078	0.001	0.026	0.010	0.150
visual "da"	0.002	0.534	0.364	0.074	0.245	0.041	0.508	0.297
auditory "ba"	0.996	0.001	0.004	0.011	0.007	0.175	0.001	0.015
auditory 2	0.953	0.001	0.009	0.001	0.050	0.274	0.004	0.021
auditory 3	0.799	0.006	0.035	0.085	0.377	0.485	0.025	0.034
auditory 4	0.095	0.136	0.054	0.040	0.992	0.177	0.037	0.050
auditory 5	0.001	0.751	0.054	0.028	0.606	0.026	0.014	0.023
auditory 6	0.001	0.991	0.054	0.008	0.105	0.030	0.006	0.009
auditory 7	0.001	0.998	0.045	0.003	0.015	0.003	0.001	0.004
auditory 8	0.001	0.926	0.060	0.009	0.001	0.001	0.001	0.001
auditory "da"	0.001	0.998	0.037	0.003	0.126	0.009	0.015	0.005

Note: There are 2 visual levels, 9 auditory levels, 8 response alternatives. The values represent the degree of support of the source of information for the alternatives.

visible speech. For example, visual "ba" supports the alternative "ba" to degree 0.946, but also the alternative "bda" to degree 0.300. Analogously, auditory 5 (the fifth level along the "ba"-"da" continuum) supports "da" to degree 0.751, but also the alternative "tha" to degree 0.606. Thus not only does the model describe the results accurately, the parameter values are meaningful. It is a challenge to alternative theories to formalize a model that can provide an equally good description of the fine details of results of unimodal and bimodal speech perception.

ALTERNATIVE MODELS

The fuzzy logical model of perception is consistent with the outcomes of the binary contrasts listed in Fig. 2.1 and also provides a good description of the quantitative results. For comparison, it is worthwhile to consider the description given by models based on the alternative outcomes of each contrast. It is at least logically possible that these alternative models might also provide an adequate description of the results, and thus it is important to assess this possibility. In addition, illustrating that the alternative models give inadequate predictions would marshal additional support against these alternatives as well as strengthen the case for the fuzzy logical model of perception. We will derive predictions for models based on non-integration, categorical perception, non-independence, and compromising integration.

NON-INTEGRATION

A model based on non-integration assumes that the subject can use only the auditory or only the visual but not both dimensions of the speech event on a given trial. Given an auditory or visual trial, the subject uses the relevant dimension with probability one. This is a reasonable assumption given that only one modality is present and no bimodal integration is required to use that dimension. On bimodal trials, the subject uses the auditory dimension with some probability p and uses the visual dimension with probability $1 - p$. On proportion p of the trials, the judgement is determined by the auditory dimension, and on proportion $1 - p$ of the trials it is determined by the visual dimension. Predicted performance given a bimodal speech event is thus a simple weighted average of the two identifications given the single-dimension speech events. If the probability of a "da" response given an auditory stimulus A_i is a_i and the probability of a "da" response given a visual stimulus V_j is v_j , then the probability of a "da" response given a bimodal speech event $A_i V_j$ is equal to

$$P(\text{"da"}: A_i V_j) = p a_i + (1 - p) v_j \quad (6)$$

An analogous prediction can be derived for each response alternative in which the p value is constant across alternatives and the a_i and v_j are unique for each response alternative. For predicting the eight-alternative study given in Fig. 2.10, the 9 auditory levels and 2 visual levels give 11 times 8, or 88, parameters plus one p parameter, for a total of 89 parameters.

CATEGORICAL PERCEPTION

The basic idea of a model based on categorical perception is that information from each of the modalities is categorical rather than continuous. Both sources are available on a single bimodal trial, but the information from each source is in categorical form. Decisions regarding the alternative percepts are made separately to the auditory and visual sources, and the identification is based on some intergration of these separate decisions. For each response alternative, there are four possible outcomes for a particular combination of auditory and visual information. Considering the "da" decision, the visual and auditory decisions could be "da"/"da", "da"/not-"da", not-"da"/"da", or not-"da"/not-"da". If the two decisions to a given speech event agree, the identification response can follow either source. When the two decisions disagree, it is assumed that the subject will respond with the decision of the auditory source on some proportion p of the trials, and with the decision of the visual source on the remainder $(1 - p)$ of the trials. The weight p reflects the relative dominance of the auditory source.

The probability of a "da" identification response, $P(\text{"da"})$, given a particular auditory-visual speech event, $A_i V_j$, would be:

$$P(\text{"da"}: A_i V_j) = (1 - a_i) v_j + [p a_i (1 - v_j)] + [(1 - p)(1 - a_i) v_j] + [0(1 - a_i)(1 - v_j)] \quad (7)$$

where i and j index the levels of the auditory and visual modalities, respectively. The a_i value represents the probability of a "da" decision given the auditory level i and v_j is the probability of a "da" decision given the visual level j . Each of the four terms in the equation represents the likelihood of one of the four possible outcomes multiplied by the probability of a "da" identification response given that outcome. In the model, each unique level of the auditory stimulus requires a unique parameter a_i , and analogously for v_j .

Equation (7) can be simplified algebraically to

$$P(\text{"da"}: A_i V_j) = p a_i + (1 - p) v_j \quad (8)$$

The predictions for the single modality conditions are given simply by the probability of an identification to that dimension. As an example, the probability of a "da" identification is predicted to be a_i given the auditory-alone condition and v_j given the visual-alone condition.

The modelling of "da" responses thus requires 9 auditory parameters plus 2 visual parameters. Each of the other 7 response alternatives needs an analogous equation to that given above, with an additional 11 free parameters. An additional p value would be fixed across all conditions, giving a total of 89 parameters. For any particular auditory-visual combination, the sum of the 4 decision probabilities to a given source also has to be constrained to be ≤ 1 . This follows from the categorical assumption that a given source is categorized as only a single category on any given presentation.

NON-INDEPENDENCE

It is very difficult to formalize and test a non-independence model, unless a particular type of dependence between the sources is specified exactly. If no type of dependence is assumed, it is necessary to estimate a unique parameter for each unique set of experimental conditions. Thus, the dependence model would require as many parameters as there are independent conditions. This violation of parsimony seems sufficient to reject a non-independence model as a meaningful description of performance. In addition, if the contribution of one source is dependent on the value of the other, any model assuming independent contributions of each source must fail. To the extent that the fuzzy logical model of perception gives an adequate description of the results, we have evidence against the non-independence assumption.

Massaro and Cohen (1983b, Experiment 2) tested a particular non-

independence model against the individual results of the seven subjects. The experiment involved identification of bimodal syllables generated by the factorial combination of visual "ba", no articulation, and visual "da", with 9 levels along a synthetic speech "ba"–"da" continuum. The results permitted an unqualified rejection of this form of dependence in the perception of bimodal speech. Until some other form of dependence is demonstrated to give an adequate description of the results, we reject non-independence in favour of independent evaluation of auditory and visual dimensions in perception of bimodal speech.

COMPROMISING INTEGRATION

The idea of compromising integration can be formulated by assuming an averaging conjunction rule, as given in Equation (3). Thus each modality provides continuous and independent evidence or truth values as in the fuzzy logical model, but the sources are averaged rather than multiplied. In this case, the probability of a "da" response would be predicted to be

$$P(\text{"da"}:A_iV_j) = (a_i + v_j)/2 \quad (9)$$

Analogous to the other models, the predictions for the single modality conditions are given simply by the truth value given that dimension. As an example, the probability of a "da" identification is predicted to be a_i given the auditory-alone condition and v_j given the visual-alone condition. As in the formulation of the other models, an analogous equation predicts the likelihood of each of the other response alternatives. Thus 11 parameters are necessary for each of the 8 alternatives for a total of 88 parameters.

A more general and realistic extension of averaging is a weighted averaging in which the auditory and visual modalities can receive different weights. In this case, representing the weight received by the auditory source as w and the weight received by the visual source as $1-w$, Equation (9) becomes

$$P(\text{"da"}:A_iV_j) = wa_i + (1-w)v_j \quad (10)$$

if w constrained to lie between zero and one. Since w is constant across all conditions, one additional parameter is necessary, for a total of 89 free parameters.

We have formulated models based on the rejected branches of the tree of binary contrasts. The hypothesis of non-independence also warrants rejection, primarily because it is not quantifiable in a parsimonious manner. What is surprising and yet convenient for our purposes is that non-integration, categorical perception, and compromising integration all make identical

quantitative predictions. That is, Equations (6), (8), and (10) representing the assumptions of non-integration, categorical perception and compromising integration, respectively, are mathematically identical. Thus a quantitative test of the predictions of the same mathematical model independently developed for non-integration, categorical perception and compromising integration should be highly informative and should provide an appropriate comparison to the fuzzy logical model.

As in the test of the fuzzy logical model of perception, the contrasting model was fit to the results of each of the eight individual subjects. Figure 2.11 gives the observed results along with the predicted results averaged across subjects. As can be seen in the figure, the model provides a very poor description of the results. The root mean squared deviation between the predicted and observed values varied between 0.195 and 0.241 across the eight subjects with an average value of 0.216. Relative to the fuzzy logical model of perception, the alternative model derived from non-integration, categorical perception, or compromising integration predicts the observed results about seven times less accurately. The advantage of the fuzzy logical model cannot be due to the number of parameters, since it actually has one fewer than the contrasting model. Thus, the model tests confirm the outcomes of the binary contrasts, as they should.

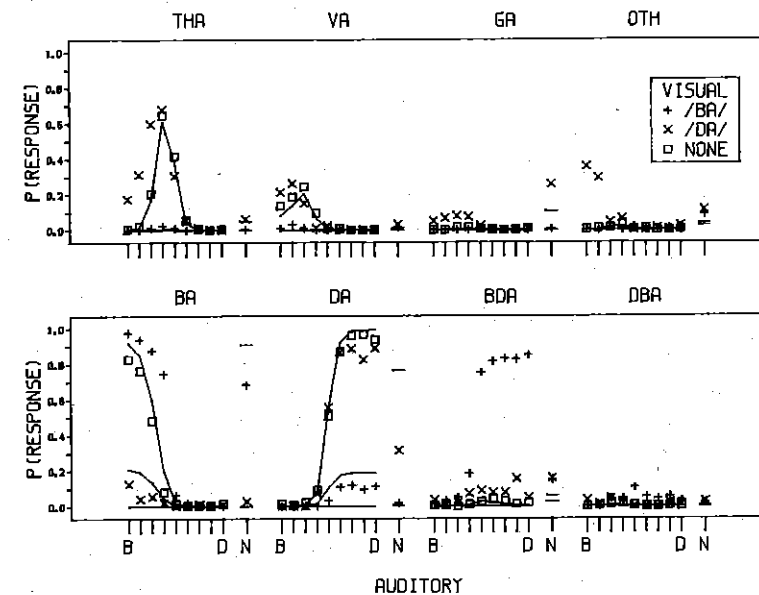


FIG. 2.11 Proportion of identifications for the eight alternatives as a function of the auditory, visual and bimodal conditions. Observations given by the points and predictions of the models based on non-integration, categorical perception and compromising integration by the lines.

INTERACTIVE-ACTIVATION MODELS

The constraints on theories of speech perception provided by the current research are apparent in their relevance to interactive-activation models, such as TRACE (McClelland & Elman, 1986). The model is similar in many respects to the fuzzy logical model of perception, and quantitative tests between the models will be difficult, if not impossible. Several outcomes of the binary contrasts, however, stand in marked contradiction to the fundamental assumptions of TRACE. Three levels of units are used in TRACE: feature, phoneme and word. The interaction among the units involves both activation and inhibition. Features activate phonemes, which activate words, and activation of some units at a particular level inhibits other units at the same level. In addition, activation of higher-order units activates their lower-order units—for example, activation of “b” activating the feature voiced. Given that multiple units at one level simultaneously activate units at a higher level, the model predicts integration and also predicts that integration can be enhancing, not just compromising. These two properties of the model agree with the outcomes of the binary contrasts.

Two other properties of the model are contradicted by the binary contrasts, however. The top-down activation from phoneme to feature produces non-independence at the featural level, a result not found for speech perception by ear and eye. The TRACE model predicts that visible speech activating “b” would result in top-down activation of the audible features of “b”, contrary to our observations of the independence of the features of audible and visible speech. Although the TRACE model assumes continuous levels of activation, the interactions among the activations tend to produce outputs that are categorical rather than continuous. This property is falsified by the ability of perceivers to transmit continuous information in speech perception. It is encouraging that binary contrasts also prove informative for evaluating new theories not developed at the time that the binary contrasts were carried out.

RETROSPECTION AND PROGNOSTICATION

We have approached the problem of speech perception by ear and by eye within the framework of falsification and strong inference. The issues that we have addressed seem fundamental to developing a psychological understanding of the phenomenon. The methods of information integration and mathematical model testing appear to be ideally suited for addressing some of the issues. The experiments have been reasonably successful in providing answers to the questions. On the basis of the outcomes, perceiving speech by ear and by eye is described within the context of a general theory of

perceptual recognition. This theory provides a common metric for evaluating and integrating multiple sources of information in pattern classification. Future work will be necessary in order to explore variations within the context of each binary contrast. We can also expect other contrasts and theoretical alternatives to present themselves as our understanding of communicating by ear and eye evolves.

ACKNOWLEDGEMENT

The writing of this chapter and the research reported in it were supported, in part, by NINCDS Grant 20314 from the Public Health Service and Grant BNS-83-15192 from the National Science Foundation. The chapter was written while I was a member of the project *Perception and Action* at the Center for Interdisciplinary Research, University of Bielefeld. I would like to thank the Center and the group members for providing an ideal atmosphere for scholarly pursuits. Michael M. Cohen made important contributions to the research enterprise.

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.
- Anderson, N. H. & Cuneo, D. O. (1978). The height+width rule in children's judgments of quantity. *Journal of Experimental Psychology: General*, **107**, 335–378.
- Chandler, J. P. (1969). Subroutine STEPIT—Finds local minima of a smooth function of several parameters. *Behavioural Science*, **14**, 81–82.
- Cohen, M. M. (1984). *Processing of visual and auditory information in speech perception*. Dissertation, University of California, Santa Cruz.
- Eimas, P. D. (1974). Auditory and linguistic processing of cues for place of articulation by infants. *Perception and Psychophysics*, **16**, 513–521.
- Eimas, P. D., Siqueland, E. R., Jusczyk, P. & Vigorito, J. (1971). Speech perception in infants. *Science*, **171**, 303–306.
- Fodor, J. A., Bever, T. G. & Garrett, M. F. (1974). *The psychology of language*. New York: McGraw-Hill.
- Ganong, W. F. III (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, **6**, 110–125.
- Gleitman, L. R. & Wanner, E. (1982). Language acquisition: The state of the state of the art. In E. Wanner & L. R. Gleitman (Eds.), *Language acquisition: The state of the art*. Cambridge: Cambridge University Press.
- Gielen, S. C., Schmidt, R. A. & Van Den Heuvel, P. J. M. (1983). On the nature of intersensory facilitation of reaction time. *Perception and Psychophysics*, **34**, 161–168.
- Goguen, J. A. (1969). The logic of inexact concepts. *Synthese*, **19**, 325–373.
- Isenberg, D., Walker, E. C. T. & Ryder, J. M. (1980). A top-down effect on the identification of function words. *Journal of the Acoustical Society of America*, **68**, AA6 (abstract).
- Krause, S. E. (1982). Vowel duration as a perceptual cue to postvocalic consonant voicing in young children and adults. *Journal of the Acoustical Society of America*, **71**, 990–995.

- Leon, M. (1982). Extent, multiplying, and proportionality rules in children's judgments of area. *Journal of Experimental Child Psychology*, **33**, 124-141.
- Luce, R. D. (1959). *Individual choice behaviour*. New York: Wiley.
- MacDonald, J. & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics*, **24**, 253-257.
- Marslen-Wilson, W. & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, **10**, 29-63.
- Massaro, D. W. (1975a). *Experimental psychology and information processing*. Chicago: Rand-McNally.
- Massaro, D. W. (Ed.) (1975b). *Understanding language: An information processing analysis of speech perception, reading and psycholinguistics*. New York: Academic Press.
- Massaro, D. W. (1979). Reading and listening (Tutorial paper). In P. A. Kollers, M. Wroldstad & H. Bouma (Eds.) *Processing of visible language, Vol. 1*. New York: Plenum, pp. 331-354.
- Massaro, D. W. (1984). Children's perception of visual and auditory speech. *Child Development*, **55**, 1777-1788.
- Massaro, D. W. (in press). Information-processing theory and strong inference: A paradigm for psychological inquiry. In H. Heuer & A. F. Sanders (Eds.), *Tutorials on perception and action*. Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Massaro, D. W. & Cohen, M. M. (1976). The contribution of fundamental frequency and voice onset time to the /zi/-/si/ distinction. *Journal of the Acoustical Society of America*, **60**, 704-717.
- Massaro, D. W. & Cohen, M. M. (1983a). Categorical or continuous speech perception: A new test. *Speech Communication*, **2**, 15-35.
- Massaro, D. W. & Cohen, M. M. (1983b). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, **9**, 753-771.
- Massaro, D. W. & Cohen, M. M. (1983c). Phonological context in speech perception. *Perception and Psychophysics*, **34**, 338-348.
- Massaro, D. W. & Oden, G. C. (1980). Speech perception: A framework for research and theory. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice*. New York: Academic Press.
- Massaro, D. W., Thompson, L. A., Barron, B. & Laren, E. (1986). Developmental changes in visual and auditory contributions to speech perception. *Journal of Experimental Child Psychology*, **41**, 93-113.
- McClelland, J. L. & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, **18**, 1-86.
- McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746-748.
- Oden, G. C. (1977). Integration of fuzzy logical information. *Journal of Experimental Psychology: Human Perception and Performance*, **3**, 565-575.
- Oden, G. C. & Massaro, D. W. (1978). Integration of featural formation in speech perception. *Psychological Review*, **85**, 172-191.
- Piaget, J. & Inhelder, B. (1967). *The child's conception of space*. New York: Basic Books.
- Platt, J. R. (1964). Strong inference. *Science*, **146**, 347-353.
- Popper, K. (1959). *The logic of scientific discovery*. New York: Basic Books.
- Raab, D. H. (1962). Statistical facilitation of simple reaction times. *Transactions of the New York Academy of Sciences*, **24**, 574-590.
- Selfridge, O. G. (1959). Pandemonium: A paradigm for learning. In *Symposium on the mechanisms of thought processes*. London: HMSO.
- Shepp, B. D. (1978). From perceived similarity to dimensional structure. In E. Rosch & B. Lloyd (Eds.), *Cognition and categorization*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

- Smith, L. B. & Kemler, D. G. (1977). Developmental trends in free classification: Evidence for a new conceptualization of perceptual development. *Journal of Experimental Child Psychology*, **24**, 279-298.
- Smith, L. B. & Kemler, D. G. (1978). Levels of experienced dimensionality in children and adults. *Cognitive Psychology*, **10**, 502-532.
- Tyler, L. K. & Wessels, J. (1983). Quantifying contextual contributions to word-recognition processes. *Perception & Psychophysics*, **34**, 409-420.
- Walden, B. E., Prosek, R. A., Montgomery, A. A., Scherr, C. K. & Jones, C. J. (1977). Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research*, **20**, 130-145.
- Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, **8**, 338-353.
- Zlatin, M. A. & Koeningsknecht, R. A. (1975). Development of the voicing contrast: Perception of stop consonants. *Journal of Speech and Hearing Research*, **18**, 541-553.