Journal of Experimental Psychology:
Human Perception and Performance
1983, Vol. 9, No. 5, 753-771

Evaluation and Integration of Visual and Auditory Information in Speech Perception

Dominic W. Massaro and Michael M. Cohen
Program in Experimental Psychology, University of California, Santa Cruz

Three experiments were carried out to investigate the evaluation and integration of visual and auditory information in speech perception. In the first two experiments, subjects identified /ba/ or /da/ speech events consisting of high-quality synthetic syllables ranging from /ba/ to /da/ combined with a videotaped /ba/ or /da/ or neutral articulation. Although subjects were specifically instructed to report what they heard, visual articulation made a large contribution to identification. The tests of quantitative models provide evidence for the integration of continuous and independent, as opposed to discrete or nonindependent, sources of information. The reaction times for identification were primarily correlated with the perceived ambiguity of the speech event. In a third experiment, the speech events were identified with an unconstrained set of response alternatives. In addition to /ba/ and /da/ responses, the /bda/ and /tha/ responses were well described by a combination of continuous and independent features. This body of results provides strong evidence for a fuzzy logical model of perceptual recognition.

Ventriloquists rely to some extent on an intersensory bias between vision and audition. When the seen and heard locations of a speaker's voice are somewhat discrepant, the speech is heard at the speaker's location (Welch & Warren, 1980). Ventriloquists do not throw their voices; rather, a listener's percept is thrown by the visual input of the apparent speaker. An analogous bias has been reported by McGurk and MacDonald (1976), who dubbed a labial speech sound /ba-ba/ onto the visual articulation of a velar speech sound /ga-ga/. Subjects viewing and listening to the videotape often reported hearing the sequence /da-da/. The bias of audition by sight presents a challenging phenomenon to current theories of perception, pattern recognition, and speech identification. The current research assesses the nature of the visual contribution to speech perception within the context of extant hypotheses about perceptual recognition of speech.

This research was supported by National Institute of Mental Health Grant MH-35334 to Dominic W. Massaro. We wish to thank Ray Gibbs for his helpful comments on an earlier version of this article and Quentin Summerfield and Arthur Samuel for their useful reviews of the manuscript.

Requests for reprints should be sent to Dominic W. Massaro, Program in Experimental Psychology, University of California, Santa Cruz, California 95064.

A speaker usually provides both auditory and visual information to the listener. Auditory information appears to be relevant to all important speech distinctions, whereas visual information is relevant primarily for place of articulation. Visual information has been shown to distinguish the subset of plosive consonants /b, p, m/ from the subsets /d, t, n/ and /g, k, n/ (Dodd, 1977; MacDonald & McGurk, 1978; McGurk & MacDonald, 1976; Summerfield, 1979). A labial articulation is easily discriminated from an alveolar or velar articulation by observing whether or not the lips are closed in articulation. It has even been demonstrated that hearing-impaired listeners can distinguish between alveolar and velar consonants, especially if the inside of the mouth is illuminated (Erber, 1972).

Although auditory information is potentially sufficient for accurate speech perception and is the only source of information available in many communication situations (e.g., the telephone), manipulating visual information in speech perception offers a valuable paradigm for gaining insights into the processes involved. First, the visual articulation might be an important source of information for the perceiver. It might be the case, in face-to-face communication, that the visual information not only supplements ambiguous auditory information but is a powerful source even when

the auditory information is unambiguous. Second, the integration of the auditory and visual information might be a fundamental process in speech perception. If so, the study of this process not only would provide insights into the process itself but also might be relevant to the general case of information integration in perception. Third, auditory information may be degraded in a number of natural situations, and it is important to assess the potential contribution of visual information to communication. Finally, the study of audio-visual speech perception is highly relevant to the general issue of intersensory perception (Welch & Warren, 1980), which is a problem that has been neglected in most research in perception.

Experiment 1

Previous studies have shown that the visual articulation provides information about place of articulation. However, the studies did not address the issue of the relative contribution of the auditory and visual sources because the quality of the auditory source was not systematically varied. By evaluating the contribution of the visual source as a function of the quality of the auditory source, we aim to measure the relative importance of the two sources. Using the logic and procedures of functional measurement (Anderson, 1974), a second goal of the present research is to determine how auditory and visual sources of information are evaluated and integrated together in speech perception.

The results of the present experiments are used to test two specific hypotheses concerning the evaluation and integration of auditory and visual information. First, we ask whether the sources of information available to the perceiver are continuous or discrete. That is, does a given source make available some continuous value of information or simply one discrete alternative? Given evidence for continuous features, the second contrast is whether or not the features are evaluated independently of one another. If the two sources are evaluated independently, the feature value of one source is not influenced by the other source. Non-independent sources imply that the value of one source is modified by the other source. The two hypotheses will be formalized as quantitative models and tested against the results of the present experiments. In the first experiment, we assess how the visual articulation of /ba/ or /da/ influences what subjects hear in a perceptual recognition task.

Method

Subjects. Six subjects were recruited from an introductory psychology class. They were given extra course credit for their participation.

Stimuli. Prior to the experiment, a videotape was prepared for presentation of the visual speech information and synchronization of the auditory information. The procedure was as follows. The speaker (D.W.M.) was seated in front of a wood-panel background that was illuminated with ordinary fluorescent fixtures in the ceiling. The speaker's head was centered in the video field and filled about two thirds of the frame in the vertical direction. Both the video and audio were recorded on a SONY AV 3650 ½-in. (3-cm) monochrome video tape recorder.

On each trial, the speaker said either /ba/ or /da/, as cued by a video terminal under control of a DEC PDP-11/34a computer. The computer cued the speaker by displaying either a large lowercase b or d on the screen for 500 msec. The cues were presented in 21 blocks of 14 stimuli with 7 b's and 7 d's in each block of 14. Prior to these 294 trials, 10 stimuli were recorded for practice trials for a total of 304 stimuli on the tape. Each trial took about 2,600 msec. The tape was reviewed to ensure that the stimuli were created according to the cues.

The auditory stimuli for the experiment were created as follows. Four each of the speaker's /ba/s and /da/s were digitized at 10,000 samples per sec. Each consonant–vowel (CV) segment was about 400 msec in length. These samples were analyzed using linear prediction with cepstrally based pitch estimation. In this analysis, a 12.8-msec window was used with 12 linear-prediction parameters, .98 preemphasis and a hamming window moving on 5.0-msec centers. This analysis was followed by smoothed formant tracking and yielded estimates for the center frequency of the first 4 formants as well as the overall amplitude and pitch of the speaker for each 5-msec segment of the CV segment. The analysis values from a typical /ba/ syllable were converted into a set of parameters for driving a software-formant serial-resonator speech synthesizer (Klatt, 1980). By altering the parametric information regarding the first 80 msec of the CV, a set of seven 400-msec CVs covering the range from /ba/ to /da/ was created. During the first 80 msec, Formant 1 (F1) went from 300 Hz to 700 Hz following a negatively accelerated path. The F2 followed a negatively accelerated path to 1199 Hz from one of seven values equally spaced between 1100 and 2000 Hz from most /ba/-like to most /da/-like, respectively. The F3 followed a linear transition to 2729 Hz from one of seven values equally spaced between 2400 and 3200 Hz. The post transition F1, F2, and F3 values followed the original /ba/ CV. During the entire CV the F4, F0 and overall amplitude of the CV were equated with the values of the original sound. The F5 was fixed at 3850 Hz. The resulting stimuli, quantized at a 10,000 sample-per-sec rate were stored on the computer disk for later presentation.

Procedure. All experimental events were controlled by

a DEC PDP-11/34a computer. The synthetic speech was filtered 20-4900 Hz (KHRON-HITE 3500R) and presented via the speakers of the SONY 11-in. (.3-m) monitors, which was also used for presentation of the visual portion of the videotape. The auditory information was presented at a comfortable listening level (71 dB-A). Three subjects could be tested simultaneously in individual sound-attenuated rooms. These rooms were each illuminated by two 60-W incandescent bulbs in a frosted-glass ceiling fixture.

On each trial of the experiment, 1 of the 7 auditory stimuli on the continuum from /ba/ to /da/ was paired with 1 of the 2 possible visual stimuli, /ba/ or /da/. The stimuli were presented in 21 blocks of the 14 possible combinations, sampled randomly without replacement. Preceding the experimental trials, a partial block of 10 practice trials was presented for a total of 304 stimuli.

The presentation of the synthetic speech was synchronized with the original audio track on the videotape and gave the strong illusion that the synthetic speech was coming from the mouth of the speaker. To accomplish this synchronization, the audio signal was monitored by a schmidt trigger circuit. When the original audio channel on the videotape exceeded a preset threshold, one of the 400-msec CV syllables was played. The subjects had 2,400 msec to make their response by pressing either the "B" or "D" button on a keyboard positioned to the side of the TV monitor.

Each subject was instructed "to watch a speaker and listen to what is spoken" and "to indicate whether you heard the sound /ba/ or the sound /da/."

Results

The proportion of /da/ responses as a function of the 7 levels along the auditory speech continuum is shown in the left panel of Figure 1, with the visual /ba/ or /da/ articulation as the curve parameters. The average proportion of /da/ responses increased significantly as a function of the level of the auditory stimulus, from .342 for the most /ba/-like to .779 for the most /da/-like, $F(6, 30) = 8.18$, $p < .001$. There was also a large effect on the proportion of /da/ responses as a function of the visual stimulus, with .214 /da/ responses for the /ba/ visual stimulus and .880 for the /da/ visual stimulus, $F(1, 5) = 43.84$, $p < .001$. The interaction of these two variables was also significant, $F(6, 30) = 5.93$, $p < .001$, because the effect of the visual variable is smaller at the less ambiguous regions of the auditory dimension.

Quantitative tests of hypotheses. The first question is whether the auditory and visual sources are continuous or discrete. The hypothesis of discrete sources implies that separate discrete phonetic decisions are made to the auditory and visual sources. In the present experiment, separate /ba/ and /da/ decisions would be made to the auditory and visual sources and the identification response would be based on these separate decisions. There are four possible outcomes for a particular combination of auditory and visual information according to the discrete model. The auditory and visual decisions could be /da/-/da/, /da/-/ba/, /ba/-/da/, or /ba/-/ba/. If the two decisions to a given speech event agree, the identification response is assumed to follow

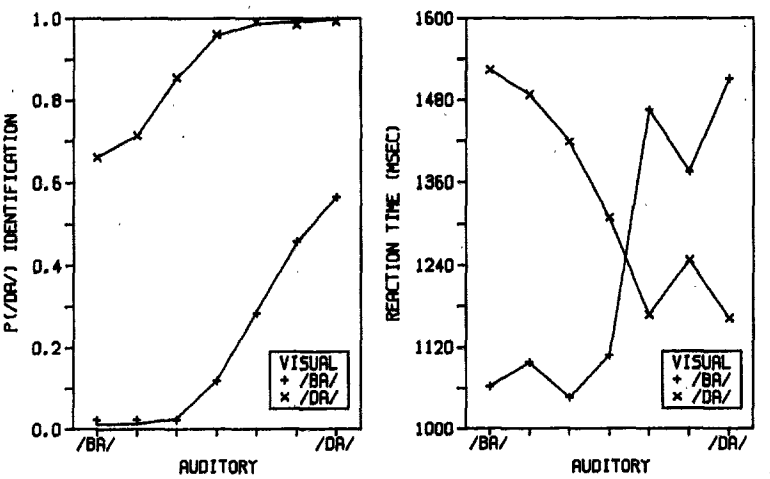

Figure 1. Observed (points) and predicted (lines) proportion of /da/ identifications for independence model (left) and reaction times (right) as a function of the auditory and visual levels of the speech event in Experiment 1.

either source. If the two decisions disagree, it is reasonable to assume that the subject will respond with the decision of the auditory source on some proportion, p, of the trials, and respond with the decision of the visual source on the remainder $(1 - p)$ of the trials. In this conceptualization, p reflects the relative dominance of the auditory source.

The probability of a D identification response, $P(D)$, given a particular auditory/visual speech event, A_iV_j, would be

$$P(D|A_iV_j) = (1 * a_iv_j) + [p * a_i(1 - v_j)] + [(1 - p) * (1 - a_i)v_j] + [0 * (1 - a_i)(1 - v_j)],$$

where i and j index the levels of the auditory and visual stimuli, respectively. The a_i value represents the probability of a D decision given the auditory level i, and v_j is the probability of a D 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 D identification response given that outcome. In this model, each unique level of the auditory stimulus would require a unique parameter for a_i and analogously for v_j. A parameter is also required for p, the relative dominance of the auditory source, which would be fixed across all conditions. This gives a total of 10 parameters for the 14 independent conditions of Experiment 1.

The second 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 nonindependence assumption claims that the visual and auditory sources are not evaluated independently but that the value of one influences the value of the other. The concept of nonindependence is most clearly illustrated by the conclusion reached by Carden, Levitt, Jusczyk, and Walley (1981), who used modified speech stimuli that could be labeled as either *stops* [b, d] or *fricatives* [f, θ]. For a given stimulus, identification of place of articulation was shown to be dependent on the identification of manner. A given stimulus might be identified as [b] (labial) if it was identified as a stop and nonlabial [θ] if it was identified as a fricative. Carden et al. proposed that the perceived manner of articulation affects the interpretation of the formant-transition cue. Accordingly, the cue value of the formant transition is dependent on the cue value (and thus the perception) of manner of articulation. Similar conclusions have been reached by Eimas, Tartter, Miller, and Keuthen (1978) and Miller (1977). When applied to the integration of auditory and visual information, the nonindependence concept would imply that the cue value of one source of information would constrain and/or be constrained by the other source.

It is very difficult to formalize and test the nonindependence model unless a particular type of dependence 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 nonindependence model would require as many parameters as there are independent conditions. Even with this limitation, we can test the nonindependence hypothesis in two ways. Given the nonindependence hypothesis that the evaluation of one feature is dependent on the value of another, any model assuming independent contributions of each feature must fail. To the extent that the independence model gives an adequate description of the results, we have evidence against the nonindependence hypothesis. A second test of nonindependence is to assume a particular form of dependence. Massaro and Cohen (1977) found a linear dependence between voicing amplitude and duration of the fricative for members of a fricative–vowel continuum going from /si/ to /zi/. Thus, it is reasonable to test this form of nonindependence between the auditory and visual sources of information. Given a good description of the independence model and a poor description of the nonindependence model, we have evidence against the hypothesis of nonindependence until an equally parsimonious form of dependence is developed and tested favorably against the experimental results.

The fuzzy logical model (Massaro & Oden, 1980b) allows a quantitative formulation of independence and nonindependence views and, therefore, a quantitative test between them. According to the fuzzy logical model of perception (FLMP), well-learned patterns are recognized in accordance with a general al-

gorithm regardless of the modality or particular nature of the patterns (Massaro, 1979; Oden & Massaro, 1978). The FLMP postulates three operations in perceptual recognition: feature evaluation, prototype matching, and pattern classification. Continuously valued features are evaluated, integrated, and matched against prototype descriptions in memory, and an identification decision is made on the basis of the relative goodness of the match of the stimulus information with the relevant prototype descriptions. DeMori and his associates (e.g., DeMori & Laface, 1980) used fuzzy logical techniques to classify speech segments in continuous speech. This approach has proven to be reasonably successful in the machine recognition of speech, and it encourages the use of similar techniques to describe human performance. Before proceeding further, we briefly discuss the concept of fuzzy logic.

It is often the case that propositions are neither entirely true nor false but rather take on continuous truth values. For example, we might say that a president is doing a "very bad" job or that a meal is not "too spicy." Ordinary logical quantification would require that the president be performing well or not and that the meal be either spicy or not. Fuzzy logic theory (Goguen, 1969; Zadeh, 1965), on the other hand, allows us to represent the continuous nature of things. In fuzzy logic we can construct a membership function, for example, short (x), which is true to the extent that item x is a member of the set short. It should be noted that fuzzy truth is different from probability. If we say that a whale is a fish to a degree of .5 that does not mean that there is a .5 probability that a particular whale is a fish. Rather, it is true that the whale is a fish to a degree of .5.

An important part of fuzzy logic theory concerns the realization of the standard logical operations of conjunction, negation, and disjunction. The range of truth values, $t(x)$, in fuzzy logic goes from 0 for perfectly false to 1 for perfectly true. Thus, a reasonable definition for negation is the additive complement

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

where $t(\sim x)$ is the truth of not x. Goguen (1969) has suggested two possibilities for the conjunction of two events, a and b:

$$t(a \wedge b) = t(a) * t(b),$$

and

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

With the help of DeMorgan's Law, we can derive the two corresponding disjunction operations:

$$t(aVb) = t(a) + t(b) - t(a) * t(b)$$

and

$$t(aVb) = \max[t(a), t(b)].$$

We should note that either of these definitions reduce to ordinary logic if we restrict truth values to 0 and 1.

Oden (1979b) reviewed the success of fuzzy logic in modeling certain classes of human performance. In a typical experiment, Oden (1977) investigated which set of definitions of fuzzy logical conjunction and disjunction best fit judgments about logical combinations of pairs of statements about class membership functions (e.g., a bat is a bird and a refrigerator is a piece of furniture). A large set of such compounds was created by factorially varying the truthfulness of each component. The data from the experiment were best explained by the multiplicative-based rules. That does not mean that humans actually carry out the process of multiplication, just that multiplication closely represents the processes involved in information integration (see Lopes, 1981).

According to the FLMP, recognition is carried out in three stages. The first stage is feature evaluation, during which the stimulus is transduced by the sensory systems and various perceptual features are derived. The features are assumed to be continuous rather than discrete. The outcome of featural evaluation is a truth value, $t(x)$, representing the degree to which each relevant feature is present in the speech stimulus.

The second stage of recognition is prototype matching, which involves the integration of the features. During this stage the featural information is compared with perceptual unit definitions, or prototypes, to determine to what degree each prototype is realized in the speech sound. Prototypes define a perceptual unit in terms of arbitrarily complex fuzzy logical propositions.

The third stage of recognition processing is pattern classification. During this stage, the merit of each potential prototype is evaluated relative to the summed merits of the other potential prototypes. The relative goodness of

a perceptual unit 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. In pandemoniumlike terms we might say that it is not how loud some demon is calling that counts but rather the relative loudness of that demon in the relevant crowd of demons. An important feature of the model is that one cue has its greatest effect when the second is at its most ambiguous level. Thus, the most informative cue has the greatest impact on the judgments.

Given this general model, we can test between the two types of integration of the auditory and visual information in speech perception. According to the independence view, the two cues are evaluated as independent sources of information. According to the nonindependence view, the values of the two cues mutually constrain one another. Both of these views can be formalized within the framework of the FLMP; the difference between the models is reflected in different prototype descriptions of the test alternatives.

For the independence model the prototypes are defined as

/da/: slightly falling F2–F3 and open lips

and

/ba/: rising F2–F3 and closed lips,

where F2–F3 represents the onsets of the second and third formants. It should be noted that the physical features listed in the prototype descriptions are actually sensory correlates of those physical features (Massaro & Oden, 1980b). Each prototype description has independent specifications for the auditory and visual sources of information. Hence the value of one source does not change the value of the other source at the prototype-matching stage. Not included in these prototype descriptions but present in the actual prototype descriptions are other features relevant to distinguishing both the consonants and the vowel /a/. In considering a particular set of prototypes, we need not consider information that is contained in common. Following our multiplicative-based operators, the merit of each of the prototypes would be evaluated according to the product (*) of the features, which are

/da/: slightly falling F2–F3 * open lips

/ba/: rising F2–F3 * closed lips.

Given only two possible levels for each of the two features in the experiment, the use of complementary features will give simpler isomorphic prototypes. That is, we can represent rising F2–F3 as (1 − slightly falling F2–F3) and closed lips as (1 − open lips), thus yielding

/da/: slightly falling F2–F3 * open lips

/ba/: (1 − slightly falling F2–F3)*

(1 − open lips).

If a_i represents the degree to which the auditory stimulus A_i has slightly falling F2–F3 and v_j represents the degree to which the visual stimulus V_j has open lips, the prototype definitions would be

$$\text{/da/: } a_i * v_j$$

$$\text{/ba/: } (1 - a_i) * (1 - v_j).$$

Given these prototype definitions, the probability of a D response given the ith level of the auditory stimulus and the jth level of the visual stimulus, $P(D|A_iV_j)$, would be

$$P(D|A_iV_j) = a_iv_j/[a_iv_j + (1 - a_i)(1 - v_j)],$$

following the pattern classification operation of the FLMP. Given seven levels of A_i and two levels of V_j, the predictions of the model require nine parameters (seven a_i values and two v_j values).

For the nonindependence model, the same two alternatives have prototype descriptions; but in this case the description is simply in terms of the auditory source of information, which is modified by the visual source,

/da/: slightly falling F2–F3 = a_{ij}

/ba/: 1 − (slightly falling F2–F3) = 1 − (a_{ij}),

where a_{ij} is the product of the auditory and visual sources,

$$a_{ij} = a_i * v_j.$$

This nonindependence formalization assumes that the visual source modifies the feature value of the auditory source and is exactly analogous to previous assumptions of nonindependence of features (Carden et al., 1981; Eimas et al., 1978; Miller, 1977). Note that in the nonindependence form the auditory and visual information is combined prior to the prototype matching stage. That is to say, only a single, multiplicatively combined feature is available for prototype matching.

Given these prototype definitions, the nonindependence model makes the following prediction given the pattern classification operation:

$$P(D|A_iV_j) = a_{ij}/[a_{ij} + (1 - a_{ij})] = a_iv_j.$$

Following the logic used for the independence model, the nonindependence model requires the same number (nine) of parameters as does the independence model.

For both the independence and the nonindependence models, a parameter is an output value of some unknown psychophysical transformation of the acoustic information made available to the sensory apparatus. It should be noted that this experiment assesses how the various sources of information are combined rather than how the feature values are determined in relation to the physical stimulus. Thus, we allow one parameter value for each hypothetical output value of each feature in our analysis of theoretical models. The parameter values can be used to determine the relative contribution of each source and to ascertain the psychophysical relationship between the stimulus source and the perceptual consequence.

To determine which model is best, the quantitative predictions of the three models were compared with the observed proportion of a /da/ response for each subject using the program STEPIT (Chandler, 1969). A model is represented to the analysis program STEPIT as a set of prediction equations and a set of unknown parameters. Initially, all parameters are set to .5. By iteratively adjusting the parameters of the model, STEPIT minimizes the squared deviations between the 14 observed and predicted points. Thus, what STEPIT does is to find a set of parameter values, which, when put in the model, come closest to predicting the observed data.

Figure 2 gives the average predicted results of the discrete and nonindependence models for the results of Experiment 1. As can be seen in the figure, both models give a poor description of the observed results. The independence model provided a much better description, as can be seen in Figure 1. Tables 1 and 2 give the root mean square deviation (RMSD) and best-fitting parameters for the three models for each subject and the average of the subject fits. The independence model gave a mean RMSD per point of .015 averaged across the six subjects, whereas the discrete and nonindependence models yielded far inferior RMSDs of .098 and .128, respectively. An analysis of variance (ANOVA) was carried out on the RMSDs, contrasting both the discrete and nonindependence models against the independence model. The independence model gave significantly lower RMSDs compared to both the discrete, $F(1, 5) = 12.58$, $p < .025$, and the nonindependence, $F(1, 5) = 18.7$, $p < .01$, models. Thus, we can reject the discrete

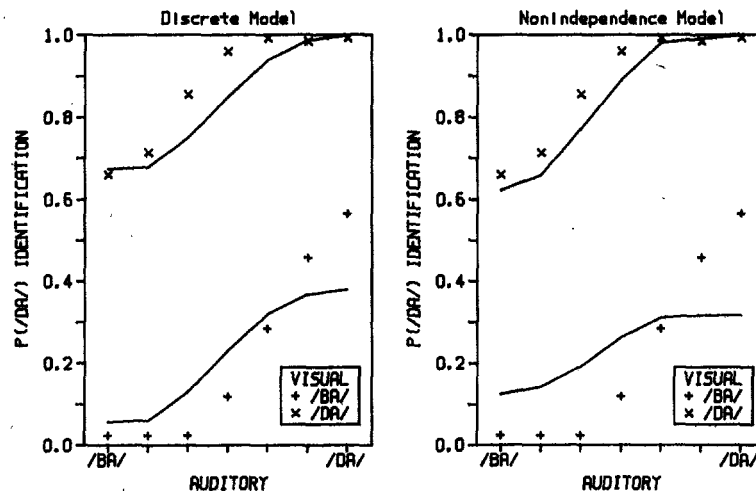

Figure 2. Observed (points) and predicted (lines) proportion of /da/ responses for discrete (left) and nonindependence (right) models of Experiment 1 results.

Table 1
Root Mean Square Deviations (RMSD) Between Observed and Predicted Values and Best-Fitting Parameter Values for Independence FLMP Model of Experiment 1 Data

		Visual		Auditory						
Subject	RMSD	ba	da	ba	2	3	4	5	6	da
1	.021	.53	.97	.05	.05	.04	.36	.78	.88	.94
2	.012	.04	.99	.11	.20	.58	.74	.82	.82	.94
3	.019	.03	.97	.16	.17	.37	.51	.51	.51	.89
4	.007	.09	.95	.00	.02	.26	.72	.95	.99	.99
5	.018	.02	.99	.01	.01	.02	.22	.71	.96	.98
6	.014	.04	.92	.18	.19	.28	.35	.73	.94	.95
M	.015	.12	.96	.08	.10	.26	.48	.75	.85	.95

Note. FLMP = fuzzy logical model of perception.

and nonindependence models in favor of the independence model. This result provides evidence consistent with the hypothesis of continuous and independent featural information.

The advantage of the independence model over the nonindependence model is consistent with the findings of Roberts and Summerfield (1981), who used an audiovisual adaptor consisting of a speech sound /bɛ/ dubbed onto a visual articulation of /gɛ/. This speech event was usually identified as /dɛ/ or some other nonlabial percept. Subjects were given 50 repetitions of this stimulus in order to evaluate their selective adaptation on a /bɛ/ to /dɛ/ auditory test continuum. The adaptation effect was identical in size and direction to that produced by a solitary auditory /bɛ/. This result indicates that the auditory component of the audiovisual stimulus maintained its integrity

Table 2
Root Mean Square Deviations (RMSDs) Between Observed and Predicted Values and Best-Fitting Parameter Values for Discrete and Nonindependence Models of Experiment 1 Data

		Visual		Auditory							
Subject	RMSD	ba	da	ba	2	3	4	5	6	da	*p*
					Discrete model						
1	.140	.33	.99	.12	.08	.00	.67	.99	.99	.99	.54
2	.061	.06	.99	.02	.02	.46	.68	.90	.90	.99	.11
3	.037	.00	.99	.00	.11	.31	.50	.50	.50	.99	.12
4	.182	.00	.99	.00	.00	.44	.63	.91	.99	.99	.76
5	.103	.02	.99	.15	.26	.36	.46	.57	.99	.99	.23
6	.063	.00	.99	.00	.02	.18	.26	.58	.99	.99	.29
M	.098	.07	.99	.05	.08	.29	.53	.74	.90	.99	.34
					Nonindependence model						
1	.187	.71	.99	.48	.46	.41	.81	.99	.99	.99	
2	.073	.12	.99	.89	.93	.99	.99	.99	.99	.99	
3	.043	.05	.99	.84	.85	.94	.95	.99	.95	.99	
4	.194	.68	.99	.00	.17	.60	.77	.97	.99	.99	
5	.129	.16	.99	.83	.84	.88	.97	.98	.99	.99	
6	.114	.17	.99	.67	.69	.78	.83	.94	.98	.99	
M	.123	.31	.99	.62	.65	.77	.89	.98	.99	.99	

Note. *p* = probability of responding with the decision of the auditory source.

and had the appropriate selective adaptation effect. This result provides strong evidence that the sources of auditory and visual information are independent at some stage of perceptual processing before perceptual recognition takes place.

An analysis was also carried out on the average reaction times (RTs) of the identification judgments. The data were pooled across /ba/ and /da/ responses before the average RTs were computed. The only significant effect was a strong interaction of the visual and auditory information, $F(6, 30) = 23.51$, $p < .001$. The right panel of Figure 1 shows that RTs given a visual /ba/ articulation increased with changes from /ba/ to /da/ along the auditory continuum. An analogous result occurred for the visual /da/ in that RTs decreased with auditory changes from /ba/ to /da/.

There are two obvious explanations for the RT results. First, subjects were slower to respond to the extent that the auditory and visual information gave conflicting information about the speech event. Second, identification time is positively related to the ambiguity of the speech event. The idea of conflicting auditory and visual information is not necessarily equivalent to the ambiguity of a speech event. That is, a given probability of a /da/ identification could result from a variety of levels of conflicting information. A very /da/-like sound paired with a good /ba/ articulation could give the same proportion of /da/ identifications as a very ambiguous sound paired with a relatively ambiguous /da/ articulation. The first case would have greater conflicting information, and yet both would be represented by the same amount of ambiguity.

These two contrasting explanations appear to be representative of the discrete and continuous models, respectively. If separate discrete decisions are made for the auditory and visual sources, then the degree of auditory and visual conflict should be an important factor. When the same decision is reached along both sources, the RTs should be relatively fast. In contrast, conflicting auditory and visual sources are likely to lead to different decisions and thus longer RTs from the additional time needed to resolve the conflict in favor of one source or the other. The discrete model has no apparent mechanism to account for increases in RT with increases in the ambiguity of the speech event. Because the listener has only discrete information, the speech events cannot be perceived with varying levels of ambiguity. However, it may be the case that the trials on which conflict most often occurs are the most ambiguous.

If the continuous information assumed by the FLMP is available, RTs should increase with increases in ambiguity of the speech event. To the extent that listeners have ambiguous information, it should take longer to decide on one of the discrete response alternatives at the pattern-classification operation. It is likely that listeners establish a criterion value distinguishing the category /ba/ from /da/ and, by necessity, the ambiguous speech events will be closer to the criterion. There is evidence that more time is required to make an identification response to the extent the information about the event is ambiguous and close to the criterion value (Norman & Wickelgren, 1969). On the other hand, the increase in RTs with increases in conflict of the visual and auditory sources is not necessarily consistent with the FLMP. With continuous sources of information and a multiplicative integration, the time for feature evaluation and prototype matching should not necessarily change with the degree of visual and auditory conflict.

Given the contrasting explanations, the RTs provide a new test between the discrete model and the continuous FLMP. To quantify the idea of conflicting information, we used the parameters corresponding to the probability values for the auditory and visual sources of information given by the discrete model. The degree of conflict was taken to be the proportion of time that auditory and visual information gave different decisions according to the discrete model:

$$C_{ij} = a_i(1 - v_j) + (1 - a_i)v_j,$$

where C_{ij} is the conflict given by the ith level of the auditory continuum and the jth level of visual articulation.

Ambiguity A_{ij} can be defined in terms of the degree to which the likelihood of a /da/ identification approximates .5;

$$A_{ij} = |P(D|A_iV_j) - .5|,$$

where $P(D|A_iV_j)$ is the probability of a /da/

identification given the ith auditory level and the jth visual level. In this conceptualization, the lowest ambiguity value would be .5, whereas the highest would be 0.

To evaluate the conflict and ambiguity explanations, a multiple regression analysis of ambiguity and conflict was carried out on the 84 RTs (14 conditions × 6 subjects). Each subject contributed 14 C_{ij}, 14 A_{ij}, and 14 RT values. Although the simple correlations of both ambiguity and conflict with RT were significant, there was a significant −.70 correlation between ambiguity and conflict. Partial correlations indicated that ambiguity accounted for 13% of the variance that could not be accounted for by conflict. At the same time, however, conflict accounted for 11% of the variance unaccounted for by ambiguity. The points in Figure 3 give the RT for identification as a function of the ambiguity for the group of six subjects. The line in Figure 3 gives the linear regression fit for the RTs predicted by ambiguity.

The problem in distinguishing between the ambiguity and conflict accounts of the RT difference is the high correlation between ambiguity and conflict. Experiment 2 extends the conditions of Experiment 1 by including neutral trials in which no articulation is given. In this case, the ambiguity of the auditory source of information will vary without a concomitant variation in the conflict between it and the visual source. Therefore, the neutral condition will be important in determining whether ambiguity or conflict is primarily responsible for RT differences observed in the identification task.

Experiment 2

Experiment 2 was carried out to test how easily the visual contribution could be attenuated by context. To the extent that subjects realized that the auditory and visual sources were not necessarily representing the same speech sound, we might expect that the visual component would be reduced. In Experiment 2, subjects sometimes heard a speech sound without the speaker moving his or her lips. One question was whether the inclusion of these "neutral" trials would decrease the magnitude of the effect of the speaker's articulation on other trials. The neutral trial also provides another level of the independent variable corresponding to the visual source of information and thus enhances the assessment provided by a factorial design. In addition, the neutral trial should decrease the correlation between the

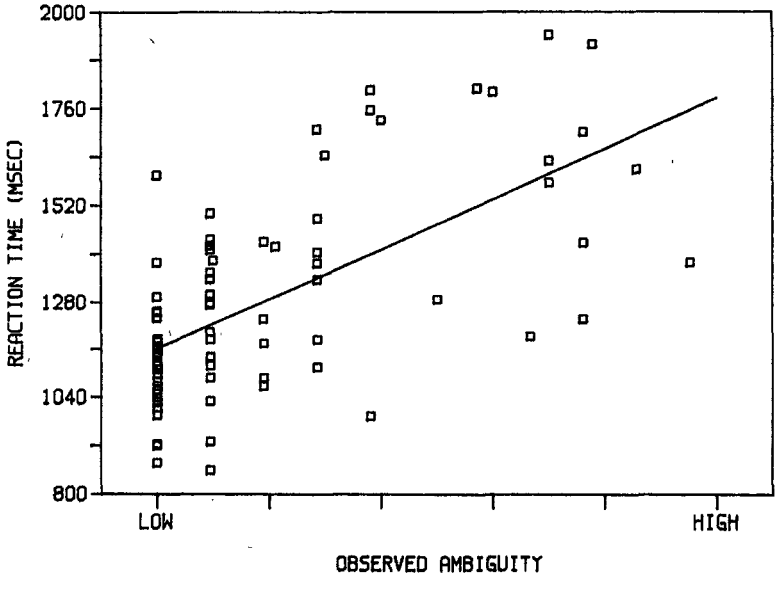

Figure 3. Reaction time in Experiment 1 as a function of ambiguity with best-fitting regression line. (Ambiguity values go from .5 for least ambiguous to 0 for most ambiguous.)

ambiguity and conflict descriptions of the RT differences.

Method

Subjects. Seven subjects were recruited from an introductory psychology class. They were given extra course credit for their participation.

Stimuli. The stimuli were equivalent to those in Experiment 1 except for the following differences. Instead of two possible levels of visual information there were three. The third visual level was neutral, and for this level the speaker kept his or her lips closed, as they were during the intertrial interval. On neutral trials, the speaker was cued to say nothing. For synchronization of the auditory stimulus during the experiment, a computer-controlled tone was recorded on the audio channel of the videotape 400 msec after the onset of the neutral cue. The nine rather than seven auditory stimuli were identical to those in Experiment 1 except that the F1 transition started at 250 Hz; the F2 transition could start at one of nine values equally spaced between 1000 and 2000 Hz; and the F3 transition could start at one of nine equally spaced values between 2200 and 3200 Hz.

Procedure. Procedural details were the same as in Experiment 1 with the following exceptions. Each trial of the experiment consisted of 1 of the 9 auditory stimuli on the continuum from /ba/ to /da/ paired with 1 of the 3 possible visual stimuli, /ba/, neutral, or /da/. There were 11 blocks of the 27 possible speech events, sampled randomly without replacement according to a prearranged order determined at the time of recording. A partial block of 10 practice trials was presented before the 297 experimental trials for a total of 307 trials. In addition, the response interval was lengthened to 2,750 msec. The instructions were identical to those given in Experiment 1 except for warning the subject that:

> On some trials, you will see the speaker say the sound and, on other trials, a sound will be presented but the speaker will not move his mouth. On all types of trials, you are simply to indicate whether you heard a /ba/ or a /da/.

Results

The points in the left panel of Figure 4 give the proportion of /da/ responses as a function of the auditory and visual levels of the stimuli. There were 12% more /da/ than /ba/ response, $F(1, 6) = 15.43$, $p < .005$. The average proportion of /da/ responses increased significantly as a function of the level of the auditory stimulus, from .108 for the most /ba/-like to .892 for the most /da/-like, $F(8, 48) = 48.03$, $p < .001$. There was also a large effect on the proportion of /da/ responses as a function of the visual stimulus, with the mean proportion of /da/ responses .356 for the /ba/ visual stimulus, .617 for the neutral stimulus, and .709 for the /da/ visual stimulus, $F(2, 12) = 9.55$, $p < .005$. The interaction of these two variables was also significant, $F(16, 96) = 3.62$, $p < .001$, because the effect of the visual variable was attenuated at the end regions of the auditory dimension.

These identification data were fit by the three models described for Experiment 1. The models had 12 parameters for the nine A_i and three V_j levels. The discrete model had an

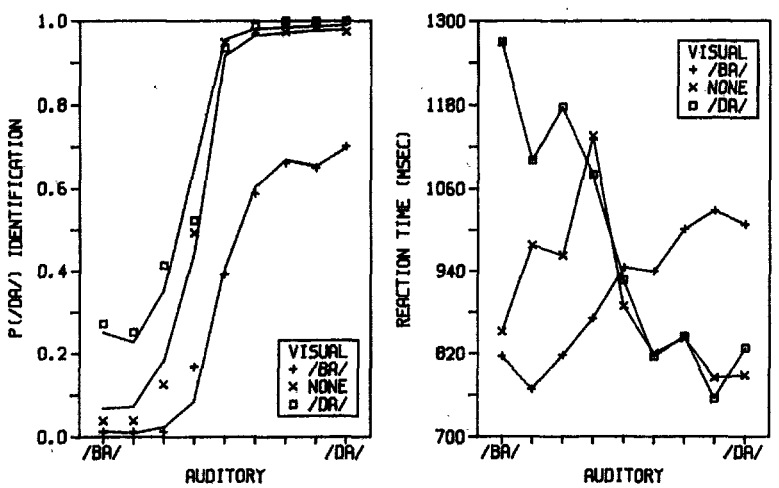

Figure 4. Observed (points) and predicted (lines) proportion of /da/ identifications for independence model (left) and reaction times (right) as a function of the auditory and visual levels of the speech event in Experiment 2.

additional parameter for p, the probability of responding with the decision reached by the auditory source.

The predictions of the discrete and nonindependence models are given in Figure 5. Both models gave a poor description of the results. The RMSDs averaged .128 for the discrete model and .112 for the nonindependence model. The lines in the left panel of Figure 4 gives the average predicted results of the independence FLMP. The model provided a good description of the observed results, giving an average RMSD per point of .058 averaged across the seven subjects. Tables 3 and 4 give the RMSD and best-fitting parameters for each subject and the average of the subject fits.

The description given by the independence model was significantly better than the description given by both the discrete, $F(1, 6) = 13.86$, $p < .01$, and the nonindependence, $F(1, 6) = 9.5$, $p < .025$, models. We conclude that the independence FLMP gives a reasonably good description of the results even when the experimental paradigm is expanded to include trials on which no visual articulation is provided.

The right panel of Figure 4 gives the average RTs of the identification responses. The RTs decreased from 979 to 874 msec as the auditory stimulus went from /ba/ to /da/, $F(8, 48) = 4.79$, $p < .001$. This was probably because the auditory continuum was biased toward the sound /da/. That overall response probability for /da/ was higher than that for /ba/. It is well known that RTs decrease with increases in response probability (Sternberg, 1969; Theios & Walter, 1974).

There was a strong interaction of the visual and auditory information, $F(16, 96) = 7.15$, $p < .001$. The visual /ba/ and visual /da/ functions were roughly monotonic and similar to those in Experiment 1. The RTs to neutral trials increased toward the middle of the auditory continuum. Accordingly, RTs increased with increases in the ambiguity of the auditory stimulus even though there was no change in conflict between the auditory and visual sources. This result seems to resolve the question remaining after Experment 1 in favor of the FLMP and against the interpretation given by the discrete model.

As in Experiment 1, the relationship between RT, ambiguity, and conflict was evaluated using multiple regression. Conflict and ambiguity were defined as in Experiment 1. In contrast to Experiment 1, there was a smaller correlation, −.45, between ambiguity and conflict. The reduction of this correlation is due to the inclusion of the neutral condition in Experiment 2, which allows changes in ambiguity without a concomitant change in conflict. However, this relationship was still significant, $t(187) = 6.88$, as were the simple correlations of both ambiguity ($t = 10.09$) and

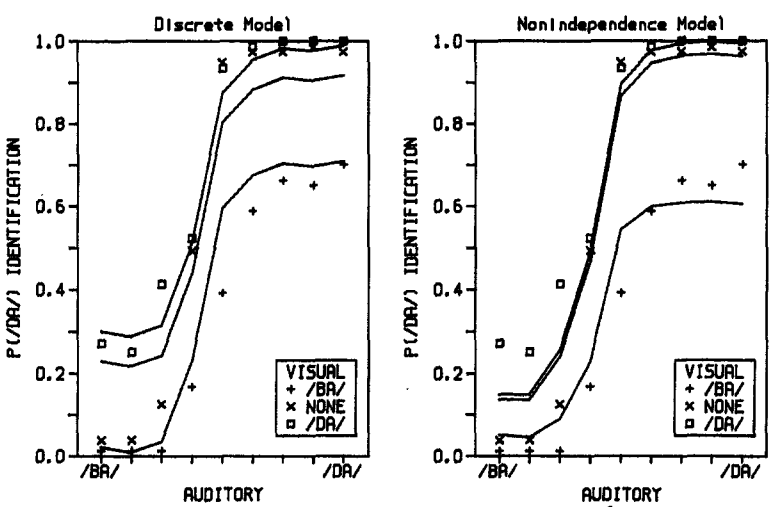

Figure 5. Observed (points) and predicted (lines) proportion of /da/ responses for discrete (left) and nonindependence (right) models of Experiment 2 results.

Table 3
Root Mean Square Deviations (RMSDs) Between Observed and Predicted Values and Best-Fitting Parameter Values for Independence Model of Experiment 2 Data

		Visual			Auditory								
Subject	RMSD	ba	nu	da	ba	2	3	4	5	6	7	8	da
1	.084	.24	.76	.79	.00	.01	.08	.54	.80	.76	.76	.81	.90
2	.086	.17	.69	.90	.06	.14	.16	.47	.64	.74	.81	.86	.89
3	.023	.00	.86	.99	.03	.01	.07	.42	.98	.99	.98	.98	.99
4	.024	.09	.34	.84	.04	.00	.00	.18	.98	.99	.99	.99	.99
5	.089	.44	.65	.81	.10	.04	.11	.18	.83	.99	.99	.99	.92
6	.025	.02	.84	.98	.00	.00	.00	.02	.96	.99	.99	.99	.99
7	.078	.06	.86	.87	.00	.01	.06	.08	.48	.90	.98	.93	.95
M	.058	.14	.71	.88	.03	.03	.06	.27	.80	.91	.93	.93	.95

conflict ($t = 5.76$) with RT. However, partial correlations indicated that ambiguity accounted for 26% of the variance that could not be accounted for by conflict, whereas conflict accounted for only 2.8% of the variance unaccounted for by ambiguity. This analysis supports ambiguity over conflict. The points in Figure 6 give the RT as a function of ambiguity, and the line gives the best linear regression fit for ambiguity.

Given the good description of the independence FLMP, the parameter values for the visual articulation can be taken as dependent measures of the contribution of the visual articulation. The average parameter values for visual v_i were .12 and .96 for the /ba/ and

Table 4
Root Mean Square Deviations (RMSDs) Between Observed and Predicted Values and Best-Fitting Parameter Values for Discrete and Nonindependence Models of Experiment 2

		Visual			Auditory									
Subject	RMSD	ba	nu	da	ba	2	3	4	5	6	7	8	da	*p*
							Discrete model							
1	.129	.00	.88	.97	.00	.00	.05	.60	.88	.88	.88	.92	.99	.77
2	.118	.00	.70	.99	.00	.04	.10	.54	.71	.94	.98	.99	.99	.54
3	.188	.00	.73	.99	.00	.00	.00	.57	.99	.99	.99	.99	.99	.21
4	.079	.00	.20	.99	.05	.00	.00	.24	.96	.99	.99	.99	.99	.96
5	.105	.20	.52	.99	.09	.02	.12	.30	.92	.99	.99	.99	.99	.88
6	.114	.00	.75	.99	.00	.00	.00	.15	.80	.99	.99	.99	.99	.92
7	.166	.00	.93	.99	.00	.00	.00	.14	.55	.76	.99	.88	.92	.73
M	.128	.03	.67	.99	.02	.01	.04	.36	.83	.94	.98	.97	.97	.72
							Nonindependence model							
1	.071	.56	.99	.99	.00	.04	.22	.70	.96	.97	.97	.99	.99	
2	.127	.42	.91	.99	.25	.41	.46	.70	.86	.99	.99	.99	.99	
3	.152	.03	.89	.99	.55	.45	.69	.91	.99	.99	.99	.99	.99	
4	.083	.97	.99	.99	.06	.00	.00	.35	.95	.98	.99	.99	.99	
5	.115	.95	.98	.99	.16	.09	.19	.34	.90	.99	.99	.99	.97	
6	.115	.89	.99	.99	.03	.00	.00	.20	.83	.99	.99	.99	.99	
7	.119	.45	.99	.99	.00	.04	.25	.35	.78	.90	.99	.99	.99	
M	.112	.61	.97	.99	.15	.15	.26	.49	.90	.98	.99	.99	.99	

Note. p = probability of responding with the decision of the auditory source.

/da/ articulations in the first experiment, whereas these values were .14 and .88 in the second. The more extreme values in Experiment 1 reflect a larger contribution of the visual source relative to that found in Experiment 2. Thus, the neutral trials produced some attenuation, although the contribution of the visual source remained very strong—even embedded—with neutral trials. The difference between the two experiments can be seen by the greater convergence of the visual /ba/ and /da/ curves in Experiment 2 than in Experiment 1 (cf. Figures 1 and 4).

Experiment 3

One potential limitation of the identification task used in Experiments 1 and 2 is the use of just two response categories. Although this procedure is used in most studies of identification of synthesized speech continua, the use of conflicting auditory and visual cues might lead to a wider variety of percepts. These percepts would go unnoticed by the experimenter because subjects had to respond with one of the categories /ba/ or /da/. Accordingly, the results supporting the integration of independent auditory and visual cues might be limited to those situations in which subjects are required to identify the sounds as one of the two alternatives. To assess the role of the available response alternatives and to provide a more comprehensive analysis of the perceptual recognition, Experiment 3 replicated Experiment 2 while providing a wider range of response alternatives. The response alternatives were determined on the basis of pilot work. Subjects identified the test items as the voiced syllables /ba/, /da/, /bda/, /tha/, /dba/, /va/, /ga/, or as "other." The instructions were identical to those used in Experiment 2, except for changes related to the response alternatives. The identification responses were analyzed and tested against the predictions of the FLMP.

Method

Subjects. Eight subjects were selected from the same pool used in the previous experiments. Two other subjects were tested but their data were not analyzed because one was a French native speaker and the other said he responded randomly.

Procedure. The subjects were tested on the same videotape and synthesized speech stimuli used in Experiment 2. The instructions were also identical except for describing the eight response alternatives:

> You will do this (indicate what you heard) by pressing one of eight buttons on the keyboards in front of you. There are buttons for the sounds ba, da, tha, bda, dba, ga, va and one for any other sound you might hear.

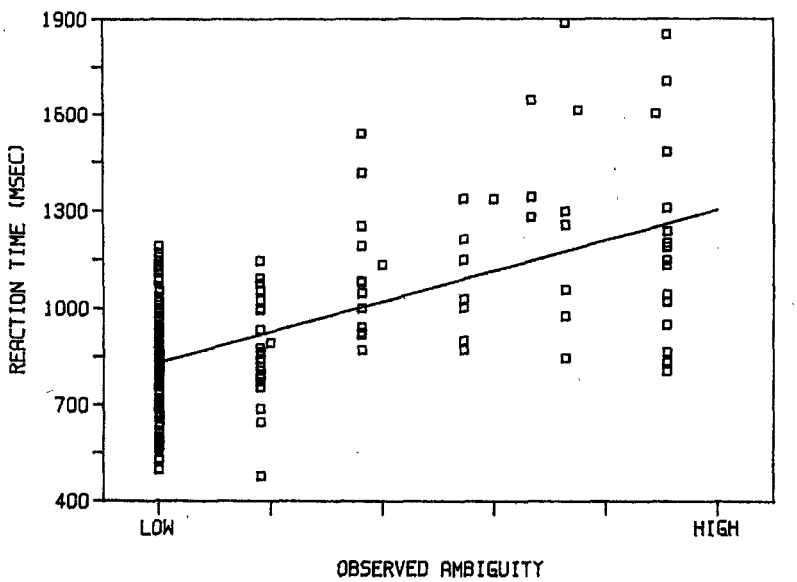

Figure 6. Reaction time in Experiment 2 as a function of ambiguity with best-fitting regression line. (Ambiguity values go from .5 for least ambiguous to 0 for most ambiguous.)

Results

The points in Figure 7 give the probability of a response as a function of the auditory and visual variables. Only the responses /ba/, /da/, /bda/, and /tha/ are given in the figure because these four alternatives accounted for 93% of the identification responses. No other alternatives gave more than 3.5% responses. An ANOVA revealed significant effects of response, $F(7, 49) = 30.34$, $p < .001$, response as a function of the auditory level, $F(56, 392) = 51.87$, $p < .001$, response as a function of the visual level, $F(14, 98) = 52.72$, $p < .001$, and the three-way interaction of response, auditory level, and visual level, $F(112, 784) = 16.83$, $p < .001$.

The results are highly informative and reveal that the percepts are not limited to just /ba/ or /da/. The percepts /bda/ and /tha/ are viable alternatives in this situation. These results are consistent with and extend the findings of MacDonald and McGurk (1978) and Summerfield (1979). One question is to what extent these results qualify the conclusions reached earlier. We see that although the present study adds significant new results, it in no way qualifies the interpretation of the results of the previous experiments.

The present results reveal more clearly an apparent asymmetry in the contribution of visual articulatory information. Relative to the neutral case of no articulation, a visual /ba/ increases the likelihood of a /ba/ response and decreases the likelihood of a /da/ response. In contrast, a visual /da/ does not increase the likelihood of a /da/ response, although it does decrease the likelihood of a /ba/ response. One explanation for this asymmetry is simply that the neutral articulation is not neutral but is more like a visual /da/ than a visual /ba/. A speaker can say /da/ without opening his or her mouth but cannot do the same for /ba/.

The percept /bda/ occurs to the extent that a visual /ba/ is paired with an auditory /da/. Given a visual /ba/, the response /bda/ increased from 0 to 70% with changes from an auditory /ba/ to an auditory /da/. The symmetrical situation did not occur; subjects did not tend to hear /dba/ when a visual /da/ was paired with an auditory /ba/. Rather, subjects tended to hear /tha/ in this case. As can be seen in Figure 7, the function was not monotonic and reached a maximum at the fourth level along the auditory continuum. The neutral articulation condition clarifies the result, indicating that the auditory source provided fairly good support for the alternative /tha/, especially for levels 3, 4, and 5. Spectrographic analysis of the speaker's syllables indicated that the second and third formants for the fourth level along the /ba/ to /da/ continuum were more like those for /tha/ than like those for /ba/ or /da/.

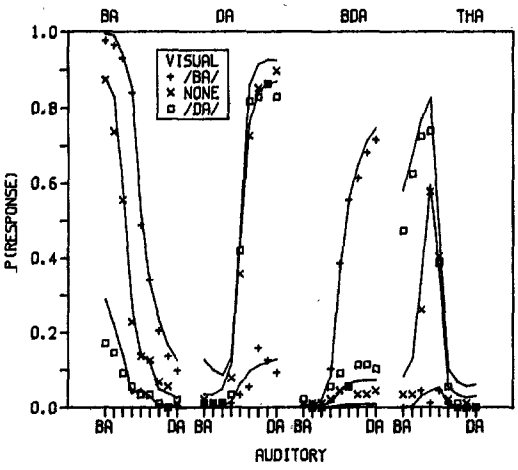

Figure 7. Observed (points) and predicted (lines) proportion of /ba/, /da/, /bda/, and /tha/ responses for the fuzzy logical model of perception of Experiment 3 results.

The question arises as to why the subjects hear a labial–alveolar sequence /bda/ rather than /dba/ when a labial visual articulation is paired with an alveolar /da/ auditory stimulus. This result is reliable and has been reported earlier by Summerfield (1979) and MacDonald and McGurk (1978). In addition, MacDonald and McGurk found that this type of labial–alveolar combination also occurs for voiceless stops and nasals. That is, subjects hear /mna/ rather than /nma/ and /pta/ rather than /tpa/. It is possible that the information about visual /b/ but not visual /d/ is processed slightly before the auditory information for the syllable. Hence, subjects will sometimes hear /bda/ because the information is more consistent with this sequence of consonants than with another sequence or a single consonant.

Fuzzy logical model. The FLMP can be applied to the results by defining prototypes for the four prevalent responses. The alter-

natives /ba/ and /da/ are defined as in the previous experiments. We also reasoned in the preceeding paragraph that /bda/ might be defined as the conjunction of the auditory features for /da/ and the visual features for /ba/;

/bda/: slightly falling F2–F3 * (1 − open lips).

For the alternative /tha/, the auditory and visual sources can be represented as

/tha/: relatively flat F2–F3 *

tongue between lips.

Given these four response alternatives and the 27 stimulus conditions, there are 108 independent data points to be predicted. In addition to the 12 parameters for predicting /ba/ and /da/ responses, 12 additional parameters are needed to predict the /tha/ responses. The alternative /bda/ does not require any additional parameters because it uses the same parameters for /ba/ and /da/. Although we have doubled the number of free parameters relative to Experiment 2, there are four times as many independent data points to be predicted. Accordingly, the results provide a much stronger test of the FLMP, both in terms of the ratio of free parameters to independent data points and in terms of predicting responses with an open-ended set of response alternatives.

The predictions of the model are given in Figure 7. The best-fitting parameter values and the RMSDs are given in Table 5. The FLMP gave an RMSD of .072 averaged across the subjects and an RMSD of .053 for the fit of an average subject. The accuracy of these predictions with unconstrained response alternatives are exactly comparable to those in Experiment 2 when only two responses were permitted.

The FLMP was also applied to the combined results of Experiments 2 and 3. This test of the model is critical because it requires that the same parameter values be used to describe the results of both experiments. Accordingly, only 24 parameters are necessary

Table 5
Root Mean Square Deviations (RMSDs) Between Observed and Predicted Values and Best-Fitting Parameter Values for Independence FLMP Model of Experiment 3 Data

		Visual			Auditory								
Subject	RMSD	ba	nu	da	ba	2	3	4	5	6	7	8	da
1	.073	.13	.93	1.00	.00	.00	.00	.12	.23	.18	.40	.80	.81
		.02	.53	1.00	.02	.00	.28	.70	1.00	.04	.10	.13	.10
2	.038	.08	.98	1.00	.00	.00	.00	.00	.30	.89	.95	.95	.99
		.02	.76	1.00	.00	.00	.01	.15	1.00	.05	.02	.02	.00
3	.098	.03	.87	1.00	.00	.01	.02	.22	.82	1.00	1.00	.97	.92
		.02	.50	1.00	.03	.11	.27	1.00	.40	.23	.13	.13	.19
4	.077	.24	.96	.99	.01	.00	.00	.06	.37	.81	.90	.86	1.00
		.01	.34	.65	.01	.03	.05	1.00	.49	.05	.06	.03	.10
5	.095	.17	.85	.99	.00	.01	.15	.29	.84	1.00	.95	.84	1.00
		.33	1.00	1.00	.01	.05	.74	1.00	1.00	.23	.06	.01	.00
6	.032	.05	.98	1.00	.00	.00	.00	.00	.24	.31	.69	.93	.98
		.12	.15	.04	.02	.02	.04	.11	1.00	.04	.00	.00	.00
7	.068	.31	.93	.99	.00	.00	.00	.00	.01	.03	.15	.17	.22
		.00	.00	.06	.04	.08	.15	.66	.21	.12	.27	.33	.25
8	.094	.25	.87	.96	.01	.02	.03	.11	.65	.98	.93	1.00	.99
		.01	.52	.75	.03	.04	.08	1.00	.37	.00	.03	.06	.12
M	.072	.16	.92	.99	.00	.01	.03	.10	.43	.65	.74	.81	.86
		.07	.47	.69	.02	.04	.20	.70	.68	.09	.08	.09	.09
F_M[a]	.053	.16	.88	.99	.01	.01	.01	.07	.41	.60	.74	.81	.86
		.06	.30	.46	.05	.08	.20	.87	1.00	.15	.10	.09	.10

Note. The first line for each subject gives the /da /-ness parameters and the second line gives the /tha /-ness parameters. FLMP = fuzzy logical model of perception.
[a] F_M = fit of the mean subject.

to predict the 135 independent data points across the two experiments. The model was fit to the average subject data from each of the two experiments and gave an RMSD of .063. The corresponding RMSDs were .045 and .053 for the separate fits of Experiments 2 and 3, respectively. Given the different subjects in the two experiments, the model did about as well as can be expected. Thus, the independence FLMP adequately captures both the integration of auditory and visual sources of information and the identification responses with a varying number of response alternatives.

Discussion

The present experiments have used the methodology of functional measurement (Anderson, 1974) and the FLMP to study the evaluation and integration of auditory and visual sources of information in speech perception. The experiments revealed a strong contribution of visual articulation even though subjects were explicitly instructed to base their decision on what they heard. These results agree with previous findings (Dodd, 1977; MacDonald & McGurk, 1978; Summerfield, 1979) reviewed in the introduction of this article. More recently, Easton and Basala (1982) found no influence of watching the articulation in the auditory perception of CVC syllables and two-syllable words. They explain the discrepancy in terms of using words in their experiment compared to CV syllables in the MacDonald and McGurk study. Although Dodd (1977) also used words, the test items were presented in white noise, which degraded the auditory source of information. It remains a possibility that the visual source of information contributes very little if the test message has words and is unambiguous. Visual contributions might be limited to the perception of sublexical segments or ambiguous or degraded speech sounds.

MacDonald and McGurk (1978) proposed the hypothesis that when viewing a speaker, the manner of articulation is detected by the ear and the place of articulation is detected by the eye. Accordingly, the visual information should dominate the auditory information along the /ba/ to /da/ continuum. If both auditory and visual information contribute to perception of place information, however, identification will be influenced by both the auditory and visual sources. Clearly, the results of the present experiments support the latter rather than the former hypothesis. More important, the contribution of the visual source has the largest impact when the auditory source is relatively ambiguous. The FLMP also provides an explanation of the pattern of errors found in previous studies of audiovisual speech perception. Dodd (1977) found a larger contribution of the visual source for labial relative to nonlabial sounds presented in noise. Other evidence indicated that the labial sounds were more ambiguous than the nonlabial sounds. It follows from the FLMP that the less ambiguous feature has the larger impact, and thus the visual source will have a larger effect with the more ambiguous labial sounds. A similar explanation might account for the number and pattern of errors in the MacDonald and McGurk (1978) study.

The results were unanimous in their support of the FLMP. The critical assumption of the FLMP is that the auditory and visual sources are evaluated independently and then integrated together to achieve perceptual recognition. The identification of syllables varying in both auditory and visual information was used to evaluate quantitative formulations of these two points of view. The independence assumption of the FLMP gave a highly accurate description of the results in contrast to the extremely poor description of a discrete and a nonindependence model.

The independence FLMP can also account for the putative nonindependence results found by Carden et al. (1971), Eimas et al. (1977), and Miller (1977). One independence explanation is in terms of modifiers on the features defining the speech prototypes in memory (Massaro & Oden, 1980a). As an example, Carden et al. found that the place boundary between b and d differed from the place boundary for /f/ and /θ/. The nonindependence interpretation was that the evaluation of the place feature was dependent on the perceived manner of articulation. However, the serial processing of features required by this explanation is unlikely. It implies no influence in the opposite direction, whereas a bidirectional influence between voicing and place was found by Massaro and Oden (1980a). The independence interpretation would simply allow

different ideal feature values in the prototypes defining the alternatives. In terms of the place feature, it might be defined as low-rising F2–F3 for /b/ and very low-rising F2–F3 for /f/. With these definitions, an ambiguous F2–F3 transition would produce more /b/ responses when the manner cue supported a stop than /f/ responses when the manner cue supported a fricitive.

The RTs for identification were also evaluated in terms of the FLMP and discrete perception views. When ambiguity and conflict were somewhat uncorrelated in Experiment 2, ambiguity accounted for over nine times as much variance as did conflict. The discrete perception model and, more generally, categorical perception provide no mechanism to account for an increase in RT with increases in the ambiguity of the test stimulus. If a test stimulus is heard as either one alternative or the other regardless of its ambiguity (as assumed by categorical perception), then the time needed for identification should also not change with ambiguity. According to the FLMP, however, the listener has information about the degree to which a speech stimulus represents one alternative or another (Massaro & Cohen, 1983; Oden & Massaro, 1978). The identification task requires a categorical decision to a continuous percept. Accordingly, it is not surprising that the listener would require additional time for this categorical decision to the extent that the continuous information was ambiguous (Norman & Wickelgren, 1969). In terms of a signal-detection framework, RTs would increase with decreases between the observation and the criterion placed along the dimension of judgment.

The FLMP assumes that the operations involved in speech perception are not special and should be explained within a general model of pattern recognition. In contrast, Liberman (1982) and his associates believed in a "special speech mode of perception." That is, "phonetic perception takes account of the common articulatory origin of diverse cues for a speech contrast" (Best, Morrongiello, & Robson, 1981, p. 205). These two views have fundamentally different explanations of categorical perception and the integration of acoustic cues. Categorical perception means that the listener's experience is limited to a discrete phonetic percept even though the speech stimulus might be varied along a continuous auditory dimension. Although it is now well acknowledged that the perception of speech only approximates the predictions of categorical perception, the concept continues to function as an important empirical and theoretical phenomenon in speech research (Aslin & Pisoni, 1980; Kuhl & Miller, 1978; Remez, Cutting, & Studdert-Kennedy, 1980). In contrast to categorical perception, the FLMP assumes that the listener has continuous information representing the degree to which the speech signal represents one speech sound or another. Categorical-perception results might simply represent the listener's decision faced with a set of discrete alternatives in natural speech or in a laboratory task. That is, the listener must map the continuous information into a discrete decision analogous to the processing assumed in the model of signal-detectability theory (Green & Swets, 1966). Categorical-perception results in no way imply that the listener's percept is limited to category information (Hary & Massaro, 1982).

What other evidence can be cited for these different points of view? One question is whether the integration of cues is unique to speech; if it is, this would be strong evidence for the idea that speech is special. However, the integration of multiple cues has been demonstrated in a variety of nonspeech situations. Oden (1979a) and Massaro (1979) found that readers integrate multiple cues for visual letter recognition, and Oden (1981) has found similar results for the perceptual recognition of dishware (cups and bowls). We are also convinced that similar findings would occur in the perception of size, shape, distance, sign language, and music. Of course, it is logically possible that the similar results do not imply similar processes or processing mechanisms across the different situations. At some level, situation-specific processes must be involved; the feature analysis of the cues in speech can not be handled by the same sensory system used for reading printed text. However, the integration of the cues and the perceptual recognition might result from the same process. The same sequence of operations such as those assumed in the FLMP might occur in all pattern-recognition situations, even those in which the experience in one sensory modality is influenced by another.

References

Anderson, N. H. Information integration theory: A brief survey. In D. H. Krantz, R. C. Atkinson, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 2). San Francisco: Freeman, 1974.

Aslin, R. N., & Pisoni, D. B. Some developmental processes in speech perception. In G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Fergusun (Eds.), *Child phonology: Vol. 2. Perception.* New York: Academic Press, 1980.

Best, C. T., Morrongiello, B., & Robson, R. Perceptual equivalence of acoustic cues in speech and nonspeech perception. *Perception & Psychophysics,* 1981, *29,* 191–211.

Carden, G., Levitt, A., Jusczyk, P. W., & Walley, A. Evidence for phonetic processing of cues to place of articulation: Perceived manner affects perceived place. *Perception & Psychophysics,* 1981, *29,* 26–36.

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

DeMori, R., & Laface, P. Use of fuzzy algorithms for phonetic and phonemic labeling of continuous speech. *IEEE Transactions on Pattern Analysis and Machine Recognition,* 1980, *2,* 136–148.

Dodd, B. The role of vision in the perception of speech. *Perception,* 1977, *6,* 31–40.

Easton, R. D., & Basala, M. Perceptual dominance during lipreading. *Perception & Psychophysics,* 1982, *32,* 562–570.

Eimas, P. D., Tartter, V. C., Miller, J. L., & Keuthen, N. J. Asymmetric dependencies in processing phonetic features. *Perception & Psychophysics,* 1978, *23,* 12–20.

Erber, N. P. Auditory, visual and auditory-visual recognition of consonants by children with normal and impaired hearing. *Journal of Speech and Hearing Research,* 1972, *15,* 413–422.

Goguen, J. A. The logic of inexact concepts. *Synthese,* 1969, *19,* 325–373.

Green, D. M., & Swets, J. A. *Signal detection theory and psychophysics.* New York: Wiley, 1966.

Hary, J. M., & Massaro, D. W. Categorical results do not imply categorical perception. *Perception & Psychophysics,* 1982, *32,* 409–418.

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

Kuhl, P. K., & Miller, J. D. Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the Acoustical Society of America,* 1978, *63,* 905–917.

Liberman, A. M. On finding that speech is special. *American Psychologist,* 1982, *37,* 148–167.

Lopes, L. Decision making in the short run. *Journal of Experimental Psychology: Human Learning and Memory,* 1981, *7,* 377–385.

Luce, R. D. *Individual choice behavior.* New York: Wiley, 1959.

MacDonald, J., & McGurk, H. Visual influences on speech perception processes. *Perception & Psychophysics,* 1978, 24, 253–257.

Massaro, D. W. Letter information and orthographic context in word perception. *Journal of Experimental Psychology: Human Perception and Performance,* 1979, *5,* 595–609.

Massaro, D. W., & Cohen, M. M. The contribution of voice-onset time and fundamental frequency as cues to the /zi/-/si/ distinction. *Perception & Psychophysics,* 1977, *22,* 373–382.

Massaro, D. W., & Cohen, M. M. Categorical or continuous speech perception: A new test. *Speech Communication,* 1983, *2,* 15–35.

Massaro, D. W., & Oden, G. C. Evaluation and integration of acoustic features in speech perception. *Journal of the Acoustical Society of America,* 1980, *63,* 81–97. (a)

Massaro, D. W., & Oden, G. C. Speech perception: A framework for research and theory. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 3). New York: Academic Press, 1980. (b)

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

Miller, J. D. Nonindependence of feature processing in initial consonants. *Journal of Speech and Hearing Research,* 1977, *20,* 510–518.

Norman, D. A., & Wickelgren, W. Strength theory of decision rules and latency in retrieval from short term memory. *Journal of Mathematical Psychology,* 1969, *6,* 192–208.

Oden, G. C. Integration of fuzzy logical information. *Journal of Experimental Psychology: Human Perception and Performance,* 1977, *3,* 565–575.

Oden, G. C. A fuzzy logical model of letter identification. *Journal of Experimental Psychology: Human Perception and Performance,* 1979, *5,* 336–352. (a)

Oden, G. C. Fuzzy propositional approach to psycholinguistic problems: An application of fuzzy set theory in cognitive science. In M. M. Gupta, R. K. Ragade, & R. R. Yager (Eds.), *Advances in fuzzy set theory and applications.* Amsterdam: North-Holland, 1979. (b)

Oden, G. C. Fuzzy propositional model of concept structure and use: A case study in object identification. In G. W. Lasker (Ed.), *Applied systems research and cybernetics,* Elmsford, N.Y.: Pergamon Press, 1981.

Oden, G. C., & Massaro, D. W. Integration of featural information in speech perception. *Psychological Review,* 1978, *85,* 172–191.

Remez, R. E., Cutting, J. E., & Studdert-Kennedy, M. Cross-series adaptation using song and string. *Perception & Psychophysics,* 1980, *27,* 524–530.

Roberts, M., & Summerfield, Q. Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory. *Perception & Psychophysics,* 1981, *30,* 309–314.

Sternberg, S. The discovery of processing stages: Extensions of Donder's method. *Acta Psychologica,* 1969, *30,* 276–315.

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

Theios, J., & Walter, D. G. Stimulus and response frequency and sequential effects in memory scanning reaction times. *Journal of Experimental Psychology,* 1974, *102,* 1092–1099.

Welch, R. B., & Warren, D. H. Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin,* 1980, *88,* 638–667.

Zadeh, L. A. Fuzzy sets. *Information and Control,* 1965, *8,* 338–353.

Received January 10, 1983
Revision received April 18, 1983 ■