Perception & Psychophysics
1996, 58 (7), 1047–1065

Perceiving speech from inverted faces

DOMINIC W. MASSARO and MICHAEL M. COHEN
University of California, Santa Cruz, California

We examined whether the orientation of the face influences speech perception in face-to-face communication. Participants identified auditory syllables, visible syllables, and bimodal syllables presented in an expanded factorial design. The syllables were /ba/, /va/, /ða/, or /da/. The auditory syllables were taken from natural speech whereas the visible syllables were produced by computer animation of a realistic talking face. The animated face was presented either as viewed in normal upright orientation or inverted orientation (180° frontal rotation). The central intent of the study was to determine if an inverted view of the face would change the nature of processing bimodal speech or simply influence the information available in visible speech. The results with both the upright and inverted face views were adequately described by the fuzzy logical model of perception (FLMP). The observed differences in the FLMP's parameter values corresponding to the visual information indicate that inverting the view of the face influences the amount of visible information but does not change the nature of the information processing in bimodal speech perception.

One of the impressive characteristics of speech perception is that the information supporting it is relatively immune to a variety of situational variables. We perceive speech under band-limited conditions (e.g., over the telephone), from a a variety of different speakers, and over a large range of speaking rates. Perceivers are also adept at integrating several sources of information. In face-to-face communication, perceivers integrate the visible speech of the talker with the auditory message. This integration has been shown to occur even if the auditory and visual sources are manipulated independently of each other (Massaro, 1987; McGurk & MacDonald, 1976). For example, the auditory syllable /ba/ is integrated with the visible syllable /da/, usually giving the perceptual experience of /va/ or /ða/ (Massaro & Cohen, 1990).

This integration of auditory and visible speech is also surprisingly robust, occurring even when the two sources are not ecologically joined. Auditory speech is integrated with synthetic (animated) visible speech in the same way it is combined with natural visible speech (Massaro & Cohen, 1983, 1990). Furthermore, auditory and visual speech are combined in a natural fashion even when the sex of the voice differs from the sex of the face doing the talking (Green, Kuhl, Meltzoff, & Stevens, 1991). Similarly, differences in the spatial location of the auditory and visual speech do not disrupt the integration process (Fisher, 1991; Massaro, 1992). It is also well known that auditory and visual speech are integrated when the auditory speech is degraded by noise (Massaro, 1987, pp. 40–45). The results of these studies demonstrate that people appear to integrate the auditory and visual sources even in relatively novel situations. Continuing in this line of investigation, we asked whether inverting the view of the face would disrupt bimodal speech perception.

This question is also of interest because, although the face provides the information for both speechreading and face recognition, there is some evidence that certain aspects of facial information are processed differently from those of speech. A left visual field/right hemisphere advantage for the recognition of identity, sex, and expression of faces is fairly well documented for both split-brain (complete commissurotomy) patients and for neurologically normal individuals (Hellige, 1993; Zaidel, 1994). On the other hand, a left hemisphere advantage is usually reported for linguistic domains (Hellige, 1993). There is also some evidence for a dissociation between speechreading and the recognition of facial expression (Campbell, 1992). One patient was unable to recognize familiar faces and facial expressions, but could recognize what phonemes were being articulated in photographs of faces. Another patient had no trouble with facial recognition but could not recognize phonemes in photographs of faces. On the other hand, de Gelder, Vroomen, and van der Heide (1991) found that accuracy of speechreading and face identification were correlated in normal participants (but not in autistic children). It is well known that recognition of faces is disrupted with rotations away from the vertical (Valentine, 1988). This makes apparent the important question of whether rotations would also disrupt speech perception.

In previous research, several investigators have explored the influence of facial rotation on speech recognition. Campbell (1994) reported a significant but weaker McGurk effect with an inverted projection of the face. Jordan and Bevan (in press) found that the McGurk effect decreased to the extent that the face was rotated from

This research was supported, in part, by grants from the Public Health Service (PHS R01 DC 00236), the National Science Foundation (BNS 8812728), and the University of California, Santa Cruz. The authors gratefully acknowledge the helpful comments of Ken Grant, Neil Macmillan, Kaoru Sekiyama, and Quentin Summerfield on an earlier version of this paper. Correspondence should be addressed to D. W. Massaro, University of California, Santa Cruz, Santa Cruz, CA 95064 (e-mail: massaro@fuzzy.ucsc.edu).

the upright. Bertelson, Vroomen, Wiegeraad, and de Gelder (1994) varied both the spatial separation between the auditory and visual speech and whether or not the face was inverted. The tendency to locate the speech at the face of the speaker was not influenced by inverting the face. Consistent with the Fisher (1991) study, the McGurk effect did not depend on the spatial separation between auditory and visual speech, but the McGurk effect did decrease with facial inversion. Green (1994) also found a significantly weaker McGurk effect when an auditory syllable /ba/ was presented with a visible /ga/ presented on a monitor in inverted position. Green (1994, p. 3014) concluded that "inverting the face... impacts on the integration of phonetic information from the auditory and visual modalities." Our study will test whether the smaller influence of visible speech from an inverted face is due to interference with the visual information, or with the integration of the auditory and visual information.

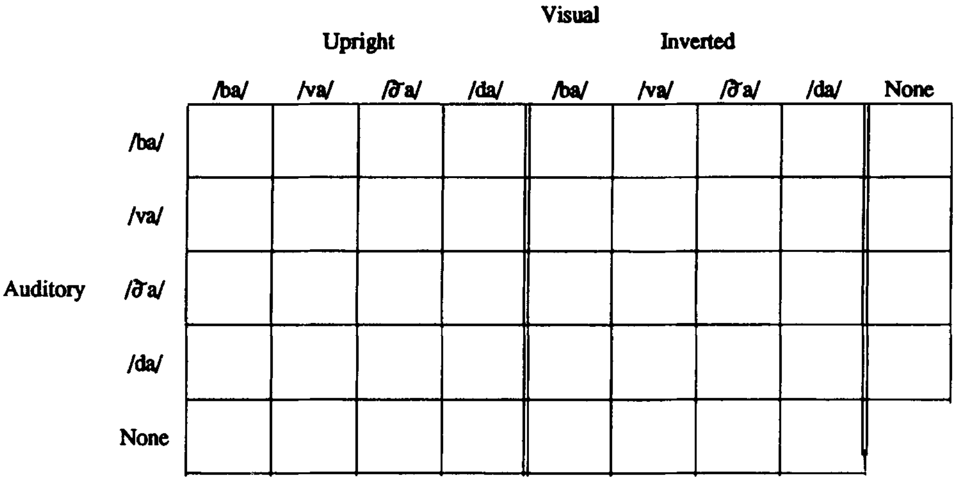
To pursue the question of inverting the front view of the face, we employed an expanded factorial design, illustrated in Figure 1. Each of the two factors corresponds to an independent variable and the different settings of the independent variables are called "levels." The design is called "expanded" because it includes the single-modality conditions as well as the factorial combinations of the two modalities. Four auditory syllables are crossed with four visual syllables. Two versions of the visible syllables are used—one upright and the second inverted. Each syllable from one modality is presented alone or paired with a syllable from the other modality. The design is more powerful than a simple factorial design for testing different models (Massaro & Friedman, 1990). It allows the investigator to address the question of how the identification of a bimodal syllable occurs as a function of the unimodal syllables that compose it. A given model has to predict both the unimodal and bimodal conditions with the same set of parameter values. Thus the design makes it easier to distinguish among different models that make similar predictions (see, e.g., Massaro, 1987, pp. 194–196).

We used the talking head developed in our laboratory (Cohen & Massaro, 1993, 1994) to control and manipulate the visible speech. We believe that it is important to randomize the upright and inverted face conditions within a trial block. This manipulation is currently easy to do with our animated face, but not with a real face. (With the advent of computer-based video systems, however, inversion of video-based stimuli should become feasible.) Exact control of the animated face also allows exact control of the alignment of the visible and audible speech.

We adhere to a falsification and strong-inference strategy of inquiry (Massaro, 1987, 1989a; Platt, 1964). Results are informative only to the degree that they distinguish among alternative theories. Thus, the experimental task, data analysis, and model testing are devised specifically to attempt to reject some theoretical alternatives. A fuzzy logical model of perception (FLMP), an auditory dominance model (ADM), an additive model of perception (AMP), and a prelabeling model (PRLM; Braida, 1991) are formalized and tested against the results. The FLMP has been the most successful model to date (Massaro, 1987, 1989b, 1990; Massaro & Friedman, 1990), but we believe it is important to provide additional tests among the extant models.

FUZZY LOGICAL MODEL OF PERCEPTION

The results from a wide variety of experiments have been described within the framework of the FLMP. Within this framework, speech perception is robust because the perceiver usually evaluates and integrates multiple sources of information to achieve perceptual recognition. The assumptions central to the model are that (1) each source of information is evaluated to give the degree to which



		Visual								
		Upright				Inverted				
		/ba/	/va/	/ða/	/da/	/ba/	/va/	/ða/	/da/	None
Auditory	/ba/									
	/va/									
	/ða/									
	/da/									
	None									

Figure 1. Expanded factorial design with four auditory syllables crossed with four visual syllables.

that source supports the relevant alternatives, (2) the sources of information are evaluated independently of one another, (3) these independent evaluations are then integrated to provide an overall degree of support for each alternative, and (4) perceptual identification follows the relative degree of support among the alternatives.

According to the FLMP, well-learned patterns are recognized in accordance with a general algorithm, regardless of the modality or particular nature of the patterns. Three operations assumed by the model are illustrated in Figure 2. Continuously valued features are evaluated, integrated, and matched against prototype descriptions in memory, and an identification decision is made on the basis of the relative goodness of match of the stimulus information with the relevant prototype descriptions.

In the FLMP, both sources are assumed to provide continuous and independent evidence for each of the prototype alternatives in the bimodal speech perception task. In the present task, /ba/, /va/, /ða/, and /da/ were used as test items. Because we do not know which features are actually used to discriminate among these alternatives, we simply mark each prototype as representing the ideal auditory and visual information for its occurrence. The prototype for /ba/ can be represented as

/ba/ : auditory /ba/ & visual /ba/,

where the two entries correspond to the ideal auditory and visual information for this alternative. The prototype for /va/ would be defined in an analogous fashion,

/va/ : auditory /va/ & visual /va/,

and so on for the other prototypes.

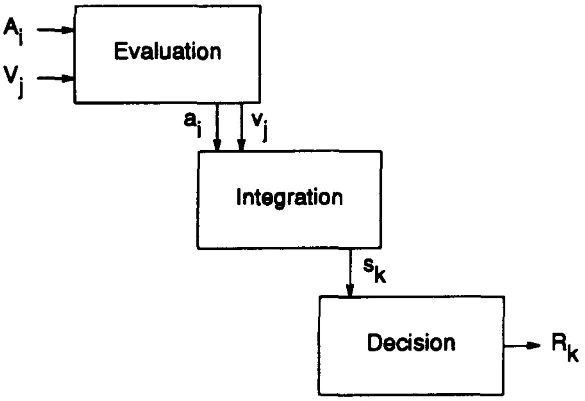


Figure 2. Schematic representation of the three stages involved in perceptual recognition. The three stages are shown to proceed left to right in time to illustrate their necessarily successive but overlapping processing. The sources of information are represented by uppercase letters. Auditory information is represented by A_i and visual information by V_j. The evaluation process transforms these sources of information into psychological values (indicated by lowercase letters a_i and v_j). These sources are then integrated to give an overall degree of support for a given alternative s_k. The decision operation maps this value into some response, R_k, such as a discrete decision or a rating.

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. If the subscripts i and j index the levels of the auditory and visual modalities, respectively, we let a_{Bi} represent the degree to which the auditory stimulus A_i supports the alternative /ba/. Similarly, v_{Bj} represents the degree to which the visual stimulus V_j supports this alternative. These degrees of support are called "feature values." When only a single modality is presented, the outcome of prototype matching is simply assumed to be equal to the outcome of the evaluation of that modality.

When two modalities are presented, the integration of the features defining each prototype is evaluated according to the product of the feature values. This multiplicative combination of feature values has been shown to give a better description of performance than do alternative combinations such as an additive one (Massaro, 1987). The multiplicative combination can also be justified on logical grounds (see Massaro, 1987, p. 193). And perhaps most importantly, when instantiated in the context of the FLMP, the multiplicative combination predicts an optimal integration of the multiple sources of information (Massaro, 1987; Massaro & Friedman, 1990).

Equation 1 gives the outcome of prototype matching for /ba/, assuming a multiplicative contribution of the auditory and visual support:

$$S(\text{/ba/} \mid A_i \text{ and } V_j) = a_{Bi} \times v_{Bj}. \tag{1}$$

where $S(\text{/ba/} \mid A_i \text{ and } V_j)$ is the support for the prototype /ba/ given auditory and visible speech, and again the subscripts i and j index the levels of the auditory and visual modalities. Analogously, if a_{Vi} represents the degree to which the auditory stimulus A_i supports /va/ and v_{Vj} represents the support for /va/ from the visual stimulus, the outcome of prototype matching for /va/ would be

$$S(\text{/va/} \mid A_i \text{ and } V_j) = a_{Vi} \times v_{Vj}, \tag{2}$$

and so on for the other prototypes. These include not only the syllables being presented but also other response alternatives.

The decision operation determines the support for one alternative relative to the sum of the support for each of the relevant alternatives. With only a single source of information, such as the auditory A_i, the probability of a /ba/ response, $P(\text{/ba/} \mid A_i)$, is predicted to be

$$P(\text{/ba/} \mid A_i) = \frac{a_{Bi}}{\sum_k a_{ki}}, \tag{3}$$

where the denominator is equal to the sum of support for all relevant (k) alternatives. With one source of information, support for other alternatives would be computed in an analogous fashion. For example, the probability of a /va/ response given visual V_j, $P(\text{/va/} \mid V_j)$, is predicted to be

$$P(/\mathrm{va}/ \mid V_j) = \frac{v_{V_j}}{\sum_k v_{kj}}. \tag{4}$$

Given two sources of information A_i and V_j, P(/ba/) is predicted to be

$$P(/\mathrm{ba}/ \mid A_i \mathrm{and} V_j) = \frac{a_{Bi} \times v_{Bj}}{\sum_k (a_{ki} \times v_{kj})}. \tag{5}$$

As can be seen in Equations 1 and 2, the absolute support for a given prototype will be less for two sources of information than just one. However, the identification judgment is a function of the relative degree of support, as shown in Equations 3–5. Thus, it is possible that a given identification will be more likely given two sources of information than given just one (Massaro, 1987, chap. 7).

The FLMP is tested against the results by fitting its quantitative predictions to the observed results of each individual participant. Its quantitative predictions are determined by estimating the values of the free parameters using the program STEPIT (Chandler, 1969). In general, a model is represented to this program in terms of a set of prediction equations and a set of unknown parameters. By iteratively adjusting the parameters of the model, the program minimizes the squared deviations between the observed and predicted points. The outcome of the program STEPIT is a set of parameter values that, when put into the model, come closest to predicting the observed results. Thus, STEPIT maximizes the accuracy of the description of a given model. The goodness-of-fit of a model is given by the root mean square deviation (RMSD), the square root of the average squared deviation between the predicted and observed values.

One important assumption of the FLMP is that the auditory source supports each alternative to some degree and analogously for the visual source. Each alternative is defined by ideal values of the auditory and visual information. The degree of support or feature value is given by how much the source matches the corresponding ideal value. For example, the feature value a_{Bi} defines the support that the auditory stimulus A_i gives the alternative /ba/. Because we cannot predict the degree to which a particular auditory or visible syllable supports a response alternative, a free parameter is necessary for each unique syllable for each unique response. However, it should be stressed that an auditory parameter value is forced to remain invariant across variation in the different visual conditions and, analogously, for a visual parameter. In the present experiment, four auditory syllables and four visual syllables are presented in either upright or inverted form. An important (testable) assumption in the fit of the models is that the visual information is assumed to differ in the upright and inverted presentations. That is, for example, the information from a visible /ba/ differs in the upright and inverted presentations. In this case, the FLMP requires four free parameters for the auditory feature values and 2 × 4 = 8 for the visual feature values for each response alternative. There were 12 reliable responses in the present experiment, so the FLMP required a total of 144 free parameters.

AUDITORY DOMINANCE MODEL

A second potential explanation of bimodal speech perception is derived from the hypothesis that an effect of visible speech occurs only when the auditory speech is not completely intelligible (Sekiyama & Tohkura, 1991, 1993; Vroomen, 1992). The hypothesis that auditory intelligibility determines whether or not visible speech will have an effect is difficult to test, primarily because intelligibility is not easily defined. Perfect identification in one test might not mean perfect intelligibility. Even given these limitations in the measure of intelligibility, we formulate one version of an intelligibility model, the auditory dominance model (ADM). The central assumption of the ADM is that the influence of visible speech given a bimodal stimulus is solely a function of whether or not the auditory speech is identified correctly. This model appears to represent extant views of auditory dominance in bimodal speech perception. Sekiyama and Tohkura (1991, p. 1804) concluded that "human beings may depend on eyes in the presence of auditory uncertainty." Similarly, Vroomen (1992) described (but did not defend) the possibility of lipreading as a backup device. In this case, the visual information "is relied on whenever the auditory signal is ambiguous" (Vroomen, 1992, p. 9).

In the current instantiation of the ADM, it is assumed that visible speech has a possible influence *only* when the auditory speech is not identified (Massaro, Cohen, & Smeele, 1995; Massaro, Tsuzaki, Cohen, Gesi, & Heredia, 1993). The probability of a response can be considered to arise from two types of trials given a speech stimulus. Consider first an auditory-alone trial. As shown in the top panel of Figure 3, the auditory speech is identified as one of the response alternatives r or not. When the participant identifies the auditory stimulus as a given alternative r, he or she responds with that alternative. In the case that no identification is made, the participant responds with a given alternative with some bias probability w_r. Therefore, the predicted probability of a response on auditory-alone trials is equal to

$$P(r \mid A) = a_r + \left(1 - \sum_k a_k\right) w_r, \tag{6}$$

where a_r is the probability of identifying the auditory source as response r, $\sum_k a_k$ is the probability of identifying the auditory source as any of the k response alternatives, and the term $(1 - \sum_k a_k)$ is the probability of not identifying the auditory source.

As shown in the middle panel of Figure 3, the situation is analogous for visual-alone trials. The visual speech is either identified as one of the response alternatives r or it is not. If the participant identifies the visual stimulus as a given alternative r, he or she responds with that alternative. If no identification is made, the participant responds with a given alternative with the bias probability

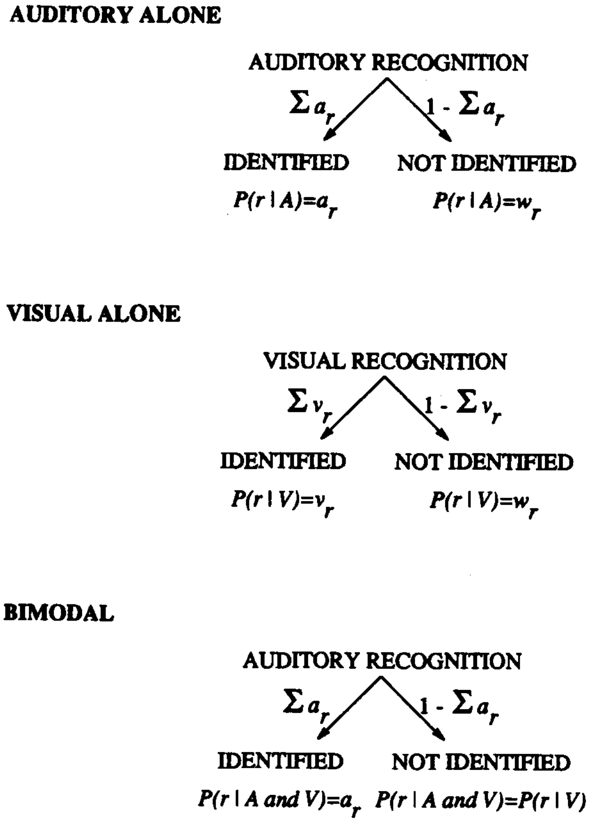


Figure 3. Decision trees for the auditory dominance model for auditory alone, visual alone, and bimodal trials. See text for explanation.

w_r. The predicted probability of a response on visual-alone trials is equal to

$$P(r \mid V) = v_r + \Big(1 - \sum_k v_k\Big) w_r, \qquad (7)$$

where v_r is the probability of identifying the visual source as response r, $\sum_k v_k$ is the probability of identifying the visual source as any of the k response alternatives, and the term $(1 - \sum_k v_k)$ is the probability of not identifying the visual source.

The bottom panel of Figure 3 illustrates the ADM for bimodal trials. The auditory speech is identified as one of the response alternatives r or not. When the participant identifies the auditory stimulus as a given alternative r, he or she responds with that alternative. In the case that no identification is made, the participant responds according to the visual information as described above. The ADM is thus capable of predicting a response that does not correspond to the stimulus presented in either individual modality. Given a bimodal stimulus, if the auditory syllable is not identified, the response is based on the visual information. If the visual information is also not identified, the participant can respond with an alternative that agrees with neither the auditory nor the visual stimuli. The predicted probability of a response on bimodal trials is equal to

$$P(r \mid A \text{ and } V) = a_r + \Big(1 - \sum_k a_k\Big)\Big(v_r + \Big(1 - \sum_k v_k\Big) w_r\Big). \qquad (8)$$

Equation 8 represents the theory that the auditory stimulus is either identified or else the decision is based on the visual information. The visible speech has an influence only when the auditory speech is not identified as one of the alternatives in the task. The model requires an a_r for each of four auditory syllables for each response alternative. Similarly, the model requires a v_r for each of four visual syllables at each orientation for each response alternative. Finally, a w_r is required for each response alternative. Given that the a_r must sum to 1, the total number of free auditory parameters is $4(n - 1)$, where n is the number of response alternatives. This constraint also holds for the v_r and w_r parameters. With 12 responses in the current tests, the model requires $4 \times 11 = 44$, plus $8 \times 11 = 88$, plus 11, for a total of 143 free parameters.

One might wonder why an ADM is necessary, because auditory dominance could be built into the FLMP and other models. However, the central assumption of auditory dominance in the ADM is qualitatively different from the corresponding assumption in the FLMP. Both modalities are always integrated in the FLMP, whereas only a single source is used on a given trial in the ADM. Thus, the FLMP and ADM provide distinctly different accounts of bimodal speech perception.

ADDITIVE MODEL OF PERCEPTION

Four different psychological models turn out to be mathematically equivalent to one another, and we call this class of models the additive model of perception (AMP). In the categorical model of perception (CMP), it is assumed that only categorical information is available from the auditory and visual sources and that the response is based on separate categorizations of the auditory and visual sources. These four cases are shown in Table 1. If the two categorizations to a given speech event agree, the identification response can follow either source. When the two categorizations disagree, it is assumed that the participant responds with the categorization to the auditory source on some proportion p of the trials, and with the categorization to the visual source on the remainder $(1 - p)$ of the trials. The weight p reflects the relative dominance of the auditory source. Considering a /ba/ response, the visual and auditory categorizations could be /ba/–/ba/, /ba/–not /ba/, not /ba/–/ba/, or not /ba/–not /ba/.

Table 1
Probabilities of the Four Possible Outcomes of the Two Unimodal Categorizations of a Bimodal Speech Stimulus for the Categorical Model of Perception

	Visual	
Auditory	/b/	not /b/
/b/	$a_{Bi} v_{Bj}$	$a_{Bi}(1 - v_{Bj})$
not /b/	$(1 - a_{Bi}) v_{Bj}$	$(1 - a_{Bi})(1 - v_{Bj})$

The probability of a /ba/ identification response given a bimodal speech event is predicted to be

$$P(\text{/ba/} \mid A_i \text{ and } V_j) = (1) a_{Bi} v_{Bj} + (p)\, a_{Bi} (1 - v_{Bj}) + (1 - \mathrm{p})(1 - a_{Bi}) v_{Bj} + (0)(1 - a_{Bi})(1 - v_{Bj}), \quad (9)$$

where *i* and *j* represent or index the levels of the auditory and visual modalities, respectively. The value a_{Bi} represents the probability of a /ba/ categorization given the auditory level *i*, and v_{Bj} is the probability of a /ba/ categorization given the visual level *j*. The value *p* reflects the amount of bias to respond with the categorization of the auditory source. Each of the four terms in Equation 9 represents the likelihood of one of the four possible outcomes multiplied by the probability of a /ba/ identification response given that outcome. Note that Equation 9 reduces to

$$P(\text{/ba/} \mid A_i \text{ and } V_j) = (p)(a_{Bi}) + (1 - p) v_{Bj}. \quad (10)$$

For each response alternative, this model requires four free parameters for the auditory source and eight for the visual. A single bias value *p* is also a necessary free parameter. Fitting the 12 responses in the present experiment thus requires 145 free parameters.

It has also been proposed that sources of information are added together to achieve perceptual recognition (Bruno & Cutting, 1988; Cutting, Bruno, Brady, & Moore, 1992). In our formulation of the AMP, the integration of the auditory and visual information is assumed to be additive rather than multiplicative. Multiplicative integration in the FLMP leads to the prediction that the contribution of one source of information is larger to the extent that the other source is ambiguous. In the AMP, on the other hand, the contribution of one source remains fixed regardless of the contribution of the other. The AMP is mathematically equivalent to the CMP, in which only categorical information is available from the auditory and visual sources. The AMP is also equivalent to (1) a single-channel model in which only a single source of information contributes to the decision on any trial (Thompson & Massaro, 1989) and (2) a weighted averaging model in which the participant simply performs a weighted averaging of the two modalities (Massaro, 1987). Thus, Equation 10 tests four different psychological models.

PRELABELING MODEL

Braida (1991) proposed a prelabeling model (PRLM), which is similar in certain respects to the FLMP. In the taxonomy of Massaro and Friedman (1990) and Cohen and Massaro (1992), the PRLM is a multidimensional version of the theory of signal detectability (TSD). A presentation of a stimulus in a given modality ideally locates that stimulus at a stimulus center in a multidimensional space. Given that the process is noisy (Gaussian), however, the actual stimulus location on a given presentation may be displaced from the stimulus center. There is also a *response center* (prototype) in the multidimensional space. The multidimensional space for a bimodal presentation is simply the combination of the spaces for the two unimodal presentations. For example, if the auditory and visual sources are each represented in 4-D space, the bimodal information is represented in 8-D space. In all cases, the participant chooses the response alternative whose response center (or prototype) is closest to the location of the stimulus in the multidimensional space.

Given the random noise involved in the PRLM, the fits involve a Monte Carlo simulation during each step of the model fit. For the final fit, for each of the stimulus conditions, 1,000 random Gaussian noise cases were averaged to produce the predicted proportion of response categorizations. This final fit was preceded by initial estimates from a preliminary closed-form multidimensional scaling (MDS) model and a 100-case PRLM model. Our tests of the PRLM turned out to be extremely time intensive, with about 30 h required for each participant's data for the 1,000-case fit, even with the incorporation of precompiled Gaussian random deviates. About 30 sec were required in order to test each participant for the FLMP and other closed form models.

In his tests of the PRLM, Braida (1991) used an MDS technique to find the locations of stimulus and response centers in order to minimize the errors in prediction of each of the two unimodal conditions. The response prototypes were assumed to be equal to their respective stimulus centers. The bimodal judgments were predicted from the combined spaces of the unimodal judgments. For his fits of the FLMP, Braida simply used the unimodal data to directly predict the bimodal points. Neither of these two tests is an optimal test because only the unimodal results are used. In Braida's test of the PRLM, the bimodal results cannot influence the location of the stimulus centers in the multidimensional space. In the test of the FLMP, he assumed that the unimodal results are an error-free measure of the parameters of the FLMP. In the current paper, however, minimization model-fitting techniques have been applied to both the unimodal and bimodal results for the tests of all models. Thus, we should have a direct comparison between the models when all models are performing as optimally as possible.

INDIVIDUAL DIFFERENCES, INFORMATION, AND INFORMATION PROCESSING

It is well known that individual differences in perception exist. Usually, experimental investigations are aimed at reducing these differences as much as possible, and/or the results are averaged across participants. This approach may preclude discovery of important properties of the processes of interest. Our research strategy is to test fewer individuals for a longer period, and to analyze the results of each participant separately. A sufficient number of observations is, therefore, recorded at each condition to justify tests of the models against the results of each participant. This approach illuminates the degree to which there are individual differences of interest, and the extent to which a given model captures the performance of all participants.

Individual differences can be meaningless or misleading, however, unless the investigator has available a good process model of the task. We all know individuals differ, but we want to know how they differ. Individuals might simply differ with respect to the information they have, or they might differ in how they process the information. The FLMP framework makes apparent an important distinction between information and information processing. *Information* refers to what the stimulus input means to the perceiver and can be thought of as the amount of information extracted. Because of unique life histories, a given stimulus event will have different degrees of meaning for different individuals. Information can be equated with the inputs and outputs of the operations in the FLMP, and is represented by the auditory and visual parameter values a_{Bi}, v_{Bj}, and so on for the other response alternatives. *Information processing* refers to the nature of the evaluation, integration, and decision operations, not to the input to or output from these operations. The degree to which the information processing is consistent with the assumptions of the FLMP is given by the goodness-of-fit of the model. The model allows for individual differences at the level of input or output, but not in the nature of the processes of evaluation, integration, and decision. Individual perceivers might differ with respect to either or both information and information processing.

Consider a naturally spoken auditory /ba/. For any arbitrarily chosen participant, it is not possible to predict how much it supports the alternative /ba/. People have unique representations of speech categories, given their unique speech histories. We can guess that the stimulus will be perceived as more /ba/-like than /va/-like, but we cannot exactly quantify how much—even given the results of hundreds of other observers. Similarly, we cannot predict how much inverting the face will disrupt the recognition of a visible /ba/. The FLMP makes a very strong prediction, however. Regardless of the amount of /ba/-ness from a given source of information, it will be evaluated and combined with other sources of information, as prescribed by the evaluation, integration, and decision operations. Thus, testing the FLMP against the results also tests whether individual differences can be located entirely at information differences.

This distinction between information and information processing is directly relevant to the question of bimodal speech perception with an inverted view of the face. Will inverting the face view simply degrade visible speech perception, or will it also disrupt the integration of visible speech with audible speech? A direct test of this question involves tests of the FLMP under upright and inverted views of the face. The FLMP has been shown to

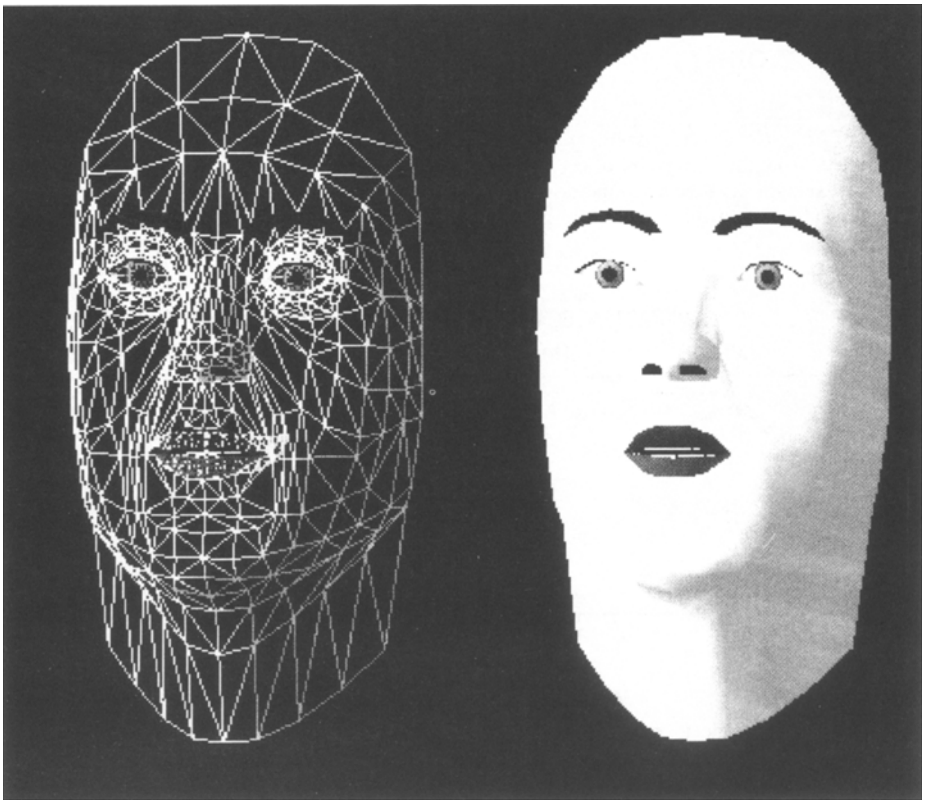


Figure 4. Framework (left) and Gouraud shaded (right) renderings of polygon facial model.

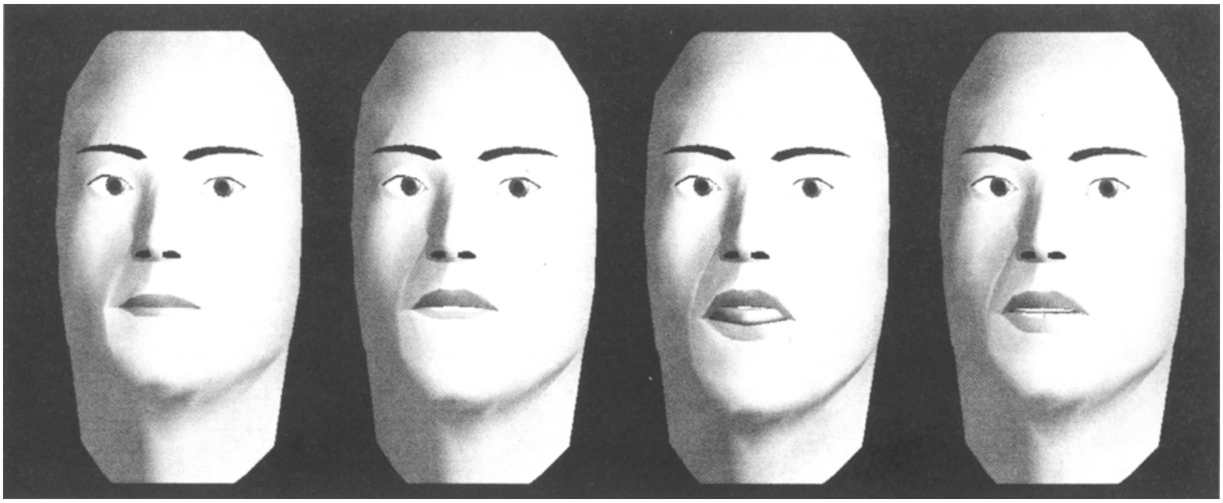


Figure 5. The facial model at the onset of the syllable for each of the four consonants. The lips are closed at the onset of /ba/, much of the lower lip is hidden by the teeth in /va/, the tongue is between the teeth in /ða/, and the mouth is slightly open at the onset of /da/.

give a good description of bimodal speech perception under a variety of conditions and in a variety of paradigms (Massaro, 1987, 1992). We expect that it will also give a good description of bimodal speech perception across different facial orientations. That is, we make the strong prediction that inverting the view of the face might degrade the visible information but it will not disrupt the robust integration of the auditory and visible speech.

METHOD

Participants

Twenty native speakers of American English participated in this experiment. All were students from the University of California, Santa Cruz. They reported having normal hearing and normal or corrected-to-normal vision. Four participants were working in the research laboratory and the other 16 participated as an option to fulfill a course requirement. All participants were unfamiliar with the specific goals of the study. They were not screened for auditory and visual sensitivity because the unimodal conditions in the current experimental task provide a measure of each.

We use synthetic visible speech in order to control the visible speech exactly and to align it appropriately with the audible speech. In addition, it is possible to present the animated face in upright or rotated position by changing the viewing angle. Previous studies have had to turn the video monitor and were not able to vary the orientation of the face randomly within a block of trials. A parametrically controlled polygon topology was used to generate a fairly realistic animation facial display (Cohen & Massaro, 1990, 1993, 1994) for the visible speech stimuli. The animation display was created by modeling the facial surface as a polyhedral object composed of about 900 small surfaces arranged in 3-D, joined together at the edges (Parke, 1974, 1975, 1982). The left panel of Figure 4 shows a framework rendering of this

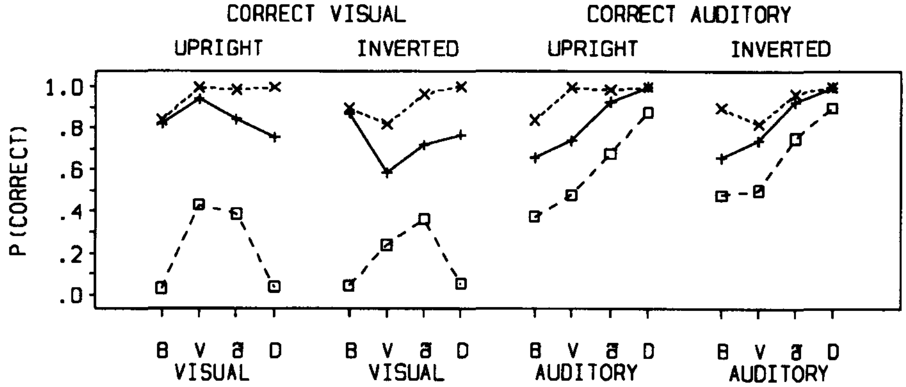


Figure 6. Average performance scored in terms of accuracy with respect to a given modality. Proportion correct on unimodal trials (solid lines and plus signs) is given for the visual syllable in the upright and inverted (left two panels) conditions. These two panels also give accuracy for the visual syllable for the bimodal trials when the auditory information is consistent with the visual (small dashed line and x) and when the auditory information is inconsistent with the visual (large dashed line and squares). Analogous measures are given for accuracy on the auditory syllable in the right two panels.

Table 2
Proportion of Responses as a Function of the Visual Stimulus for Upright and Inverted Orientations in the Unimodal Visual Condition

Visual	Response	Upright	Response	Inverted
b	b	0.8200	b	0.8675
	v	0.0775	v	0.0550
	d	0.0600	d	0.0375
	ð	0.0250	ð	0.0250
v	v	0.9350	v	0.5825
	b	0.0400	d	0.2425
	d	0.0200	b	0.0825
			ð	0.0725
d	ð	0.8375	ð	0.7150
	d	0.0750	d	0.1350
	v	0.0300	v	0.0550
	bð	0.0200	dð	0.0350
			b	0.0300
d	d	0.7525	d	0.7650
	ð	0.1100	ð	0.0850
	v	0.0800	v	0.0800
	b	0.0400	b	0.0475

Note—Responses are listed in order of magnitude.

model. To achieve a natural appearance, the surface was smooth shaded using Gouraud's (1971) method (shown in the right panel of Figure 4). To achieve a more realistic synthesis, a tongue with four control parameters was added to the facial model.

The face was animated by altering the location of various points in the grid under the control of 58 parameters, 17 of which were used for speech animation. Each phoneme is defined in a table according to target values for the 17 control parameters and segment duration. Examples of the control parameters include jaw rotation, mouth width, lip protrusion, lip corner width (width from the inner to outer lip margins at the corner), mouth corner protrusion, mouth corner horizontal offset, mouth corner height, lower lip "f" tuck (which slides the lower lip up and over the lower teeth), upper and lower lip raise, tongue angle (moves the front of the tongue up and down) and length (which moves the front of the tongue forward and back), and jaw thrust (which moves the jaw forward and back). Parke's software, revised by Pearce, B. Wyvill, G. Wyvill, and Hill (1986) and ourselves (Cohen & Massaro, 1990, 1993), was implemented on a Silicon Graphics Inc. Crimson-VGX computer. For the inverted face views, the computer animation software simply twisted the view by 180° in the frontal parallel plane. This is analogous to inverting the monitor, as was done in the previous studies with natural inverted faces.

Apparatus and Materials

Synthetic visible speech and natural audible speech were used as test stimuli. The stimuli were the consonant-vowel (CV) syllables /ba/, /va/, /ða/, and /da/. The computer was programmed to pronounce these four syllables in either upright or inverted orientations. Figure 5 shows the inverted view of the synthetic face at the onset of the articulation of the four syllables. For viewing the upright face, the reader should simply invert the figure. The natural auditory speech stimuli were four syllables, /ba/, /va/, /ða/, and /da/, taken from the Bernstein and Eberhardt (1989) videodisk database.

On each trial, the face with the default parameter values (neutral face) was played for 1,300 msec preceding the presentation of the visible test syllable. The neutral face was either presented upright or inverted to agree with the orientation of the test syllable on that trial. The neutral face also remained on the display during the response and intertrial intervals. A 100-msec, 1000-Hz warning tone was played 600 msec into the initial 1,300-msec static facial display. The visible syllable was presented without auditory speech on visual trials. Audiovisual stimuli were created by combining the auditory speech of the four syllables with the visual speech of each of these syllables (in the two orientations). The synthetic visible speech was made to mimic the natural visible syllables. The durations of the synthetic visible speech agreed with the corresponding natural syllables and were approximately 730 msec for /ba/, 730 msec for /va/, 900 msec for /ða/, and 667 msec for /da/. The durations of these four auditory syllables were 396, 470, 506, and 422 msec for /ba/, /va/, /ða/, and /da/, respectively. For the bimodal trials, the onset of the auditory syllable was aligned with the visible syllable at exactly the point the auditory speech would have begun in the original unaltered natural syllable. For the auditory trials, a dark screen was presented starting 1,300 msec before the auditory stimulus. As with the visual or bimodal trials, a 100-msec 1000-Hz warning tone was played 600 msec into this initial 1,300-msec interval.

Design and Procedure

Natural auditory and synthetic visual speech were manipulated in the expanded factorial design illustrated in Figure 1. There were four possible auditory speech syllables, four visible speech syllables, and the visible syllables could be presented upright or inverted 180°. Thus, there were 4 auditory trials, 8 visual trials, and 4 × 8 or 32 bimodal trials. Each of these 44 trial types was presented once in every block of 44 trials. Unknown to the participants, there were 10 unanalyzed practice trials before each experimental session. There were five blocks per session, and two sessions per day. Participants were tested on 2 days. This design gives a total of 20 observations per participant per condition.

Participants were instructed to listen and to watch the speaker, and to identify the consonant of the syllable as /b/, /v/, /ð/, or /d/, or as any combination of two of these consonants (i.e., a consonant cluster). This gave a total of 20 possible responses. The participants made their responses by pressing a key labeled as "b," "v," "th," or "d" on the terminal keyboard for single responses or pressed two keys successively for consonant cluster responses. The experiment was participant driven (e.g., a next trial would occur only after all of the simultaneously tested participants had responded to the previous trial).

The display monitor subtended a horizontal visual angle of 27° at a viewing distance of 50 cm. The synthetic face was displayed in color in the center of the screen, and subtended a horizontal visual angle of 11.8°. The experimental stimuli were presented to the participants over individual NEC Model C12-202A 12-in. color monitors. The intensity of the auditory stimuli was 67 dB-A (slow) measured in the approximate position of the observer's right ear for the repeated vowel /a/. The measurement was done with the sound level meter (B&K 2231, with the Microphone Type 4133).

Table 3
Proportion of Responses as a Function of the Auditory Stimulus in the Unimodal Auditory Condition

Auditory	Response	
b	b	0.6550
	v	0.3325
v	v	0.7375
	ð	0.2200
ð	ð	0.9225
	dð	0.0275
	v	0.0200
d	d	0.9950

Note—Responses are listed in order of magnitude.

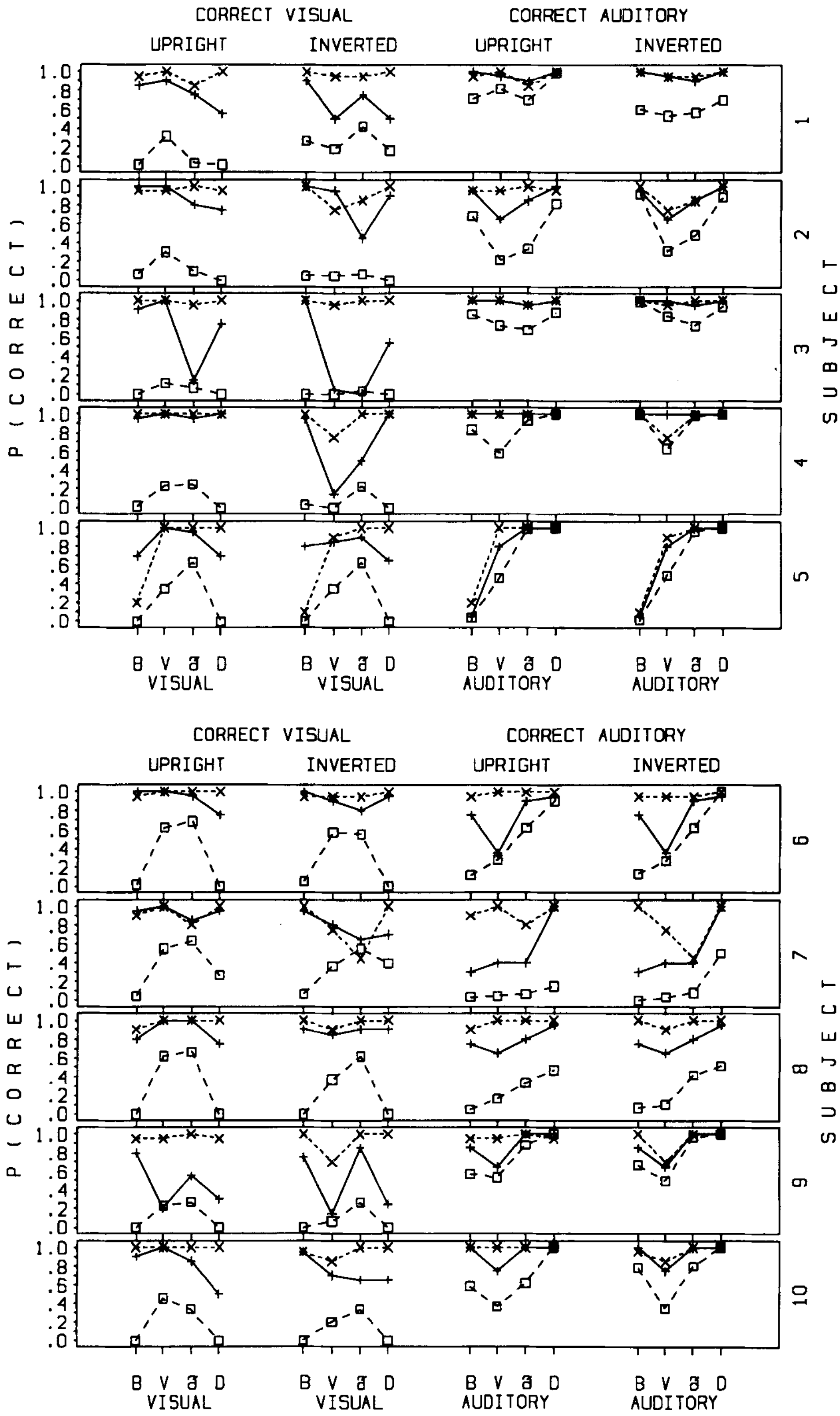
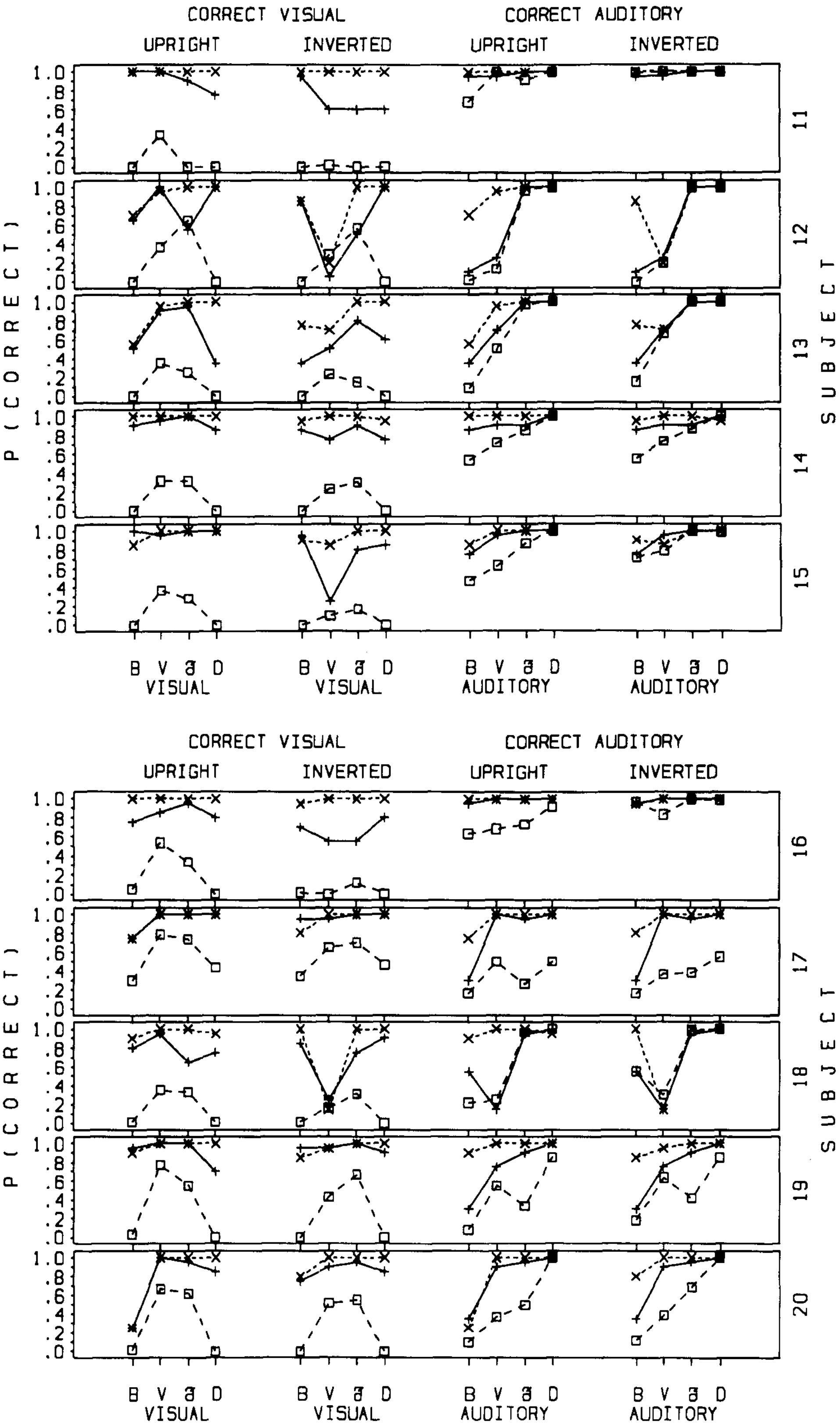


Figure 7. Individual performance for the 20 participants scored in terms of accuracy with respect to a given modality. Proportion correct on unimodal trials (solid lines and plus signs) is given for the visual syllable in the upright and inverted (left two panels) conditions. These two panels also give accuracy for the visual syllable for the bimodal trials when the auditory information is consistent



with the visual (small dashed line and x) and when the auditory information is inconsistent with the visual (large dashed line and squares). Analogous measures are given for accuracy on the auditory syllable in the right two panels.

Up to 4 participants 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.

RESULTS AND DISCUSSION

Identification judgments were recorded for each stimulus. The mean observed proportion of identifications was computed for each participant for the unimodal and bimodal conditions by pooling across all 20 replications of each condition.

We first measured performance in terms of accuracy with respect to a given modality, as shown in Figure 6. The solid lines with plus signs give performance on the unimodal conditions. The accuracy of lipreading the upright and inverted faces in the unimodal condition can be used to assess whether inverting the face degrades the visual information. This primary manipulation was successful: Inverting the face disrupted the identification of the visible syllables (left two panels of Figure 6). Accuracy of lipreading for the visual-alone condition was .836 with the upright and .733 with the inverted face [$F(1,19) = 14.67$, $p = .001$], although the interaction between orientation and syllable indicated that this effect was due to poorer performance only for the inverted /va/ and /ða/ syllables [$F(3,57) = 18.18$, $p < .001$].

Overall accuracy does not completely represent performance, however, and it is important to evaluate the confusions among test syllables in the bimodal conditions. Thus, we also analyzed the proportion of times each response was given to each of the unique stimulus conditions. Table 2 gives the proportion of responses to the four syllables presented in the upright and inverted orientations. All responses that were given more than 2% of the time to a given stimulus are presented. In each cell, the responses are given in decreasing order of response probability. In general, the confusions were similar for the upright and inverted orientations. Table 2 shows that inverting /va/ dramatically increased the misidentification of this syllable as /da/. Thus, inverting the visible speech decreased the discriminability between /va/ and /da/. Inverting the syllable /ða/ increased the likelihood that it was misidentified as /da/ or /va/. Although the meaning of these confusions is uncertain, it is clear that inverting the face decreases the information value of the visible speech.

The table shows that inversion disrupted performance on only the syllables /va/ and /ða/. This interaction between syllable and orientation can be reasonably explained by considering the visible characteristics that are used to distinguish these syllables. As depicted in Figure 5, an obvious feature for /ba/ is mouth closure at its onset whereas /da/ has an open mouth at onset. Inverting the face should not degrade these primary cues since they would be functionally equivalent in the inverted face. The syllables /va/ and /ða/, on the other hand, have visible characteristics that are more easily distorted by inverting the face. More of the upper lip is exposed in /va/ than in the other three syllables, and only the upper teeth are visible. Although only a narrow portion of the upper teeth is visible in /ða/, no part of the lower teeth can be seen. Inverting these two syllables would necessarily distort these functionally important cues. Stated in another manner, the cues for /ba/ and /da/ are more symmetric when rotated about a horizontal line, whereas the cues for /va/ and /ða/ are less so. It appears that inverting the face tends to disrupt processing of the asymmetric cues more than the symmetric ones. Perceivers appear to have difficulty normalizing for the orientation of the face, using a process such as mental rotation, for example. To summarize, this analysis documents that inverting the face decreases the visual information in speech perception. This result is similar to the findings that face recognition decreases with inverted faces (Rhodes, 1994; Valentine, 1988).

Analogous measures are given for correct performance on the auditory syllable in the right two panels of Figure 6. In this case, performance was scored in terms of accuracy with respect to the auditory modality. The solid lines with plus signs give performance on the unimodal condition. Note that the unimodal auditory condition is actually the same data in the upright and inverted face conditions. Table 3 gives the proportion of responses to the four auditory syllables. Table 3 shows that the natural auditory speech was not always identified accurately when presented alone. There were large individual differences, however, and some of the participants identified all four syllables almost perfectly. Correct performance in the unimodal auditory condition differed for the four syllables, averaging .655, .738, .923, and .995 for the syllables /ba/, /va/, /ða/, and /da/ [$F(3,57) = 12.26$, $p < .001$]. As can be seen in Table 3, the auditory syllable /ba/ was misidentified as /va/, whereas the auditory syllable /va/ was misidentified as /ða/. When /ða/ was misidentified, it was in the form of a consonant cluster or /va/.

Bimodal accuracy is given for consistent and inconsistent trials in Figure 6; the left two panels are scored with respect to accuracy on the visual syllable, whereas the right two panels are scored with respect to the auditory syllable. The results show a large influence of both modalities on performance. The small dashed lines with x's give performance when the two modalities are consistent with each other. When scored with respect to visual performance, overall performance was more accurate with two sources of consistent information relative to just the unimodal visual condition in both the upright [$F(1,19) = 16.41$, $p = .001$] and the inverted [$F(1,19) = 24.76$, $p < .001$] conditions. When scored with respect to auditory performance, overall performance was more accurate with two sources of consistent information relative to just the unimodal auditory condition in both the upright [$F(1,19) = 20.14$, $p < .001$] and the inverted [$F(1,19) = 18.60$, $p < .001$] conditions.

In agreement with the unimodal visual results, the bimodal results in the two left panels of Figure 6 support

Table 4
Proportion of Responses as a Function of the Auditory and Visual Stimuli for Upright and Inverted Orientations in the Bimodal Condition

Visual	Auditory	Response	Upright	Response	Inverted
b	b	b	0.8350	b	0.8925
		v	0.1475	v	0.0975
v	b	v	0.8100	v	0.5100
		b	0.1775	b	0.4650
ð	b	b	0.4175	b	0.4725
		ð	0.4075	ð	0.3500
		v	0.1600	v	0.1550
d	b	b	0.5225	b	0.4950
		v	0.2525	v	0.2625
		ð	0.1800	ð	0.1675
		d	0.0350	d	0.0475
b	v	v	0.7575	v	0.7325
		bv	0.1150	bv	0.1275
		ð	0.0650	ð	0.0575
		b	0.0300	b	0.0500
v	v	v	0.9900	v	0.8150
				ð	0.1600
ð	v	ð	0.7100	ð	0.6650
		v	0.2525	v	0.2850
		ðv	0.0225		
d	v	ð	0.4875	v	0.4775
		v	0.4225	ð	0.4150
		d	0.0350	d	0.0575
		dv	0.0325		
b	ð	ð	0.6300	ð	0.6350
		bð	0.1200	v	0.1075
		v	0.1125	bð	0.1050
		vð	0.0400	bv	0.0525
		b	0.0350	b	0.0400
v	ð	ð	0.5150	ð	0.7350
		v	0.4175	v	0.1775
		vð	0.0550	vð	0.0275
				dð	0.0200
				d	0.0200
ð	ð	ð	0.9800	ð	0.9600
				dð	0.0325
d	ð	ð	0.8825	ð	0.8725
		dð	0.0400	d	0.0500
		d	0.0400	v	0.0375
		v	0.0300	dð	0.0300
b	d	d	0.8700	d	0.8625
		bd	0.1025	bd	0.0950
		b	0.0225	b	0.0375
v	d	d	0.8725	d	0.9600
		v	0.0650	v	0.0300
		vd	0.0600		
ð	d	d	0.8750	d	0.8625
		ðd	0.0675	ð	0.0700
		ð	0.0425	ðd	0.0550
d	d	d	0.9925	d	0.9975

Note—Responses are listed in order of magnitude.

the conclusion that the inverted face was a less effective source of information than the face in upright form. When scored with respect to the visual stimulus, visual accuracy was somewhat higher with the upright face when the visual and auditory information were consistent with each other [$F(1,19) = 15.26, p = .001$]. If the upright face is more influential than the inverted face, the upright face should give more benefit to auditory performance when it is consistent with the auditory information. In agreement with this prediction, the two right panels of Figure 6 show that, when scored with respect to the auditory stimulus, auditory accuracy on consistent trials was somewhat higher given the upright than the inverted face [$F(1,19) = 4.99, p = .036$].

When the two modalities were inconsistent with each other (the large dashed lines with squares in Figure 6), performance was disrupted relative to the unimodal condition. Two sources of inconsistent information disrupted performance in that the unimodal condition was always more accurate than the inconsistent bimodal condition. When scored with respect to visual performance, overall performance was less accurate with two sources of inconsistent information relative to just the unimodal visual condition in both the upright [$F(1,19) = 375.41, p < .001$] and the inverted [$F(1,19) = 452.73, p < .001$] conditions. When scored with respect to auditory performance, overall performance was less accurate with two sources of inconsistent information relative to just the unimodal auditory condition in both the upright [$F(1,19) = 54.06, p < .001$] and the inverted [$F(1,19) = 25.12, p < .001$] conditions. To relate the results to those of Campbell (1994), Jordan and Bevan (in press), Bertelson et al. (1994), and Green (1994), it is important to assess whether the so-called McGurk effect is weaker with inverted than with upright faces. When scored with respect to auditory performance, auditory accuracy with inconsistent information was significantly less (.055) with the upright than with the inverted face [$F(1,19) = 9.63, p = .006$]. Thus, our results are consistent with other extant results.

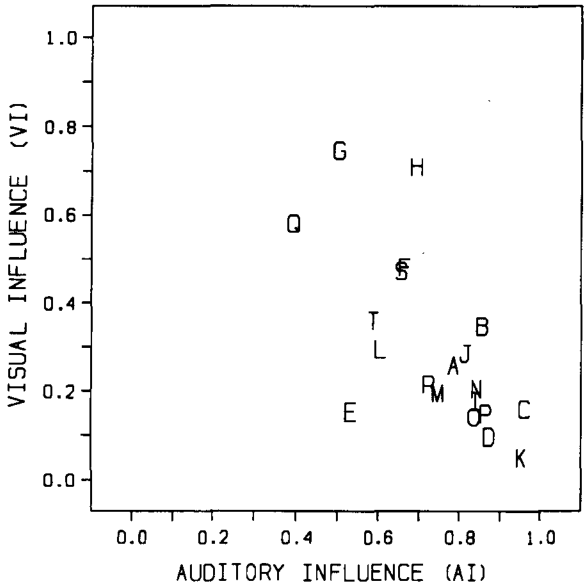


Figure 8. Visual influence as a function of the auditory influence for the 20 participants represented by the letters A through T.

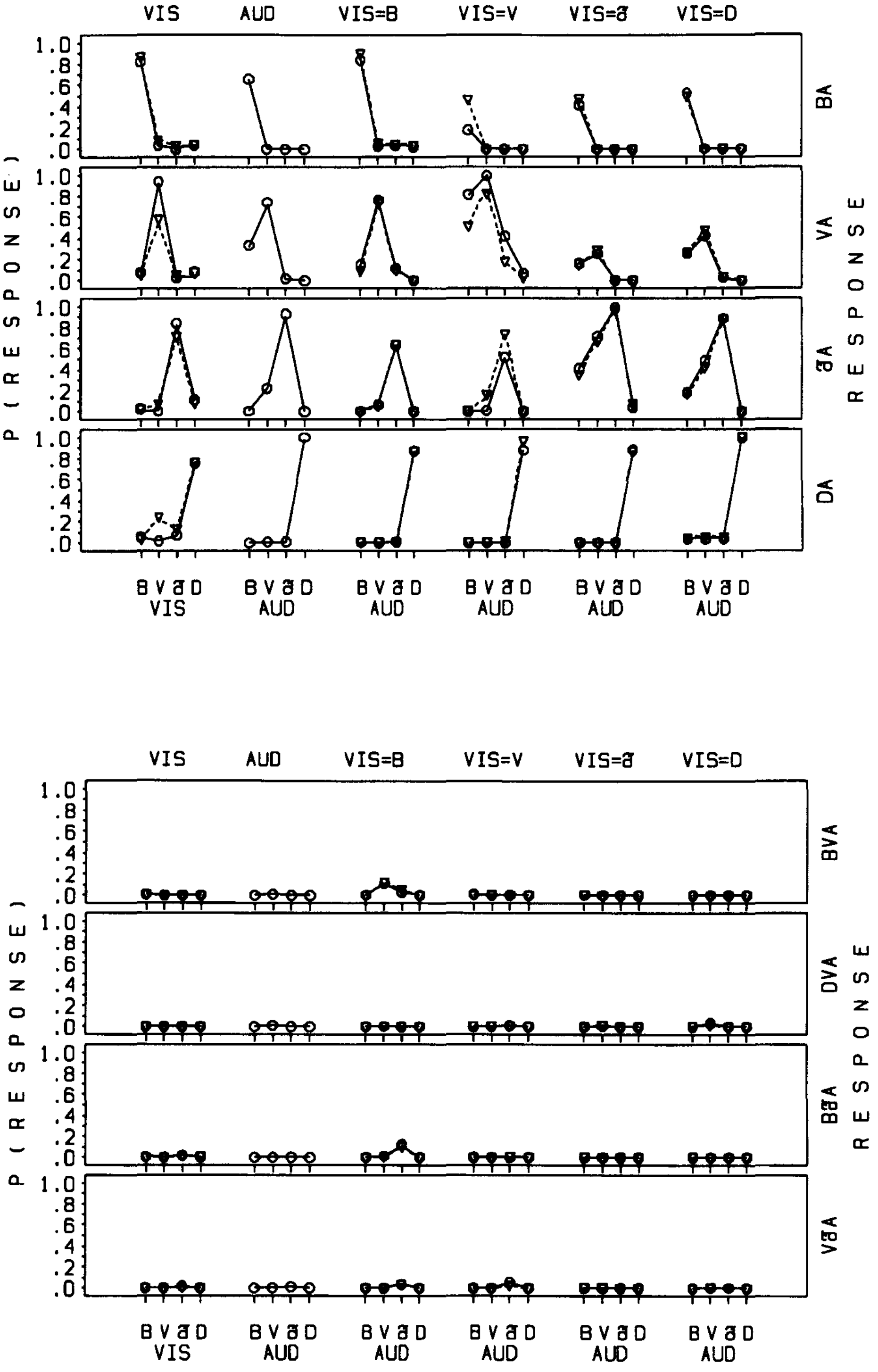


Figure 9. Observed (points) and predicted (lines) proportion of identifications for the visual-alone (far-left panel), the auditory-alone (second from left panel), and the factorial auditory–visual (right four panels) conditions as a function of the four levels of the synthetic auditory and visual speech. The lines give the predictions for the fuzzy logical model of perception (dashed line for inverted and solid line for upright face).

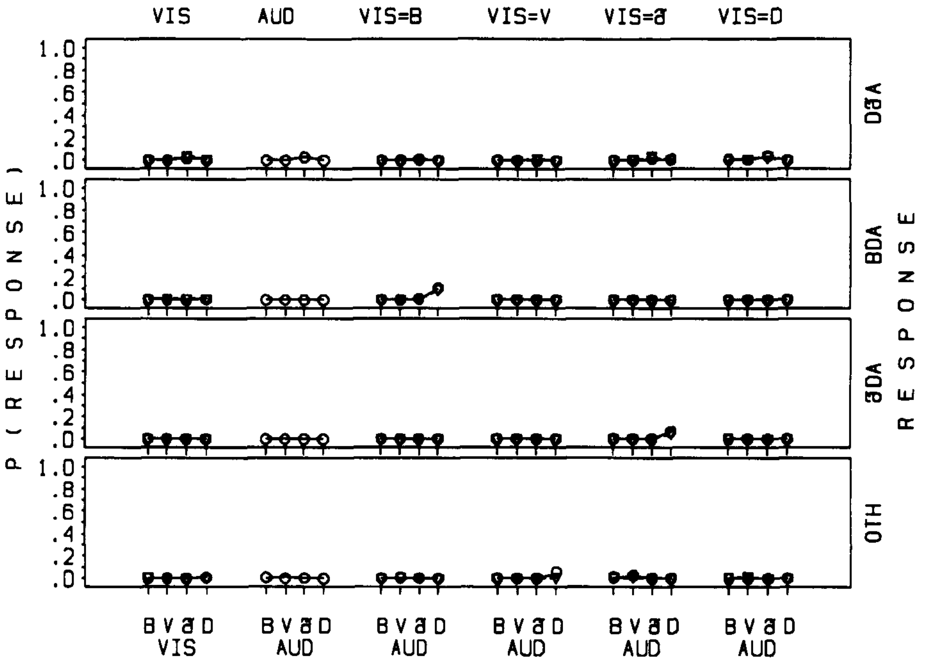


Figure 9. (continued)

With 20 observations for each participant for each condition, it is reasonable to analyze the results from individual participants. As can be seen in Figure 7, there were substantial individual differences with respect to the relative influence of the two modalities. Some participants showed a larger influence from the visual source whereas others were more influenced by the auditory. A global measure of influence is simply the vertical separation among the three curves in each of the panels. For example, Participant 11 showed a very small visual effect and a very large auditory effect. Participant 7, on the other hand, showed a larger visual effect than auditory effect.

Results shown in Tables 2–4 and Figures 6 and 7 are inconsistent with the idea that visible speech has an influence only when the auditory speech is not perfectly intelligible. When presented alone, auditory /da/ was misidentified only 2 out of 400 times, but it was identified correctly only 87% of the time when paired with inconsistent auditory information (see right two panels in Figure 6). Although auditory /da/ is relatively robust (Massaro, 1987, p. 46), inconsistent visible speech can make a substantial contribution for some participants when paired with this syllable. Some participants identified several of the syllables perfectly on auditory-alone trials, and yet the identification of these syllables was influenced by visible speech on the inconsistent trials. For example, Participants 3, 4, and 16 identified all four auditory syllables about perfectly, but were still influenced by the visible speech. If visible speech influences bimodal perception only when the auditory speech is unintelligible and we take perfect accuracy as indicating good intelligibility, there should have been no influence of the visible speech. Thus, these results provide evidence against the hypothesis that visible speech has an influence only when the auditory speech is unintelligible.

Table 4 gives the proportion of times each response was given to each of the 4 × 8 bimodal conditions. The confusions in Table 4 show a substantial number of consonant clusters in the bimodal condition. For example, a visual /ba/ paired with an auditory /da/ was identified as /bda/ .103 when upright and .095 when inverted. In all cases, the first consonant of the cluster response was articulated more forward in the mouth than was the second consonant. These results also show that the same bimodal syllable can produce both so-called fusion and combination judgments (McGurk & MacDonald, 1976). For example, an upright visual /ba/ paired with an auditory /ða/ gave .120 /bða/ and .113 /va/ judgments. This result is consistent with the FLMP assumption that a given bimodal syllable supports several alternatives in parallel.

Table 4 also shows that even the inverted face was effective in influencing bimodal speech perception. The confusions were very similar to those given with the upright face, although the overall influence of the inverted face was smaller. The upright /va/ was recognized correctly more often than was the inverted /va/, and we predict that the so-called McGurk effect would be smaller for the inverted /va/ than for the upright /va/. This prediction was substantiated. For example, a visual /va/ paired with an auditory /ba/ gave .810 /va/ responses with the upright face and only .510 /va/ responses with the inverted face. Similarly, a visual /va/ paired with an auditory /ða/ gave .418 /va/ responses with the upright face and only .178 /va/ responses with the inverted face. These results agree with the general premise that the influence of visible speech is a function of its information value (Massaro, 1984; Massaro & Cohen, 1990). According to this logic, the poorer speechreading of the inverted face necessarily implies a smaller influence on bimodal trials.

The large effect of visible speech on the bimodal trials documents previous findings of a strong influence of visible speech in bimodal speech perception (Massaro, 1987). Given that we did not also use a natural face, there is no direct measure of the effectiveness of the animated face. Other experiments with a much larger repertoire of alternatives have revealed that our current animated face has somewhat less visual information; thus, we expect that the influence of visible speech was somewhat diminished relative to what would have been found with a natural face. Even so, the present results show that this visible influence is substantial even when the auditory speech is perfectly intelligible when presented alone. This result again contradicts the viewpoint that visible speech is effective only when the auditory speech is not intelligible.

Influence of Audible and Visible Speech

We computed a measure of performance to reflect the influence of the visible and audible speech. Letting *Ua*, *Ca*, and *Ia* correspond to auditory accuracy under the unimodal, consistent bimodal, and inconsistent bimodal conditions, the visual influence *VI* is defined to be equal to

$$VI = (Ca - Ua) + (Ua - Ia) = Ca - Ia. \quad (11)$$

Given that *Ua* is the accuracy under the auditory-alone condition, the increase in auditory accuracy given consistent visual information (Condition *Ca*) gives a measure of visual influence. Similarly, the decrease in auditory accuracy given inconsistent visual information (Condition *Ia*) gives a second independent measure of visual influence. The additive combination of these two measures is then taken as the measure *VI* for each participant. Analogously, if *Uv*, *Cv*, and *Iv* correspond to visual accuracy under the unimodal, consistent bimodal, and inconsistent bimodal conditions, the auditory influence *AI* is equal to

$$AI = (Cv - Uv) + (Uv - Iv) = Cv - Iv. \quad (12)$$

The measures *Ca*, *Ia*, *Cv*, and *Iv* were computed for each participant and subjected to analyses of variance (ANOVAs). These analyses revealed that inconsistent auditory information disrupted visual performance more than inconsistent visual information disrupted auditory performance (all $ps < .001$). Similarly, but to a smaller degree, consistent auditory information improved visual performance more than consistent visual information improved auditory performance (all $ps < .001$). These results are consistent with the almost ubiquitous finding that, for perceivers with normal hearing, audible speech is more influential than is visible (Massaro, 1987, 1992).

The measures *VI* and *AI* were also computed for each participant and subjected to analyses of variance. Figure 8 gives the *VI* and *AI* values for each of the 20 participants. As can be seen in the figure, there was a larger influence from the auditory than from the visual source of information. Furthermore, some participants showed a larger influence of one or both modalities. For example, Subject G gave a larger value of *VI* than did Subject Q, and Subject B was more influenced by both modalities than was Subject E.

The *AI* and *VI* measures were also submitted to a correlation analysis. As expected from previous research (Massaro, 1992), the two types of influence were inversely correlated with each other [$r = -.672, r^2 = .451, t(18) = -3.85, p < .005$]. Thus, participants who showed a larger influence of one modality tended to show a smaller influence of the other.

Model Tests

The FLMP, ADM, AMP, and PRLM were tested against each of the individual participant's results. The models were fit to 12 responses: 11 responses whose overall proportion on bimodal trials across all participants exceeded .007, plus a 12th alternative "other" for the remaining responses. Given the 44 experimental conditions, this gives 12 × 44 or 528 data points to be predicted for each participant. For all of the models, unique visual feature parameters were estimated for each of the four syllables for the upright and inverted conditions. That is, it was assumed, for example, that the information given by an upright visual /ba/ differed from the information given by an inverted visual /ba/. No other parameters were permitted to vary across the upright and inverted conditions. The FLMP can be fit with 12 free parameters for each of the 12 response alternatives for each of the 4 auditory syllables and the 8 (4 syllables × 2 orientations) visual conditions, for a total of 144. The number of free parameters for the fit of the ADM was 143. The fit of the AMP required 145, or one more free parameter than did the fit of the FLMP.

As with the FLMP, the PRLM model we tested requires a total of 144 parameters: for each of 4 dimensions, 4 visual syllables × 2 orientation centers, 12 visual response centers, 4 auditory stimulus centers, and 12 auditory response centers. That is, there were 36 stimulus and response centers per dimension × the 4 dimensions.

The lines in Figure 9 give the predictions of the FLMP. The average RMSD values for the fit of the FLMP, AMP, ADM, and PRLM were .018, .080, .074, and .026, respectively. ANOVAs were carried out on the RMSD values with model as a factor. The FLMP gave a significantly better fit than the ADM [$F(1,19) = 167.85, p < .001$], the AMP [$F(1,19) = 142.43, p < .001$], and the PRLM [$F(1,19) = 9.11, p = .007$]. Although the FLMP gave a significantly better fit than all three of the competitors, the PRLM gave a better fit than did the ADM [$F(1,19) = 124.04, p < .001$] and the AMP [$F(1,19) = 97.78, p < .001$].

Table 5
Best Fitting Visual Parameter Values for the Fuzzy Logical Model of Perception as a Function of the Visual Stimuli for Upright and Inverted Orientations for the 12 Responses

	Upright				Inverted			
Response	b	v	ð	d	b	v	ð	d
b	0.9617	0.0610	0.0212	0.0824	0.9879	0.1577	0.0493	0.0635
v	0.1224	0.9851	0.0408	0.0691	0.0744	0.7419	0.0660	0.0698
ð	0.0130	0.0126	0.9602	0.2060	0.0166	0.1480	0.9452	0.1489
d	0.0954	0.0261	0.1031	0.9786	0.0526	0.3753	0.2200	0.9762
bv	0.0129	0.0007	0.0006	0.0096	0.0154	0.0073	0.0052	0.0073
dv	0.0039	0.0015	0.0003	0.0090	0.0039	0.0084	0.0047	0.0079
bð	0.0121	0.0007	0.0493	0.0110	0.0089	0.0035	0.0221	0.0147
vð	0.0085	0.0039	0.0281	0.0100	0.0032	0.0057	0.0087	0.0107
dð	0.0035	0.0009	0.0443	0.0133	0.0032	0.0084	0.0856	0.0227
bd	0.0112	0.0008	0.0003	0.0139	0.0092	0.0188	0.0245	0.0230
ðd	0.0038	0.0011	0.0061	0.0059	0.0039	0.0070	0.0137	0.0086
Other	0.0071	0.0036	0.0081	0.0163	0.0102	0.0103	0.0049	0.0211

Of theoretical interest is the extent to which the inversion of the face changed the nature of processing bimodal speech or simply influenced the information available in visible speech. To test this idea, we compared the fit of the FLMP predictions separately for the upright and inverted faces. If inverting the face disrupts the processes postulated by the FLMP, then the fit should be significantly poorer for the inverted than for the upright face. To test this, new RMSDs were computed from the predicted and observed results to give separate RMSDs for the upright and inverted conditions. The RMSD values for the fit of the FLMP to the upright-face and inverted-face conditions were .0158 and .0187, respectively. These differences were not statistically significant [$F(1,19) = 3.38$], indicating that inverting the face does not disrupt information processing postulated by the FLMP. Apparently, it is only the visible information that is degraded by inverting the face.

Although the effect of inverting the face was statistically significant, it is still of interest to ask whether this effect was large enough to challenge the models. To address this question, we fit a new version of the FLMP that assumes there is no effect of inverting the face. That is, this model assumes that the same visual information is available in the upright and inverted conditions. Thus, a new set of parameters was estimated with the constraint that the visual parameter for each syllable was the same in the upright and inverted conditions. The fit of this model was significantly poorer than the FLMP with different visual parameters for the upright and inverted syllables, RMSD of .041 versus .018 [$F(1,19) = 63.42, p < .001$]. Thus, we can conclude that inverting the face significantly degraded the visual information, but did not change the nature of the information processing.

Given the good fit of the FLMP, the corresponding parameter values can give a measure of the degree to which inverting the face disrupts the visual information. Tables 5 and 6 show the mean parameter values for the visual and auditory parameters, respectively. It should be noted that the parameter values in Tables 5 and 6 do not equal the predicted response proportions for the unimodal conditions. As can be seen in Equation 3, for example, $P(/\text{ba}/ \mid Ai)$ is equal to the auditory support, a_{Bi}, for /ba/ divided by the sum of support for all relevant alternatives. The sum of support does not have to add up to 1, and thus the predicted response proportion will probably differ from the corresponding parameter value.

An ANOVA was carried out on the the parameter values supporting the correct alternative for the 8 visual conditions (4 syllables × 2 orientations). There was a significant effect of orientation with more support for the correct response in the upright (.971) versus inverted (.913) presentations [$F(1,19) = 5.86, p = .024$], and a syllable × orientation interaction [$F(3,57) = 8.06, p < .001$]. This result shows that the upright face was more influential than the inverted face and this differential influence was larger for some syllables than others. This analysis, based on the FLMP's parameter values, agrees with the analyses carried out on accuracy of performance and the response confusions. This agreement between the traditional analyses and the FLMP analysis further supports our conclusions.

In summary, we studied the robustness of bimodal speech perception by varying the orientation of the face. Participants identified auditory syllables, visible syllables, and bimodal syllables with either an upright or an

Table 6
Best Fitting Auditory Parameter Values for the Fuzzy Logical Model of Perception as a Function of the Auditory Stimuli for the 12 Responses

Response	b	v	ð	d
b	0.8066	0.0031	0.0002	0.0011
v	0.4624	0.8981	0.0190	0.0017
ð	0.0258	0.3114	0.9476	0.0006
d	0.0010	0.0046	0.0048	0.9999
bv	0.0031	0.0270	0.0043	0.0002
dv	0.0023	0.0277	0.0051	0.0004
bð	0.0011	0.0029	0.0110	0.0002
vð	0.0006	0.0083	0.0249	0.0002
dð	0.0009	0.0133	0.0628	0.0016
bd	0.0006	0.0008	0.0009	0.0068
ðd	0.0033	0.0018	0.0087	0.0057
Other	0.0056	0.0175	0.0071	0.0022

inverted face. Although inverting the face disrupted performance, the FLMP gave a good description of the results with both the upright and the inverted faces. Inverting the face did not appear to change the nature of processing bimodal speech, but simply influenced the information available from the face. Thus, inverting the face is simply a method of degrading the visual information in the same way that auditory noise or bandpass filtering have been used to degrade the auditory information. The FLMP has also been successful in describing bimodal speech perception under different levels of auditory noise (Massaro, 1987, chap. 2). Parallel to the good fit of the current results, the good fit of the FLMP was achieved with the strong assumption that auditory noise did not change the nature of processing bimodal speech.

The distinction between information and information processing also speaks to issues in face processing. It is often assumed that different psychological systems recognize different aspects of the face. As described in the introduction, there appears to be a hemispheric asymmetry that differs for face recognition and speech perception. This asymmetry might simply reflect differences in information as opposed to differences in information processing. Similarly, a different system is putatively used for face identification from the ones used for expression, face matching, and lipreading (Etcoff & Magee, 1992; Tanaka & Farah, 1993). However, there is no reason to assume that dissociations of these behaviors necessarily reflect different systems (Levine, Banich, & Koch-Weser, 1988; Sergent, 1994). It might simply be the case that different types of information are used. For example, different cues are used for identifying sex and person identity. Thus, observed dissociations might be due to differences in information rather than differences in information processing. More generally, we believe that the distinction between information and information processing is important for experimental and theoretical progress.

REFERENCES

Bernstein, L. E., & Eberhardt, S. P. (1986). *Johns Hopkins lipreading corpus I-II: Disc I.* Baltimore: Johns Hopkins University Press.

Bertelson, P., Vroomen, J., Wiegeraad, G., & de Gelder, B. (1994, September). *Exploring the relation between McGurk interference and ventriloquism.* Paper presented at the 1994 International Conference on Spoken Language Processing, Yokohama, Japan.

Braida, L. D. (1991). Crossmodal integration in the identification of consonant segments. *Quarterly Journal of Experimental Psychology*, **43A**, 647-677.

Bruno, N., & Cutting, J. E. (1988). Minimodularity and the perception of layout. *Journal of Experimental Psychology: General*, **117**, 161-170.

Campbell, R. (1992). The neuropsychology of lipreading. *Philosophical Transactions of the Royal Society London: Series B*, **335**, 39-45.

Campbell, R. (1994). Audiovisual speech: Where, what, when, how? *Current Psychology of Cognition*, **13**, 76-80.

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

Cohen, M. M., & Massaro, D. W. (1990). Synthesis of visible speech. *Behavior Research Methods, Instruments, & Computers*, **22**, 260-263.

Cohen, M. M., & Massaro, D. W. (1992). On the similarity of categorization models. In F. G. Ashby (Ed.), *Probabilistic multidimensional models of perception and cognition* (pp. 395-447). Hillsdale, NJ: Erlbaum.

Cohen, M. M., & Massaro, D. W. (1993). Modeling coarticulation in synthetic visual speech. In N. M. Thalmann & D. Thalmann (Eds.), *Models and techniques in computer animation* (pp. 139-156). Tokyo: Springer-Verlag.

Cohen, M. M., & Massaro, D. W. (1994). Development and experimentation with synthetic visible speech. *Behavior Research Methods, Instruments, & Computers*, **26**, 260-265.

Cutting, J. E., Bruno, N., Brady, N. P., & Moore, C. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General*, **121**, 364-381.

de Gelder, B., Vroomen, J., & van der Heide, L. (1991). Face recognition and lip-reading in autism. *European Journal of Cognitive Psychology*, **3**, 69-86.

Etcoff, N. L., & Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition*, **44**, 227-240.

Fisher, B. (1991). *Integration of visual and auditory information in perception of speech events.* Unpublished doctoral dissertation, University of California, Santa Cruz.

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

Green, K. P. (1994). The influence of an inverted face on the McGurk effect. *Journal of the Acoustical Society of America*, **95**, 3014.

Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophysics*, **50**, 524-536.

Hellige, J. B. (1993). *Hemispheric asymmetry: What's right and what's left.* Cambridge, MA: Harvard University Press.

Jordan, T. R., & Bevan, K. (in press). Seeing and hearing rotated faces: Influence of facial orientation on visual and audio-visual speech recognition. *Journal of Experimental Psychology: Human Perception & Performance.*

Levine, S. C., Banich, M. T., & Koch-Weser, M. P. (1988). Face recognition: A general or specific right hemisphere capacity? *Brain & Cognition*, **8**, 303-325.

Massaro, D. W. (1984). Children's perception of auditory and visual speech. *Child Development*, **55**, 1777-1788.

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

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

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

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

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

Massaro, D. W., & Cohen, M. M. (1983). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception & Performance*, **3**, 753-771.

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

Massaro, D. W., Cohen, M. M., & Smeele, P. M. T. (1995). Cross-linguistic comparisons in the integration of visual and auditory speech. *Memory & Cognition*, **23**, 113-131.

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

Massaro, D. W., Tsuzaki, M., Cohen, M. M., Gesi, A., & Heredia, R. (1993). Bimodal speech perception: An examination across languages. *Journal of Phonetics*, **21**, 445-478.

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

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

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

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

Pearce, A., Wyvill, B., Wyvill, G., & Hill, D. (1986). Speech and expression: A computer solution to face animation. *Proceedings of Graphics Interface '86* (pp. 136-140).

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

Rhodes, G. (1994). Secrets of the face. *New Zealand Journal of Psychology*, **23**, 3-17.

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

Sekiyama, K., & Tokhura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics*, **21**, 427-444.

Sergent, J. (1994). Cognitive and neural structures in face processing. In A. Kertesz (Ed.), *Localization and neuroimaging in neuropsychology* (pp. 473-494). San Diego: Academic Press.

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, **46A**, 225-245.

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

Valentine, T. (1988). Upside-down faces: A review of the effect of inversion upon face recognition. *British Journal of Psychology*, **79**, 471-491.

Vroomen, J. H. M. (1992). *Hearing voices and seeing lips: Investigations in the psychology of lipreading*. Doctoral dissertation, Katholieke Universiteit Brabant, Nijmegen.

Zaidel, D. W. (1994). Worlds apart: Pictorial semantics in the left and right cerebral hemispheres. *Current Directions in Psychological Science*, **3**, 5-8.

(Manuscript received February 3, 1995;
revision accepted for publication January 15, 1996.)