Memory & Cognition
1995, 23 (1), 113-131

Cross-linguistic comparisons in the integration of visual and auditory speech

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

and

PAULA M. T. SMEELE
Delft University of Technology, Delft, The Netherlands

We examined how speakers of different languages perceive speech in face-to-face communication. These speakers identified synthetic unimodal and bimodal speech syllables made from synthetic auditory and visual five-step /ba/–/da/ continua. In the first experiment, Dutch speakers identified the test syllables as either /ba/ or /da/. To explore the robustness of the results, Dutch and English speakers were given a completely open-ended response task. Tasks in previous studies had always specified a set of alternatives. Similar results were found in the two-alternative and open-ended task. Identification of the speech segments was influenced by both the auditory and the visual sources of information. The results falsified an auditory dominance model (ADM) which assumes that the contribution of visible speech is dependent on poor-quality audible speech. The results also falsified an additive model of perception (AMP) in which the auditory and visual sources are linearly combined. The fuzzy logical model of perception (FLMP) provided a good description of performance, supporting the claim that multiple sources of continuous information are evaluated and integrated in speech perception. These results replicate previous results found with English, Spanish, and Japanese speakers. Although there were significant performance differences, the model analyses indicated no differences in the nature of information processing across language groups. The performance differences across languages were caused by information differences due to different phonologies in Dutch and English. These results suggest that the underlying mechanisms for speech perception are similar across languages.

Earlier research has shown that visual information from a talker's face can improve speech intelligibility over that obtained with the presentation of only auditory information. This improvement is most noticeable when the auditory signal is degraded by hearing impairment, the presence of noise, or bandwidth filtering (Binnie, Montgomery, & Jackson, 1974; Breeuwer & Plomp, 1984; Massaro, 1987; Summerfield, 1979). Speech perception is superior with visual information for sentences (Reisberg, McLean, & Goldfield, 1987; Summerfield, 1979), words (Campbell & Dodd, 1980), or nonsense words (Binnie et al., 1974; Smeele & Sittig, 1991a). The generality of these results is particularly informative because it indicates a contribution of visible speech regardless of the lexical status or the sentential context of the speech.

Although the influence of visible speech is substantial when auditory speech is degraded, visible speech also contributes to performance when it is paired with intelligible audible speech. Experiments in which conflicting auditory and visual information have been presented (Green & Kuhl, 1989; Massaro & Cohen, 1983; McGurk & MacDonald, 1976) have shown that vision strongly influences speech perception. Furthermore, experiments in which the synchronization between the audio and visual channels has been varied (Campbell & Dodd, 1980; Cohen, 1984; Massaro & Cohen, 1993b; Smeele & Sittig, 1991b; Smeele, Sittig, & van Heuven, 1992), have indicated that visual information can still be successfully integrated with audition even when there is a severe time delay. Thus, the strong influence of visual speech is not limited to situations with degraded auditory input; it also appears to have an important influence even when paired with perfectly intelligible speech sounds or when presented asynchronously with auditory speech.

Although the study of how humans perceive bimodal speech has been primarily carried out with English talkers, there has been one recent cross-linguistic examination of speech perception (Massaro, Tsuzaki, Cohen, Gesi, & Heredia, 1993). Identification responses of English, Spanish, and Japanese speakers to synthetic auditory and visual syllables were compared. The synthetic speech was manipulated in an expanded factorial design, shown in Figure 1. Five levels of audible speech

The research reported in this paper and the writing of the paper were supported, in part, by grants from the Public Health Service (PHS R01 NS 20314), the National Science Foundation (BNS 8812728), and the graduate division of the University of California, Santa Cruz. We thank Brain MacWhinney for his comments on an earlier version of this paper. Correspondence should be addressed to D. W. Massaro, Program in Experimental Psychology, Clark Kerr Hall, University of California, Santa Cruz, CA 95064 (e-mail: massaro@ fuzzy.ucsc.edu).

Auditory \ Visual	/ba/	2	3	4	/da/	None
/ba/						
2						
3						
4						
/da/						
None						

Figure 1. Expanded factorial design used in the present experiments to include both bimodal speech and auditory and visual conditions presented alone. The five levels along the auditory and visual continua represent auditory and visual speech syllables varying in equal physical steps between B and D. For the auditory continuum, B corresponds to rising $F2$ and $F3$ transitions and D corresponds to falling $F2$ and $F3$ transitions. For the visual continuum, B corresponds to closed lips at the onset of the syllable and D corresponds to open lips at onset.

varying between /ba/ and /da/ were crossed with five levels of visible speech varying between the same alternatives. The onsets of the second and third formants were varied to give an auditory continuum between the syllables /ba/ and /da/. In analogous fashion, parameters of an animated face were varied to give a continuum between visual /ba/ and /da/. This design allows one to address the question of how the identification of a bimodal syllable occurs as a function of the unimodal syllables that compose it. The design is more powerful than a simple factorial design for testing different models (Massaro & Friedman, 1990). In the present experiments, the same stimuli were used to extend this cross-linguistic research to another language group: Dutch speakers. This research should allow us to determine the degree to which we can generalize our conclusions to a new language. In addition, the second experiment permitted a comparison between allowing subjects a completely open-ended set of response alternatives and giving subjects eight alternatives in advance.

MODELS OF BIMODAL SPEECH PERCEPTION

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 were devised specifically to reject some theoretical alternatives. A fuzzy logical model of perception (FLMP), an auditory dominance model (ADM), and an additive model of speech perception (AMP) were formalized and tested against the results. The FLMP has been the most successful model to date (Massaro, 1987, 1989b, 1990; Massaro & Friedman, 1990) and we begin with the description of this model.

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 there are usually multiple sources of information that the perceiver evaluates and integrates to achieve perceptual recognition. The following assumptions are central to the model: (1) Each source of information is evaluated to give the degree to which that source specifies the relevant alternatives, (2) the sources of information are evaluated independently of one another, (3) the sources are integrated to provide an overall degree of support for each alternative, 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 applying the FLMP to the bimodal speech perception task, both sources are assumed to provide continuous and independent evidence for each of the prototype alternatives. With the onsets of the second ($F2$) and third ($F3$) formants defined as the important auditory feature

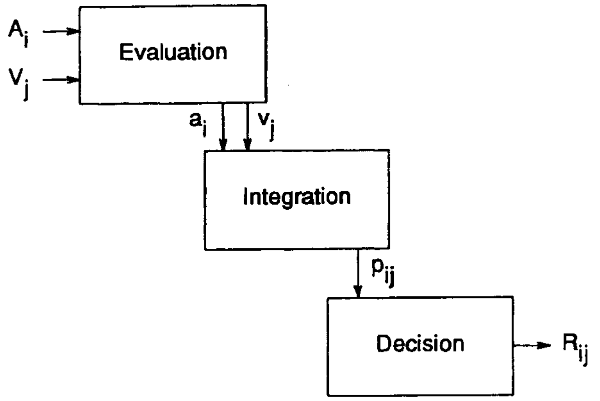

Figure 2. Schematic representation of the three operations involved in perceptual recognition. The evaluation of an auditory source of information A_i produces a truth value a_i, indicating the degree of support for alternative R. The visual source V_j is evaluated similarly to give v_j. Integration of the truth values gives an overall goodness of match p_{ij}. The response R_{ij} is equal to the value p_{ij} relative to the goodness of match of all response alternatives.

and the degree of initial opening of the lips defined as the important visual feature, the prototype for /da/ might be something like

/da/ : Slightly falling $F2$–$F3$ and open lips.

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

/ba/ : Rising $F2$–$F3$ and closed lips,

and so on for the other prototypes.

At the evaluation stage, each source of information is assigned some truth value indicating the degree of support for each relevant alternative. The truth values lie between zero and one, with zero being no support and one being complete support (Massaro, 1987). The value .5 is neutral or completely ambiguous support. We let a_{Di} represent the degree to which the auditory stimulus A_i supports the alternative /da/, that is, has *slightly falling* $F2$–$F3$. Similarly, v_{Dj} represents the degree to which the visual stimulus V_j supports the alternative /da/, that is, has *open lips*. Given a prototype's independent specifications for the auditory and visual sources, the evaluation of one source cannot change the evaluation of the other source.

The integration of the features defining each prototype is computed by taking the product of the feature values. It is assumed that the outcome of prototype matching for /da/ would be a multiplicative contribution of the auditory and visual support:

$$S(/\mathrm{da}/|A_i \text{ and } V_j) = a_{Di} \times v_{Dj}, \qquad (1)$$

where $S(/\mathrm{da}/|A_i$ and $V_j)$ is the support for the prototype /da/, given auditory and visible speech, and the subscripts i and j index the levels of the auditory and visual modalities, respectively. Analogously, if a_{Bi} represents the degree to which the auditory stimulus A_i has *rising* $F2$–$F3$ and v_{Bj} represents the degree to which the visual stimulus V_j has *closed lips*, the outcome of prototype matching for /ba/ would be

$$S(/\mathrm{ba}/|A_i \text{ and } V_j) = a_{Bi} \times v_{Bj}, \qquad (2)$$

and so on for the other prototypes.

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 one A_i, the probability of a /da/ response, $P(/\mathrm{da}/)$, is predicted to be

$$P(/\mathrm{da}/|A_i) = \frac{a_{Di}}{\sum_k a_{ki}}, \qquad (3)$$

where the denominator is equal to the sum of support for all relevant (k) alternatives. Similarly,

$$P(/\mathrm{da}/|V_j) = \frac{v_{Dj}}{\sum_k v_{kj}}. \qquad (4)$$

Given two sources of information A_i and V_j, $P(/\mathrm{da}/)$ is predicted to be

$$P(/\mathrm{da}/|A_i \text{ and } V_j) = \frac{a_{Di} \times v_{Dj}}{\sum_k (a_{ki} \times v_{kj})}. \qquad (5)$$

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 for just one. However, the identification judgment is a function of the relative degree of support as shown in Equations 3, 4, and 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).

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 is given by how much the source matches the corresponding ideal value. 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. An auditory parameter is forced to remain invariant across variation in the different visual conditions and, analogously, for a visual parameter. Given five levels of auditory and visual speech, the FLMP requires 5 free parameters for the visual feature values and 5 for the auditory feature values for each response alternative. With just two mutually exclusive response alternatives, the support for the second alternative is one minus the support for the first. Only a total of 10 free parameters is needed. (The procedure for estimating the free parameters for the fit of the models is given in the section Tests of the Models.)

Finally, the FLMP, as depicted in Figure 2, allows one to make an important distinction between *information* and *information processing* (Massaro, 1987, 1989). One component of information corresponds to the outcome of evaluation: how much a particular stimulus presented to a given input channel supports the various alternatives. One component of information processing corresponds to the process of integration: how the various sources of information are combined. Perceivers of different linguistic groups might differ with respect to either or both of these characteristics. Consider the second level along a synthetic auditory speech continuum between /ba/ and /da/. This stimulus might support the alternative /ba/ for one language significantly more than for another language. We cannot hope to equate the amount of support for a given category across different linguistic groups. We simply synthesize the same range of speech stimuli for the different languages and have the subjects categorize these stimuli.

Given the unique phoneme inventories and phonologies of the different languages, we may see different response patterns from the different linguistic groups. The

FLMP makes a very strong prediction, however. Regardless of the amount of /ba/-ness from a given source of information, it will be combined with other sources of information, as prescribed by the integration and decision operations. With respect to the integration of audible and visible speech, the information value of a given modality might differ, but it will be combined with the other modality in the same manner for all languages. Thus, the model allows for linguistic differences in the truth values or degrees of support assigned at the level of evaluation, but not in the processes of integration and decision. Thus, testing the FLMP against the results also tests whether linguistic differences can be located entirely at the evaluation stage of processing.

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

Auditory Dominance Model

A second potential explanation of the influence of visible speech is that an effect of visible speech occurs only when the auditory speech is not completely intelligible (Sekiyama & Tohkura, 1991; Vroomen, 1992). Given the reasonable observation that speech is primarily auditory (Studdert-Kennedy, 1989), it would be only natural to believe that visible speech must necessarily play a secondary role in bimodal speech perception. Some authors have recently proposed that visible speech will influence perception *only* when the auditory information is ambiguous (Sekiyama & Tohkura, 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 easy to define and implement in a model. Perfect identification in an auditory test might not mean perfect intelligibility. Even given these limitations in the measure of intelligibility, we can formulate one version of the intelligibility hypothesis, the 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. It should be noted that the all-or-none assumption about auditory identification in the ADM is *not* inconsistent with the assumption that intelligibility is a continuous measure. Intelligibility is determined from a set of identification trials. Even though identification is all-or-none on any given trial, the proportion of identifications over a set of trials would give a continuous measure of intelligibility.

According to the ADM, 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 is shown in the top panel of Figure 3, the auditory speech is identified as one of the response alternatives r or not. When the subject identifies the auditory stimulus as a given alternative r, he/she responds with that alternative. In the case that no identification is made, the subject 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

$$P(r|A) = a_r + \left(1 - \sum_r a_r\right) w_r, \qquad (6)$$

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

For visual-alone trials, the situation is analogous. As is shown in the middle panel of Figure 3, the visual speech is identified as one of the response alternatives r or not. When the subject identifies the visual stimulus as a given alternative r, he/she responds with that alternative. In the case that no identification is made, the subject responds

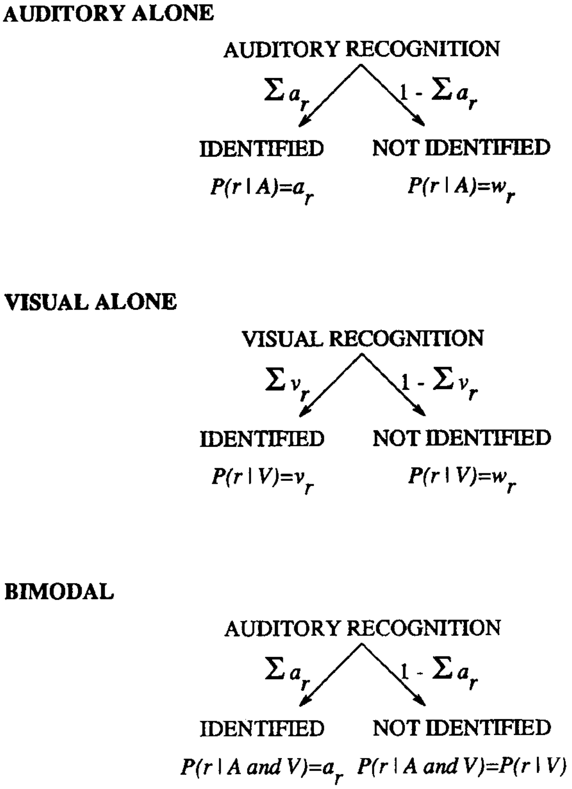

Figure 3. Probability trees for ADM for auditory alone, visual alone, and bimodal trials. See text for explanation.

with a given alternative with the bias probability w_r. Therefore, the predicted probability of a response on visual alone trials is equal to

$$P(r|V) = v_r + \left(1 - \sum_r v_r\right)w_r, \qquad (7)$$

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

Finally, we consider the bimodal case, shown in the bottom panel of Figure 3. For these trials, the auditory speech is identified as one of the response alternatives r or not. When the subject identifies the auditory stimulus as a given alternative r, he/she responds with that alternative. In the case that no identification is made, the subject responds according to the visual information as described above. Therefore, the predicted probability of a response on bimodal trials is equal to

$$P(r|A \text{ and } V) = a_r + \left(1 - \sum_r a_r\right)\left[v_r + \left(1 - \sum_r v_r\right)w_r\right]. \qquad (8)$$

Equation 8 represents the theory that either the auditory stimulus is identified or else the subject bases his/her decision 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, v_r, and w_r for each response alternative. Relative to the FLMP, this model has an additional five parameters for each response alternative.

If speakers of a given language use visible speech only when the auditory speech is *not* identified correctly, this model should give a better description of the results than should the FLMP. This model has the potential of accounting for a small use of visual speech by speakers of a given language.

Finally, 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 the ADM is qualitatively different from the FLMP assumption that both auditory and visual speech contribute to speech perception on a given trial. In the ADM, either the auditory or the visual speech (or neither) controls the judgment on a given trial. Furthermore, the concept of dominance is redundant in the FLMP, because the degree of support from a given source of information (and the support from other sources) determines the outcome of integration.

Additive Model of Perception

Additive models have been proposed and tested to explain perception and pattern recognition in several domains (Cutting, Bruno, Brady, & Moore, 1992; Massaro, 1988; Massaro & Cohen, 1993a). In the AMP, it is assumed that the sources of information are added rather than multiplied as in the FLMP.

Table 1
The Probabilities of the Four Possible Outcomes of the Two Unimodal Categorizations of a Bimodal Speech Stimulus for the Single-Channel Model and the Categorical Model

	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

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

where i and j index the levels of the auditory and visual modalities, respectively. The a_{Bi} value represents the support for /ba/, given the auditory level i; and v_{Bj} is the support, given the visual level j. The value p reflects the amount of bias to respond with the support from the auditory source. For each response alternative, the AMP requires five free parameters for the auditory source, five for the visual. A single bias value p is also a necessary free parameter.

It should be noted that the AMP is mathematically equivalent to both a single-channel model and a categorical model. In the single-channel model, the subject attends to just one modality on bimodal trials (Thompson & Massaro, 1989) and responds on the basis of this modality. Table 1 gives the probabilities of the four possible outcomes of the two unimodal categorizations of a bimodal speech stimulus for the single-channel model. Combining these probabilities reduces the single-channel model to the AMP given by Equation 9. The same mathematical equivalence holds for a categorical model in which the subject categorizes each of two modalities. The subject then responds with the outcome of the auditory categorization with some bias p and the outcome of the visual categorization with bias $1 - p$ (Massaro, 1987). The categorical model can also be described by Table 1 and reduces to the AMP described by Equation 9.

PREVIOUS RESULTS AND BASIS FOR EXPERIMENT 1

Sekiyama and Tohkura (1991) evaluated confusion matrices for auditory and bimodal speech in Japanese. Relative to previous results with English subjects, there seemed to be a smaller influence of visible speech for Japanese speakers. The authors concluded that "the 'Japanese McGurk effect' is less easily induced than the English one, and that it depends on the auditory intelligibility of the speech signal" (p. 1797). However, it is first necessary to realize that the influence of visible speech in bimodal speech perception occurs to some degree rather than being simply present or absent. When analyzed from this perspective, Sekiyama and Tohkura's (1991) findings are not surprising and can be described within the context of the FLMP. First, the influence of visible speech was greater for the less intelligible syllables and increased when noise was added to the auditory stimulus. In the FLMP, the integration process leads to the outcome that the influence of a source of information will necessarily increase to the extent that other sources are made more ambiguous. Thus, it is not necessary to assume that the auditory source has some a priori dominance but rather just that it might be less ambiguous in many contrasts. Second, Japanese speakers gave fewer consonant clusters as responses. This is not surprising if we accept that people's responses will reflect the phonemes and phonology of their language. Consonant clusters do not occur in Japanese, and we can expect fewer of them as responses than in English. For these reasons, we cannot conclude that information processing of Japanese speakers differs from that of English speakers. Information differences might be responsible for the performance differences.

The methodology of the present experiments allows us to separate information differences from information processing differences. The experiments with native-English Americans and native Spanish and Japanese speakers (Massaro & Cohen, 1990; Massaro et al., 1993) indicated important contributions of auditory and visual speech in bimodal speech perception. Most importantly, the experiments revealed both differences and similarities in performance across the different languages. The English speakers gave mostly /ba/ and /da/, /bda/, /ða/, and /va/ responses. Visible speech had a strong influence on the perceptual judgments of the English speakers. Visible articulations on the /ba/ end of the continuum increased the number of /ba/ judgments. The number of /bda/ judgments increased when a visible /ba/ was paired with an auditory syllable from the /da/ end of the continuum. Visible /da/ articulations increased the likelihood of /da/, /ða/, and /va/ responses. Although the Japanese speakers were also highly influenced by visible speech, they gave a different set of responses. These subjects mainly responded not only with /ba/ and /da/, but also gave frequent /wa/ and /za/ judgments. The latter two syllables are psychophysically similar to the former. The likelihood of a /ba/ judgment decreased as the visible stimulus went from the /ba/ end to the /da/ end of the continuum. This effect of the visible speech also occurred in the bimodal condition. Similarly, the likelihood of a /da/ judgment increased as the visible stimulus went from the /ba/ end to the /da/ end of the continuum. There were more /za/ responses for the visible speech at the /da/ end of the continuum. The number of /wa/ responses increased at the /ba/ end of the visible continuum.

These differences between Japanese and English speakers reflect the differences in the phonemic repertoires, phonetic realizations of the syllables, and phonotactic constraints in the two languages. Although different responses are given, speakers of both languages were influenced by visible speech. In addition, the contribution of one source was largest to the extent that the other source was ambiguous. The details of these judgments were nicely captured in the predictions of the FLMP, which gave a significantly better fit than did the ADM or the AMP. Given the outcome, the experiments substantiated the distinction made between information and information processing. The information made available by evaluation differs naturally for different languages. However, the information processing involved in integration and decision is identical across languages. Thus, these results provide some of the first findings that the FLMP provides a good account of bimodal speech perception in languages other than English.

Needless to say, the positive results from Japanese subjects does not warrant a generalization across all languages. As in cross-linguistic research in other domains (MacWhinney & Bates, 1989), it is necessary to test a variety of linguistic groups. In the present study, we continued this research with a new linguistic group from another linguistic family (Maddieson, 1984): Dutch speakers. We test the hypothesis that the Dutch will give a different repertoire of responses, but that the judgments will be accounted for by the FLMP. Of course, a viable alternative hypothesis is that the FLMP will fail. Like English, Spanish, and Japanese, the Dutch language also has /b/ and /d/ phonemes. Because of the cross-linguistic differences in the phonemes, however, the ideal auditory cues and visual cues specifying these phonemes differ across these languages. For example, the /d/ is somewhat more dental in Dutch. Also, the long-vowel /a/ in the stimuli is longer and more open in Dutch. The synthetic speech was modeled after American English and we might expect a less perfect match with Dutch than with English.

Another consideration in comparing perceivers of different languages is that the phoneme inventories differ across languages. In an open-ended task, subjects will naturally choose a native speech segment that gives the best fit to the auditory and visual information. For example, English has the phoneme /ð/, whereas Dutch does not. In the study of Massaro et al. (1993) English subjects often responded /ða/ when an auditory /ba/ was paired with a visual /da/. Since Dutch does not have a /ð/ phoneme, judgments of Dutch speakers will necessarily differ from those of English speakers.

Languages also differ in their phonotactic rules, the rules for combining phonemes into admissible words. Massaro et al. (1993) found that English subjects gave a consonant cluster answer /bda/ when an auditory /da/ was paired with a visual /ba/. The consonant cluster /bd/ exists in English and Dutch, but does not occur in the word-initial position. The cluster occurs primarily at morpheme boundaries—that is, a morpheme-final consonant followed by a morpheme-initial consonant. In Dutch but not in English, /bd/ is mostly pronounced as /pd/, owing to assimilation rules. We might therefore expect that Dutch subjects would be somewhat less likely to respond /bda/, given a visual /ba/ and an auditory /da/, showing a different influence from the visible speech.

A distinction has to be made between information and information processing. *Information* refers to just the output of the evaluation operation in the FLMP (see Figure 2). *Information processing* refers to the nature of the evaluation, integration, and decision operations, not to the input to or output from these operations. This study primarily addresses differences in information processing across different languages. Although perceivers of different languages might process speech in the manner described by the FLMP, a given level of auditory or visual information will not necessarily have equivalent effects across the different languages. Given the phonetic differences in the segments /ba/ and /da/ and the phonological differences across the languages, it is unlikely that a given speech stimulus will be identified equivalently. The hypothesis of no differences in information processing predicts only that the FLMP can accurately describe the results for speakers of different languages.

An alternative hypothesis predicts that the FLMP will fail to describe differences across language groups. For example, suppose that the Dutch are less influenced by visible speech in bimodal speech perception but have accurate identifications of visible speech (without the auditory signal); then, Equation 10, below, should fail. In this case, the FLMP should give a poor description of the results, because it cannot predict this type of selective weighting of one of the two sources of information.

The discussion concerning the FLMP and linguistic differences also applies equally to the ADM and AMP. These models also predict the information processing underlying speech perception, not the information available to a given speaker of a given language. For each speaker, the model only specifies how the information is processed, given unimodal and bimodal speech.

In comparing the two language groups, it is necessary to control for or account for the differences in phoneme inventories and phonotactic rules. The most direct method is to conduct an experiment in which subjects are limited to just two responses, /ba/ and /da/, similar to the task used in Massaro et al. (1993). The rationalization for this experiment will be described in the next section.

EXPERIMENT 1
Two Alternatives

According to the FLMP, the relative goodness rule (RGR) at decision predicts that performance should be a function of only the possible alternatives in the task (in this case, only /ba/ and /da/). Consider an auditory /da/ paired with a visual /ba/. Even though there would be significant support for different prototypes in the different languages, the probability of a /da/ judgment fo r all speakers is predicted to be

$$P(/\text{da}/|A_i \text{ and } V_j) = \frac{a_{Di} \times v_{Dj}}{a_{Di} \times v_{Dj} + (1-a_{Di}) \times (1-v_{Dj})}, \tag{10}$$

where a_{Di} is the auditory support for /da/ and v_{Dj} is the visual support for /da/. With just two alternatives, it is sufficient to assume in the FLMP that the support for /da/ is given by one minus the support for /ba/ (Massaro, 1987). In terms of Equation 10, it can be shown that the model makes equivalent predictions if a_{Bi} is assumed to be equal $1 - a_{Di}$ and v_{Bj} is assumed to be equal to $1 - v_{Dj}$ (Massaro, 1989a). Given the success of the FLMP with English, Spanish, and Japanese speakers, the hypothesis that all speakers process speech in the same manner predicts that the equation will give an equally good description of Dutch speakers.

Method

Subjects. Twenty native-Dutch speakers participated in this 2-h experiment as subjects. They were recruited by posted advertisements on the campus, and they were paid Hfl.20,–. The subjects were tested for normal hearing and had normal or corrected-to-normal vision. Nineteen of them were students from the Delft University of Technology in the Netherlands, and one was a nurse. Their ages ranged from 20 to 26 years (average, 22.6 years). Education in English as a foreign language started for most of the subjects (12) at age 12 in high school, but for some of them it began earlier, at age 10 (2) or 11 (6). The average duration of formal education in English was 6.6 years. The subjects also reported having various degrees of experience, ranging from low to high, in speaking and reading English.

Apparatus and Materials. The test stimuli were the audible and visible synthetic speech used by Massaro et al. (1993). With an auditory speech synthesizer, a continuum of five sounds was created to vary linearly between a good /ba/ and a good /da/. In an exactly analogous manner, by using computer animation, a synthesized face was programmed to say /ba/ and /da/ and also three intermediate syllables. Thus, a five-step visible continuum going from /ba/ to /da/ was created.

Synthetic audible speech. Tokens of the first author's /ba/ and /da/ were analyzed by using linear prediction to derive a set of parameters for driving a software formant serial resonator speech synthesizer (Klatt, 1980). By altering the parametric information specifying the first 80 msec of the consonant–vowel syllable, a set of five 400 msec syllables covering the range from /ba/ to /da/ was created. The center and lower panels of Figure 4 show how some of the acoustic synthesis parameters changed over time for the most /ba/-like and /da/-like of the five auditory syllables. During the first 80 msec, the $F1$ went from 250 to 700 Hz following a negatively accelerated path. The $F2$ followed a negatively accelerated path to 1199 Hz, beginning with one of five values equally spaced between 1187 and 1437 Hz from most /ba/-like to most /da/-like, respectively. The $F3$ followed a linear transition to 2729 Hz from one of five values equally spaced between 2387 and 2637 Hz. All other stimulus characteristics were identical for the five auditory syllables. Figure 5 gives the spectrograms of the five syllables along the continuum.

Synthetic visible speech. Like Parke (1974), we used a parametrically controlled polygon topology to generate a fairly realistic animation facial display (Cohen & Massaro, 1990). 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 6 shows a framework rendering of this model. To achieve a natural appearance, the surface was smooth shaded according to Gouraud's (1971) method (shown in the right panel of Figure 6). The face was animated by altering the location of various points in the grid under the control of 50 parameters, 11 of which were used for speech animation. Control parameters for several demonstration sentences were selected and refined by the investigator by studying his own articulation frame by frame and estimating the control parameter values (Parke, 1974). Each phoneme is defined in a table according to target values for segment

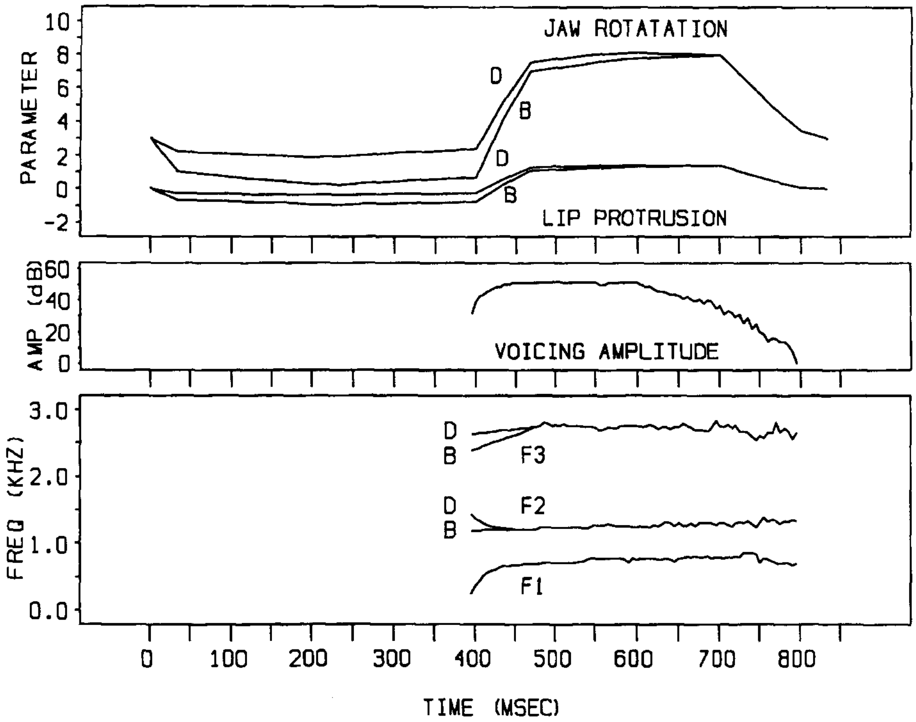

Figure 4. Visual and auditory parameter values over time for visual /ba/ and /da/ stimuli and auditory /ba/ and /da/ stimuli. Bottom panel shows formants $F1$, $F2$, and $F3$; middle panel shows voicing amplitude; and top panel shows jaw rotation and lip protrusion. See text for details.

duration, segment type (stop, vowel, liquid, etc.), and 11 control parameters. The parameters that are used are jaw rotation, mouth x scale, mouth z offset, lip corner x width, mouth corner z offset, mouth corner x offset, mouth corner y offset, lower lip "f" tuck, upper lip raise, and x and z teeth offset. Parke's software, revised by Pearce, Wyvill, Wyvill, and Hill (1986) and ourselves (Cohen & Massaro, 1990) was implemented on a Silicon Graphics Iris 3030 computer. We adapted the software to allow new intermediate test phonemes. To create an animation sequence, each frame was recorded with a broadcast quality Betacam video recorder under control of the Iris.

Figure 7 gives pictures of the facial model at the time of maximum stop closure for each of the five levels between /ba/ and /da/. Table 2 gives the parameter target values used in the visual synthesis for the consonant portion of each visual stimulus, the default resting parameter values, and the values for the vowel /a/. The top panel of Figure 4 shows how the visual synthesis parameters changed over time for the first (/ba/) and last (/da/) visual levels. For the sake of clarity, only two of the visual parameters are shown: jaw rotation (larger parameter means more open), and lip protrusion (Mouth z offset in Table 2; smaller number means more protrusion). Not shown in the figure, the face with the default parame-

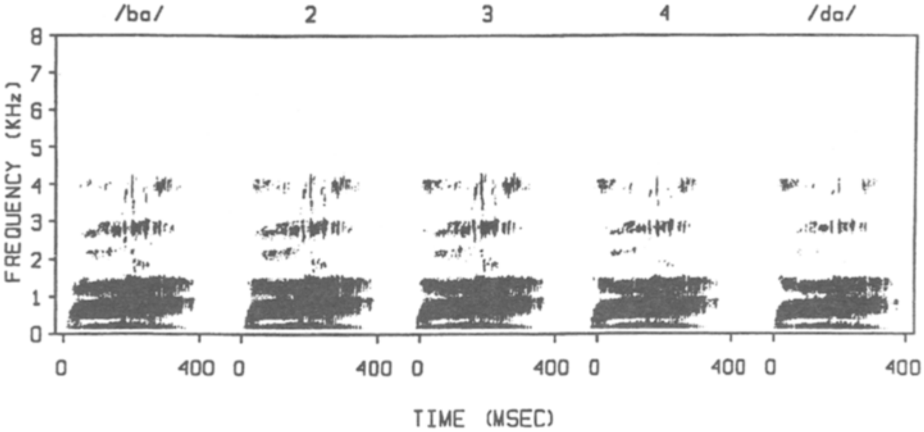

Figure 5. Spectrograms for the five levels of auditory speech between /ba/ and /da/.

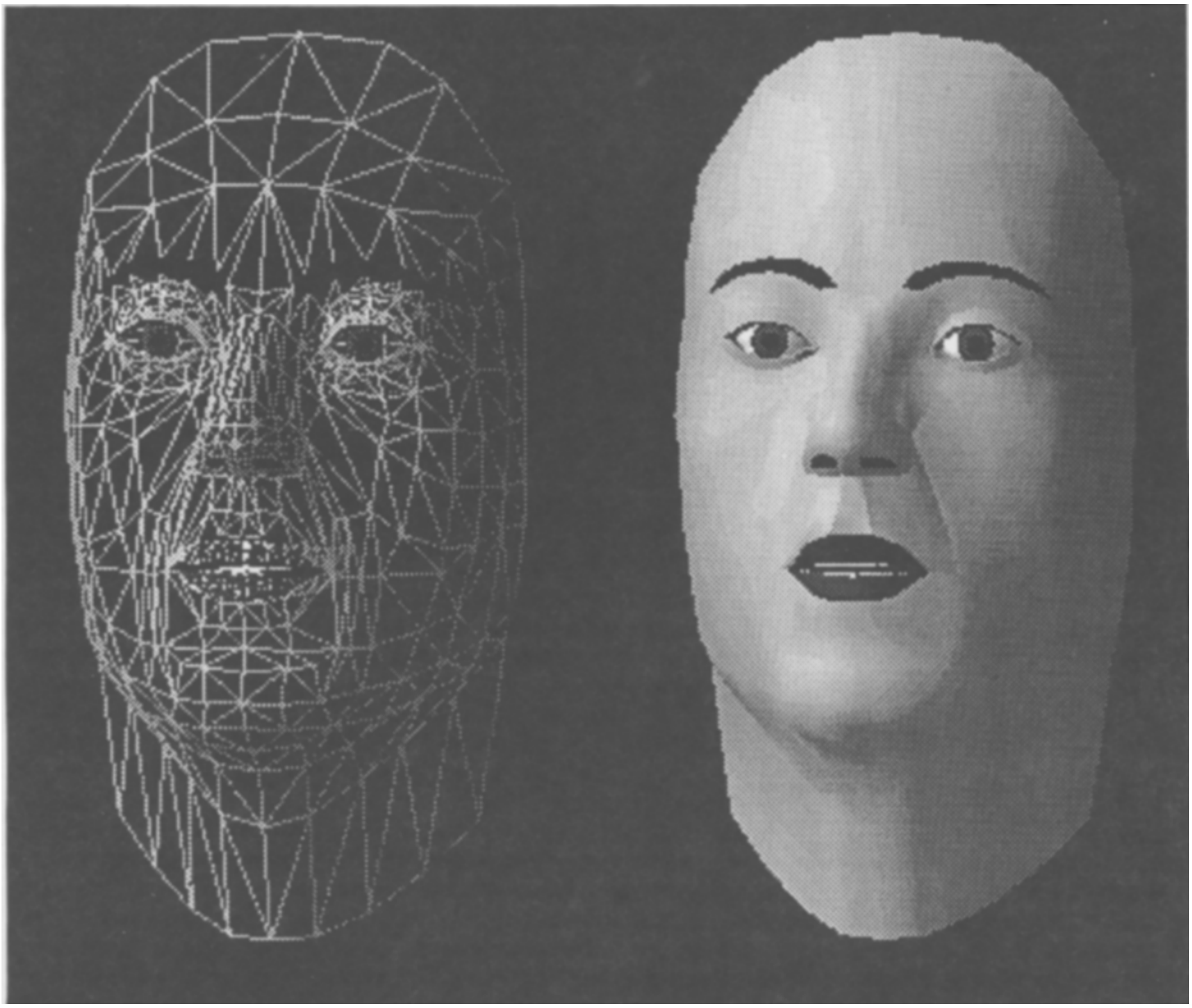

Figure 6. Framework (left) and Gouraud shaded (right) renderings of polygon facial model.

ter values was recorded for 2,000 msec preceding and 2,000 msec following the time shown for a total visual stimulus of 4,866 msec. A dark screen was presented for the auditory alone trials.

Following the synthesis, a Betacam tape was dubbed to 3/4-in. U-Matic for editing. Only the final 4,766 msec of each video sequence was used for each trial. A tone marker was dubbed onto the audio channel of the tape at the start of each syllable to allow the playing of the 400-msec auditory speech stimulus just following the consonant release of the visual stimulus. The marker tone on the video tape was sensed by a Schmidt trigger on a PDP-11/34A computer, which presented the auditory stimuli from digitized representations on the computer's disk. Figure 4 shows the temporal relationship between the auditory and visual parts of the stimulus. As can be seen in the figure, the parameter transitions specifying the consonantal release occurred at about the same time for both modalities.

Table 2
Visual Synthesis Parameters for the Five Stops, Default Position, and /a/

Parameter	Default	/b/	2	3	4	/d/	/a/
Jaw rotation	3.00	0.00	0.45	0.90	1.35	1.80	10.00
Mouth *x* scale	1.00	1.00	1.03	1.06	1.09	1.12	1.00
Mouth *z* offset	0.00	−1.00	−0.85	−0.70	−0.55	−0.40	2.00
Lip corner *x* width	0.00	0.00	0.75	1.50	2.25	3.00	20.00
Mouth corner *z* offset	0.00	−15.00	−15.00	−15.00	−15.00	−15.00	0.00
Mouth corner *x* offset	0.00	2.00	3.50	5.00	6.50	8.00	0.00
Mouth corner *y* offset	0.00	0.00	0.45	0.90	1.35	1.80	−5.00
Lower lip "f" tuck	0.00	−5.00	−5.00	−5.00	−5.00	−5.00	0.00
Upper lip raise	0.00	2.00	3.65	5.30	6.95	8.60	2.00

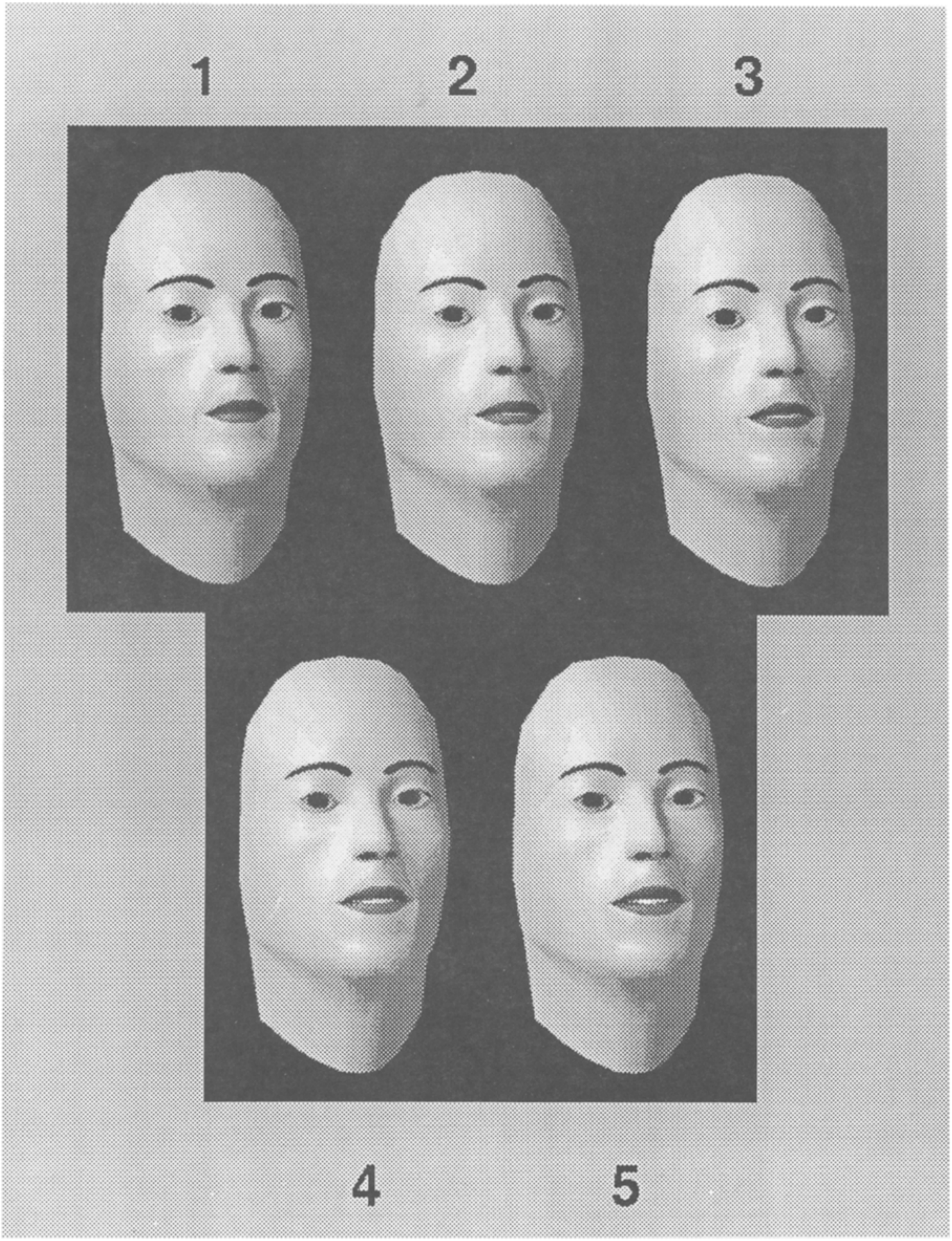

Figure 7. The facial model at the onset of the syllable for each of the five levels of visible speech between /ba/ and /da/.

Design and Procedure. In this experiment, synthetic auditory and visual speech were manipulated in the expanded factorial design previously illustrated in Figure 1. The onsets of the second and third formants were varied to give an auditory continuum between the syllables /ba/ and /da/. In analogous fashion, we systematically varied parameters of the facial model to give a continuum between visual /ba/ and /da/. Five levels of audible speech varying between /ba/ and /da/ were crossed with five levels of visible speech varying between the same alternatives. In addition, the audible and visible speech also were presented alone for a total of

25 + 5 + 5 = 35 independent stimulus conditions. Six random sequences were determined by sampling the 35 conditions without replacement, giving six different blocks of 35 trials. These trials were recorded on videotape for use in the experiments.

The subjects were instructed to listen to and to watch the speaker, and to identify the syllable as /ba/ or /da/. They all received 6 practice trials; the number of test trials was 840 (35 × 6 × 4). Thus there were 24 observations at each of the 35 unique experimental conditions. The subjects were given a short break after every 210 trials. The display monitor subtended a visual angle of 19.5°. The experimental tape was played on a video recorder (Panasonic S-VHS Hi-Fi NV-FS100EV HQ). The subjects sat at a distance of 1.46 m facing the monitor (Sony Triniton PVM-2130 QM, 52 cm). The loudness level of the auditory stimuli was 79 dB (A). The measurement was done with the sound-level meter (B&K 2231, with microphone Type 4133: Time Weighting, "Fast"; Frequency Weighting, "A"; Display Parameter, "SPL"). The background noise level was 47.5 dB (A).

The subjects were tested individually in a normally illuminated room. They gave their answers by marking either "ba" or "da" on a prepared answer sheet. In all cases, the experimenter was a native speaker of Dutch and all instructions and interactions were in the Dutch language. The subjects were not told that the stimuli were based on the American English language.

Results

The subjects' forced-choice response identifications were recorded for each stimulus. The mean observed proportion of identifications was computed for each subject for the unimodal and bimodal conditions. Separate analyses of variance were carried out on the auditory, visual, and bimodal conditions. Both the auditory and the visual sources of information had a strong impact on the identification judgments. The points in Figure 8 give the observed proportion of /da/ responses for the auditory alone (left plot), the bimodal (middle plot), and the visual alone (right plot) conditions as a function of the five levels of the synthetic auditory and visual speech varying between /ba/ and /da/. As illustrated in the figure, the proportion of responses changed systematically across the visual continuum, for both the unimodal [$F(4,76) = 208.59$, $p < .001$] and the bimodal [$F(4,76) = 39.98, p < .001$] conditions. Similarly, the pattern of responses changed in an orderly fashion across the auditory continuum, for both the unimodal [$F(4,76) = 440.36$, $p < .001$] and the bimodal [$F(4,76) = 128.48$, $p < .001$] conditions. Finally, the auditory and visual effects were *not* additive in the bimodal condition, as demonstrated by the significant auditory–visual interaction on response probability [$F(4,76) = 17.63, p < .001$].

Relative influence of visible and audible speech. One question of interest comprises the relative contributions of visible and audible speech in the bimodal condition. An index of the magnitude of the effect of one modality can be described by the difference in response probabilities for the two endpoint stimuli that are from that modality. This difference was computed for each subject for each level for both audible and visible sources of information. As an example, given some auditory level, a .9 probability of /da/, given the visual /da/ endpoint stimulus, and an overall .2 probability of /da/, given the visual /ba/ endpoint stimulus, would give a visual effect of .7. In Figure 9, the size of the mean visual effect is plotted

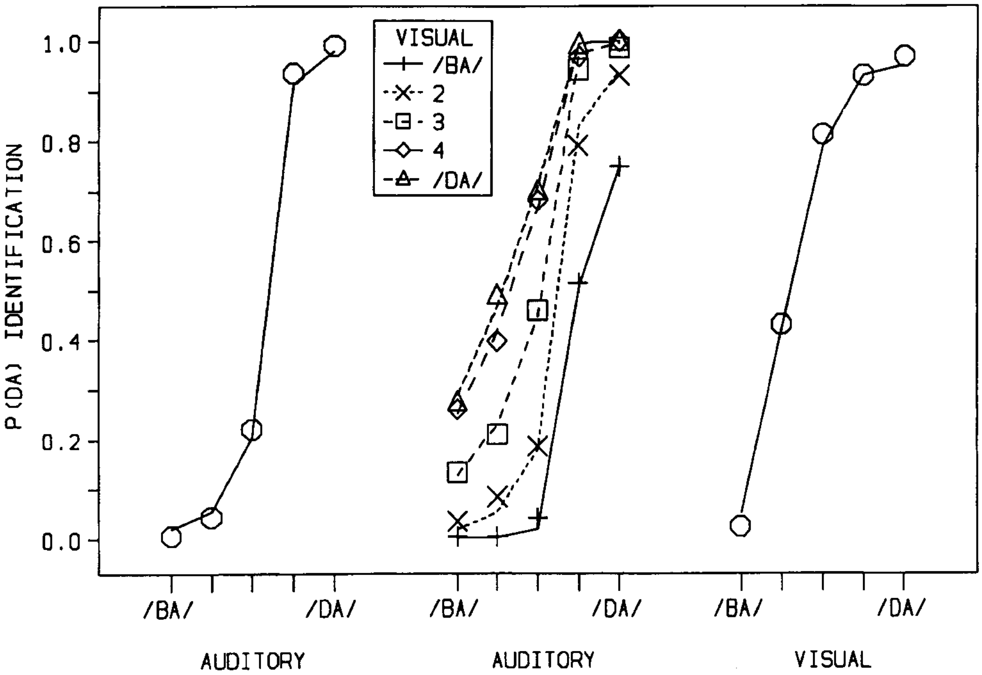

Figure 8. Observed (points) and predicted (lines) probability of a /da/ response for the auditory alone (left plot), bimodal (middle plot), and visual alone (right plot) conditions as a function of the five levels of the synthetic auditory (AUDITORY) and visual (VISUAL) speech varying between /ba/ (BA) and /da/ (DA) for the Dutch speakers. Predictions are for the FLMP model.

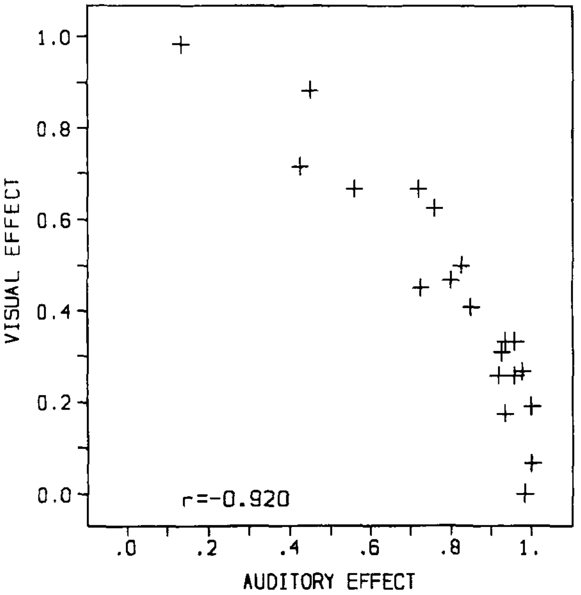

Figure 9. Mean visual effect as a function of the mean auditory effect for each of the subjects in the bimodal condition for the Dutch speakers. The *r* value gives the correlation between the two effects.

as a function of the size of the mean auditory effect. Audible speech had a larger influence than did visible speech. The size of the effects varied significantly across subjects. In addition, there was a strong negative correlation (given in the plot of Figure 9) between the two effects. To the extent that one modality had a large effect, the other had a small effect.

Tests of the models. The FLMP, ADM, and AMP were fit to the individual results from each of the 20 subjects. The quantitative predictions of the model are determined by using the program STEPIT (Chandler, 1969). A model is represented to the program in terms of a set of prediction equations and a set of unknown parameters. By iteratively adjusting the parameters of the model, the program minimizes the squared deviations between the observed and predicted points. The outcome of the program STEPIT is a set of parameter values which, when put into the model, come closest to predicting the observed results. Thus, STEPIT maximizes the accuracy of the description of a given model. We report the goodness-of-fit of a model by the root mean square deviation (RMSD)—the square root of the average squared deviation between the predicted and observed values.

The points in Figure 8 give the observed proportion of /da/ responses for the auditory alone (left plot), audiovisual (middle plot), and visual alone (right plot) conditions as a function of the five levels of the synthetic auditory and visual speech varying between /ba/ (BA) and /da/ (DA) for the Dutch speakers. The continuous lines in Figure 8 give the average predictions of the FLMP. The FLMP was fit to the results by estimating five a_{Di} and five v_{Dj} values for the five levels of each modality. The a_{Bi} and v_{Bj} values were set to one minus these values. The FLMP model provides a good description of the identifications of both the unimodal and the bimodal syllables (average RMSD value was 0.0409).

Table 3 gives the average best-fitting parameters of the FLMP. These parameter values index the degree of support for each response alternative by each level of the audible and visible stimuli. As can be seen in Table 3, the parameter values change in a systematic fashion across the five levels of the audible and visible synthetic speech. For both modalities, the support for the alternative /da/ increases systematically from the /ba/ to /da/ level along the continuum. The larger spread for the auditory than for the visual parameters indicates a larger influence for auditory than for visual speech. The same result is apparent in Figure 9.

It should be noted that the predicted points for the FLMP shown in Figure 8 cannot be recovered from the parameter values shown in Table 3. The figure and table give the predicted points and parameter values averaged across the model fits of the individual subjects. Given that the FLMP is nonlinear, the average predictions cannot be exactly computed from the average parameter values.

To fit the ADM to the results, each unique level of the auditory stimulus requires 2 unique parameters, a_b and a_d, for each of the five levels along the auditory continuum. Two free parameters are necessary for each of the five levels along the visual continuum. Finally, an auditory bias parameter, w_b, is necessary, for a total of 21 parameters for the ADM. The continuous lines in Figure 10 give the average ADM predictions of the observed results. The average RMSD value was 0.0797. An analysis of variance of the RMSD values showed that the FLMP, with just 10 free parameters, gave a significantly better description of the results [$F(1,19) = 22.34$, $p < .001$].

A test of the AMP also allows a test of whether the inputs are added or combined in an additive or a nonadditive manner. To fit the AMP to the results, each unique level of the auditory stimulus requires a unique parameter a_{Bi}, and analogously for v_{Bj}. The modeling of /da/ responses thus requires 5 auditory parameters plus 5 visual parameters. The p value would be fixed across all conditions, for a total of 11 parameters. Thus, we have a fair comparison to the FLMP, which requires 10 parameters.

The AMP was fit to the individual results in the same manner as in the fit of the FLMP. The predicted lines in Figure 11 show that the AMP gave a poor description of the observed results. The average RMSD was 0.1062.

Table 3
Average Best-Fitting Parameters for the FLMP Model for the Dutch Speakers

	Level				
Modality	B	2	3	4	D
Visual	0.0491	0.4194	0.7889	0.9315	0.9509
Auditory	0.0200	0.0547	0.2026	0.9148	0.9793

Note—The values index the degree of support for the alternative /da/. The support for /ba/ is 1 minus each value.

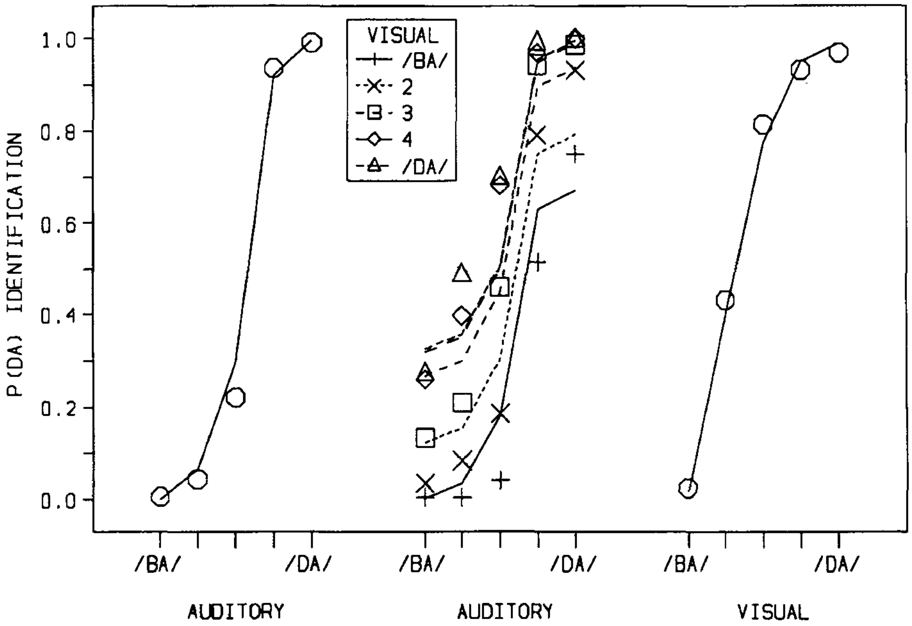

Figure 10. Observed (points) and predicted (lines) probability of a /da/ response for the auditory alone (left plot), bimodal (middle plot), and visual alone (right plot) conditions as a function of the five levels of the synthetic auditory (AUDITORY) and visual (VISUAL) speech varying between /ba/ (BA) and /da/ (DA) for the Dutch speakers. Predictions are for the ADM model.

Given RMSDs for each subject and each model, it is reasonable to test for statistically significant differences among the models. A single-factor analysis of variance (with the FLMP vs. the AMP as the factor) on the individual-subject RMSD values showed that the FLMP gave a significantly better description of the results than did the AMP [$F(1,19) = 69.79, p < .001$].

The good fit of the FLMP relative to the AMP is evidence against additive integration. The integration of the multiple sources appears to follow a multiplicative com-

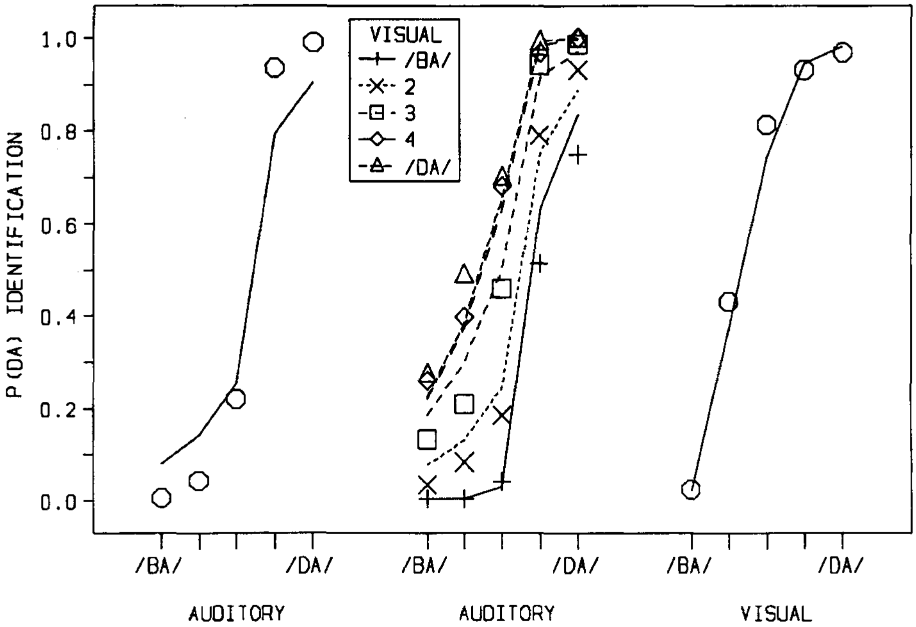

Figure 11. Observed (points) and predicted (lines) probability of a /da/ response for the auditory alone (left plot), bimodal (middle plot), and visual alone (right plot) conditions as a function of the five levels of the synthetic auditory (AUDITORY) and visual (VISUAL) speech varying between /ba/ (BA) and /da/ (DA) for the Dutch speakers. Predictions are for the AMP model.

bination, with the result that the less ambiguous source has the larger impact on performance. Given that the FLMP is mathematically equivalent to Bayes's theorem—an optimal algorithm for integrating multiple sources of information—the good fit of the model to the present results is evidence for optimal speech recognition by humans (Massaro, 1987, 1989a; Massaro & Friedman, 1990).

EXPERIMENT 2
Open-Ended Response Task

We now ask whether similar results will be found when subjects are permitted an open-ended set of response alternatives. An open-ended response task is of interest for several reasons. First, the nature of the responses is of interest. Research with English, Spanish, and Japanese speakers (Massaro et al., 1993) has shown that subjects respond with alternatives from their native language. Will the same hold true for the Dutch speakers? Second, will the FLMP continue to give a better description than the other models when speakers are given a completely open-ended set of response alternatives? Finally, it is of interest whether visible speech will still have an important influence when responses are not constrained. In addition to Dutch subjects, we tested English speakers in the same task. English speakers were tested because all of the earlier experiments had actually limited the number of responses to eight alternatives derived from pilot studies. We can expect that the response alternatives given by the Dutch and English speakers will reflect differences in the phoneme inventories and the phonological structure of the two languages. Therefore, a direct comparison of Dutch and English performance will not permit us to distinguish between information and information processing. Tests of the models, however, will indicate whether or not Dutch and English speakers process speech in a similar manner. Furthermore, comparison of the performance of English speakers in the present task with performance in an earlier task permitted us to observe changes in performance due to the number of specified alternatives.

Method

The Dutch speakers were 10 students of the Delft University of Technology in the Netherlands. None of them had participated in the two-alternative experiment. Their ages ranged from 19 to 24 years (average, 21.2). These subjects were tested for normal hearing and had normal or corrected-to-normal vision. They received Hfl.10,– for participating. The English speakers were 9 students of the University of California in Santa Cruz. The ages of these subjects ranged from 17 to 20 years (average, 18.8). None of them reported any hearing or seeing loss. Their participation was one of the options to fill a course requirement.

All procedural details were the same as those described for the two-choice task, except that now subjects were allowed to give any possible response. The stimulus tapes from Experiment 1 were used. Each of the 35 possible syllables was presented a total of 12 times during two sessions, and the subject identified each stimulus with a written response during a 3-sec response interval. Prior to presentation of the experimental stimuli, the subjects were given 6 practice trials to familiarize them with the task. The subjects were given a short break of approximately 5 min after completing the tape of 210 trials. Unknown to the subjects, the tape was rewound and played again, repeating the 210 trials.

Results

The Dutch subjects gave a variety of responses—for example, 24 alternatives. Voiceless responses occurred on 4.24% of the trials. These responses were pooled with their voiced cognates. In addition to /ba/ and /da/, /*v*a/, /za/, /va/, /*v*ha/, and /ma/ were the most frequent responses. After the voiceless judgments were pooled, these alternatives accounted for 98.84% of the responses. The remaining alternatives were placed in an "other" category. The points in Figure 12 give the observed proportion of responses for the visual alone (left plot), auditory alone (second plot), and bimodal (remaining plots) conditions as a function of the five levels of the synthetic auditory and visual speech varying between /ba/ (B) and /da/ (D) for the Dutch speakers. The proportion of /ba/ and /da/ judgments changed systematically in the expected direction across the levels of the visual and auditory continua. Synthetic speech at the /ba/ end of the visible speech continuum also gave some support for the alternative /va/. The number of /za/ responses increased somewhat at the /da/ end of the auditory continuum. There were very few /ma/ and /ha/ responses. On the other hand, the number of /ma/ judgments slightly increased when there was auditory /ba/ information only. These bimodal judgments reflect the contribution of both auditory and visual speech. Furthermore, the judgments are more or less in line with the psychophysical similarity between the test stimuli and the response alternatives.

The points in Figure 13 give the observed proportion of responses for the visual alone (left plot), auditory alone (second plot), and bimodal (remaining plots) conditions as a function of the five levels of the synthetic auditory and visual speech varying between /ba/ (B) and /da/ (D) for the English speakers. The English speakers gave 29 alternatives. Voiceless responses occurred on 22.54% of the trials. These responses were pooled with their voiced cognates. After this pooling, the /ba/, /da/, /va/, /ða/, /za/, /bda/, and /la/ judgments accounted for 97.12% of the responses. The remaining alternatives formed the "other" category. Surprisingly, there were not many /ba/ and /da/ judgments in the unimodal conditions. The audible and visible speech tended to support the alternatives /va/, /ða/, and /za/. Contributions of both audition and vision can be observed in the responses of the English speakers. Visual information (especially the second and third levels) paired with the /ba/ end of the auditory continuum increased the number of /va/ judgments. The proportion of /ða/ responses was largest at the middle levels of the auditory continuum. More /za/ judgments were given when visible /da/ was paired with an auditory syllable from any level of the continuum, except the /da/ end. Visual /ba/ increased the number of /bda/ responses only at the /da/ end of the auditory continuum.

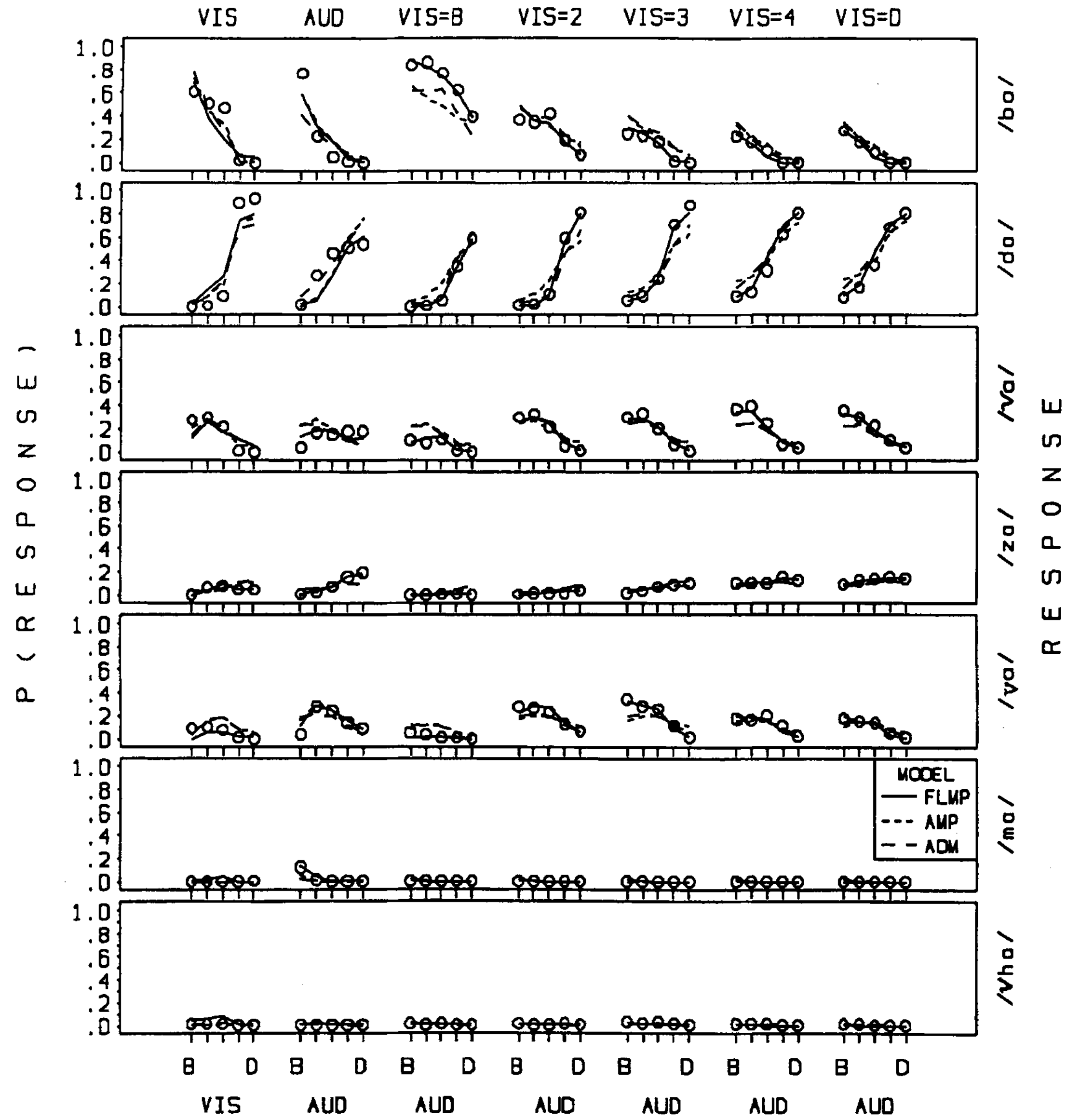

Figure 12. Observed (points) and predicted (lines) proportion of /ba/, /da/, /va/, /za/, /va/, /ma/, and /vha/ identifications for the visual alone (left plot), auditory alone (second plot), and bimodal (remaining plots) conditions as a function of the five levels of the synthetic auditory (AUD) and visual (VIS) speech varying between /ba/ (B) and /da/ (D) for the Dutch speakers. The lines give the predictions of the FLMP, AMP, and ADM.

Model Tests

The FLMP, ADM, and AMP were tested against the results. The predictions of the FLMP are given by Equations 3, 4, and 5. The fit of the FLMP requires $N(a + v)$ parameters where N is the number of response alternatives with a levels of auditory and v levels of visual stimuli. Thus, 80 free parameters are necessary with 8 alternatives (7 plus the "other" category). The number of free parameters for the fit of the ADM also depends on the number of alternatives N. For the ADM, $(N - 1)a + (N - 1)v + (N - 1)$ free parameters are necessary. With 7 specific alternatives and the "other" category, $(8 - 1)5 + (8 - 1)5 + (8 - 1) = 77$ free parameters were used in the model test. For the AMP, $N(a + v) + 1$ parameters are necessary, or $8(5 + 5) + 1 = 81$.

The lines in Figure 12 give the predictions of the FLMP, AMP, and ADM for the Dutch speakers. The average RMSD values for the fit of the FLMP, AMP, and ADM were 0.0734, 0.1085, and 0.1025, respectively. Analyses of variance were carried out on the RMSD values of the different models. The FLMP gave a significantly better fit than did both the AMP and the ADM [$F(1,9) = 38.1, p < .001$ (FLMP vs. AMP), and $F(1,9) = 13.2, p = .006$ (FLMP vs. ADM)]. The details of the judgments are best captured in the predictions of the FLMP. Table 4 gives the average best-fitting parameters of the

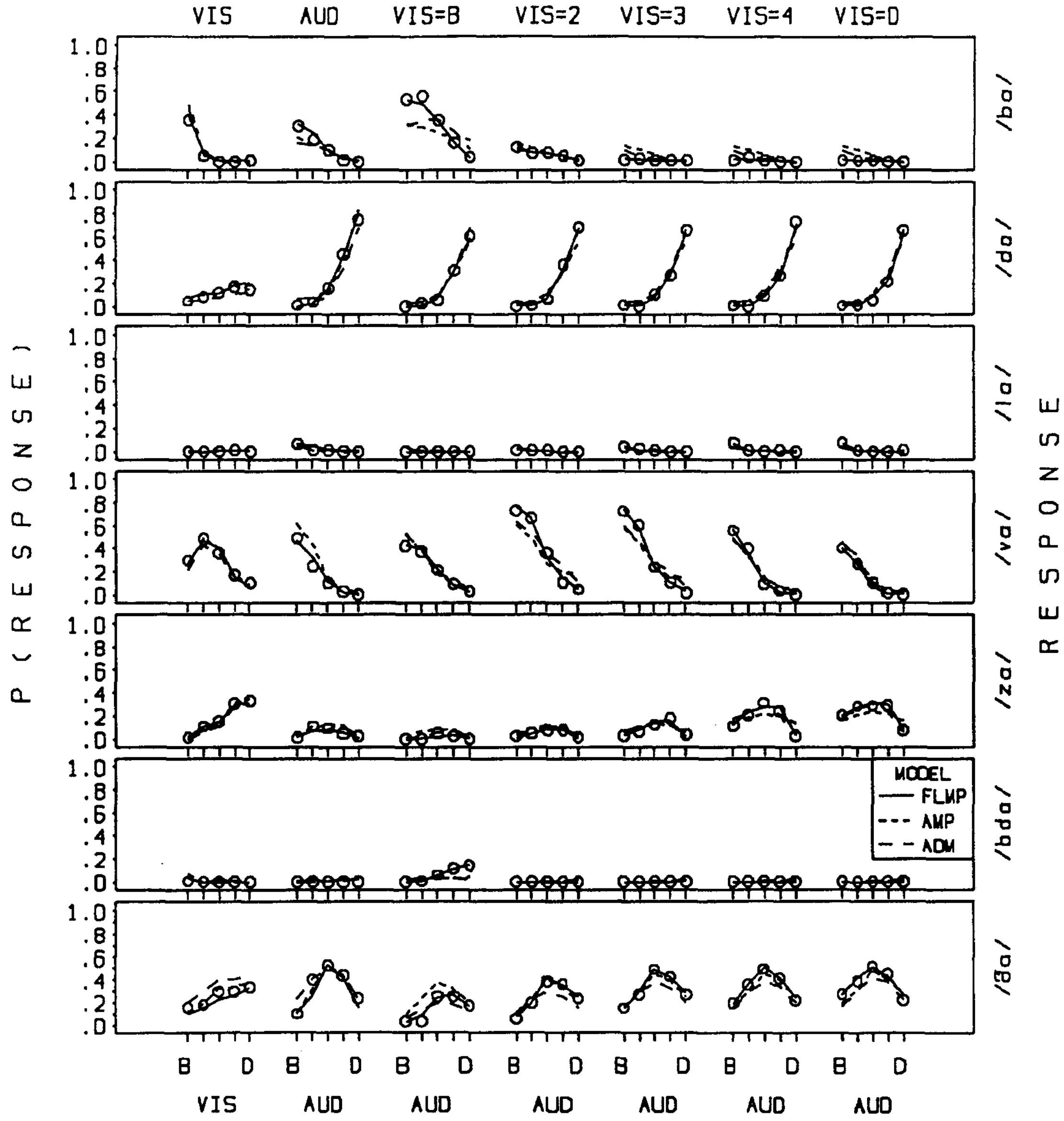

Figure 13. Observed (points) and predicted (lines) proportion of /ba/, /da/, /la/, /va/, /za/, /bda/, and /ða/ identifications for the visual alone (left plot), auditory alone (second plot), and bimodal (remaining plots) conditions as a function of the five levels of the synthetic auditory (AUD) and visual (VIS) speech varying between /ba/ (B) and /da/ (D) for the English speakers. The lines give the predictions of the FLMP, AMP, and ADM.

FLMP for the Dutch speakers. These parameter values index the degree of support for each response alternative by each level of the audible and visible stimuli. As can be seen in the table, the parameter values change in a systematic fashion across the five levels of the audible and visible synthetic speech. These parameter values account for relative contribution of the audible and visible speech to the judgments shown in Figure 12.

The lines in Figure 13 give the predictions of the FLMP, AMP, and ADM for the English language group. The average RMSD values for the fit of the FLMP, AMP, and ADM to the English speakers were 0.0512, 0.0926, and 0.0910, respectively. Analyses of variance were carried out on the RMSD values of the different models. The FLMP gave a better description of the data than did both the AMP and the ADM [$F(1,8) = 25.4, p = .001$ (FLMP vs. AMP) and $F(1,8) = 24.5, p = .001$ (FLMP vs. ADM)]. Table 5 gives the average best-fitting parameters of the FLMP for the English speakers.

DISCUSSION

One goal of the present study was to broaden the domain of inquiry in bimodal speech perception by evaluating the performance of Dutch speakers. Dutch speakers had not previously been tested on the expanded

Table 4
Average Best-Fitting Parameters for the FLMP Model for the Dutch Speakers With Open-Ended Responses

		Response							
Modality	Level	/ba/	/da/	/*v*a/	/za/	/va/	/*v*ha/	/ma/	"Other"
Visual	B	0.9480	0.0506	0.2472	0.0898	0.0027	0.0183	0.1198	0.1266
	2	0.5532	0.3875	0.4684	0.1094	0.1267	0.0686	0.1700	0.1654
	3	0.3344	0.6757	0.3309	0.3029	0.1177	0.1356	0.2584	0.2539
	4	0.0680	0.9971	0.1548	0.0708	0.0335	0.0013	0.0100	0.0151
	D	0.0666	0.9972	0.0709	0.0862	0.0224	0.0001	0.0047	0.0341
Auditory	B	0.7872	0.0294	0.2202	0.0021	0.1813	0.2002	0.0238	0.0014
	2	0.5787	0.0873	0.4448	0.0985	0.5390	0.1142	0.0625	0.1056
	3	0.2774	0.3940	0.3379	0.1268	0.4110	0.0034	0.0222	0.0491
	4	0.0518	0.5762	0.1923	0.2027	0.2518	0.0114	0.0080	0.0072
	D	0.0106	0.6774	0.1640	0.2042	0.1153	0.0038	0.0037	0.0092

factorial design with synthetic and animated speech. In the first experiment, Dutch subjects identified the syllables as /ba/ or /da/, which allowed a direct comparison with an earlier experiment with English, Spanish, and Japanese speakers (Massaro, Tsuzaki, Cohen, Gesi, & Heredia, 1993). Speakers of different languages, when asked for a response in an open-ended task, will give judgments corresponding to segments that occur in their native language. Because of the differences in the phonemic repertoire, we can expect to obtain different judgments from different language groups. A previous result consistent with this expectation is that English and Japanese speakers gave different judgments on a synthetic /ba/–/da/ continuum. The English subjects gave /ba/, /da/, /bda/, /ða/, /va/, /dba/, and /ga/ judgments, whereas Japanese subjects gave /ba/ and /da/, /ga/, /wa/, and /za/ responses. When different sets of response alternatives are used, a direct comparison between two languages is difficult. However, by limiting the subjects to only two alternatives—for example, /ba/ and /da/—and by testing models on the experimental data, one can make a comparison. The two-alternative task proved to be useful in our previous cross-linguistic study, and it was also used in the Experiment 1 with Dutch speakers.

In the two-alternative task, the results of the Dutch speakers indicated significant contributions of auditory and visual speech. Similar to the results from other languages, the contribution of one source of information was larger to the extent that the other source was ambiguous. Three different models of how auditory and visible sources of information are processed were tested against the results from the two-choice task. Given the results and interpretation of the Sekiyama and Tohkura (1993) study, it is important to test a model that assumes that the contribution of visible speech is dependent on poor-quality audible speech. This hypothesis was formulated in terms of a model in which the perceiver uses just auditory speech on auditory trials and visual speech on visual trials. On bimodal trials, the perceiver either identifies the auditory information, or else bases the decision on the visual information when the auditory information is ambiguous. The results were also used to test an additive model of speech perception in which the auditory and visual sources are linearly combined. These two models were contrasted with the FLMP, in which multiple sources of continuous information are evaluated and integrated in speech perception. The outcome of the model tests provided unambiguous support for the FLMP description of the Dutch speakers in the two-alternative task.

To allow a more convincing conclusion and to provide a stronger test of the models, a second experiment was carried out with the same stimuli but without specifying any response alternatives. The subjects were free to identify the stimuli as any possible alternative. In addition to the Dutch subjects, a group of native English speakers

Table 5
Average Best-Fitting Parameters for the FLMP Model for the English Speakers With Open-Ended Responses

		Response							
Modality	Level	/ba/	/da/	/la/	/va/	/za/	/bda/	/ða/	"Other"
Visual	B	0.6213	0.1675	0.0072	0.4899	0.0406	0.0516	0.1993	0.2071
	2	0.1611	0.2091	0.0036	0.7078	0.1690	0.0030	0.2679	0.1862
	3	0.0224	0.1864	0.0218	0.6144	0.2386	0.0263	0.3886	0.1345
	4	0.0169	0.3537	0.0139	0.3022	0.4650	0.0140	0.4748	0.1645
	D	0.0085	0.2904	0.0192	0.1146	0.4483	0.0034	0.4979	0.1639
Auditory	B	0.5703	0.0044	0.1436	0.6871	0.0334	0.0105	0.1711	0.0152
	2	0.4405	0.0610	0.0732	0.6175	0.1654	0.0363	0.4129	0.0223
	3	0.1749	0.2573	0.0365	0.2195	0.1767	0.0050	0.7296	0.0242
	4	0.0575	0.6930	0.0096	0.0368	0.1014	0.0271	0.5928	0.0163
	D	0.0053	0.9803	0.0152	0.0137	0.0081	0.0230	0.2441	0.0191

was tested because the previous studies had actually limited the English-speaking subjects to eight specified alternatives. As in the two-alternative task, the Dutch and English subjects were influenced by both auditory and visual sources of information. Because the two language groups differed with respect to the alternatives that were used in the identification judgments, a comparison between the two language groups must be theoretically motivated. To this end, the same three models tested in the two-alternative task were tested in this open-ended task. The FLMP provided a significantly better description of the identifications of the individuals for both groups of subjects.

One might argue that the English/Dutch comparison is weakened by the substantial English experience of the native Dutch speakers. Subjects with this experience were unavoidable, because of universal schooling in English in the Netherlands. Similarly, in this age of telecommunications, it is difficult to imagine individuals without some experience in English. We do not believe that this English experience substantially influenced our results, for two important reasons. First, the experiment was carried out entirely in Dutch, and there is little evidence that a second language acquired after adolescence influences first-language processing. Second, and most importantly, substantial differences *were* found between the English and Dutch subjects in terms of the responses used in the experiment. The Dutch subjects responded with valid alternatives in Dutch and did not give English alternatives such as /ða/. With respect to our distinction between information and information processing, there were substantial differences in information but no differences in information processing. In this respect, these results are analogous to those of the cross-linguistic research on sentence processing carried out in the framework of the competition model (MacWhinney & Bates, 1989).

One caveat is that our conclusions of no information-processing differences are limited to the case of response decision. Temporal measures of the dynamics of information processing might reveal some differences. However, our research has shown that there is strong correlation between reaction time in this task and the probability of a given decision (Massaro, 1987). In a forced choice task, reaction times are slow to the extent that the two responses are about equally likely, and they speed up as one of the responses becomes more likely. We have predicted this result in terms of the ambiguity of the stimulus event. We expect that it will be ambiguity, and not language, that will predict the time course of information processing in speech perception.

Given that the FLMP was also found to give superior descriptions of Spanish and Japanese speakers in the Massaro et al. (1993) study, speakers of these languages appear to process bimodal speech in fundamentally the same manner. Although there are significant differences in the languages, it appears that the underlying mechanisms for speech perception are similar for the four studied languages. In future work, a greater variety of languages should be tested so that we may draw more general conclusions about human bimodal speech perception. For example, future cross-linguistic research can include more exotic languages, such as African languages and Asian tone languages. To this end, we will be happy to make a videotape of our stimuli available to investigators who wish to test additional language groups.

It should be stressed that we accept that there are cross-linguistic differences in the outcome of speech perception. Languages have different phonologies, and meaningful segments in different languages occupy different positions in articulatory space (Lindau & Ladefoged, 1986). It follows that languages will differ in the intelligibility of their auditory and visual segments. Consider the syllables /ba/ and /da/ as an example. English, but not Japanese, has /va/ and /ða/ syllables that are psychoacoustically similar to /ba/ and /da/. For this reason, auditory /ba/ and /da/ are less discriminable or intelligible in English than they are in Japanese. Given the tradeoff between two sources of information in the FLMP, we expect a larger visual influence in English than in Japanese. This difference occurs even though the information processing is identical in the two languages.

The replication of the experiment with English speakers illuminates the influence of specifying response alternatives in advance. In general, subjects will choose a wider variety of response alternatives when the alternatives are left unspecified. On the other hand, for a given stimulus, an individual might select a response alternative that has been given as a possible alternative in the task but not otherwise. Significantly more /bda/ judgments were given when /bda/ was given as one of the eight possible response alternatives in Massaro et al.'s (1993) study than in the present study, in which the response alternatives were unspecified. Table 6 gives the proportions of different responses in the two studies. Also, a small proportion of /bda/ and /ga/ judgments was given when these were specified alternatives in the eight-alternative task, but not in the completely open-ended task. Even so, the differences between the two methods were quantitative and not qualitative. As can be seen in the table, the two response profiles bear a great deal of similarity. Somewhat more /bda/, /dba/, and /ga/ responses were observed when these were specified as

Table 6
Weighted Average Proportion of Response Alternatives in the Completely Open-Ended (Present Study) and in the Eight-Alternative Task (Massaro et al., 1993) for the English Speakers

Response	Open-Ended	Eight Alternatives
/ba/	0.09179	0.12879
/da/	0.20426	0.22126
/va/	0.25423	0.29606
/bda/	0.01006	0.06644
/ða/	0.28514	0.18931
/la/	0.01141	
/za/	0.11431	
/dba/	0.00000	0.02901
/ga/	0.00056	0.02471
"Other"	0.02880	0.04436

possible alternatives, and /la/ and /za/ were given in the truly open-ended study. We can conclude that the nature of speech processing is not changed in any fundamental way by constraining the number of possible response alternatives. Similarly, the advantage of the FLMP in the two-alternative, eight-alternative, and completely open-ended study supports the idea that the manner of processing bimodal speech is not dependent on the number of alternatives used in an experiment.

REFERENCES

Binnie, C. A., Montgomery, A. A., & Jackson, P. L. (1974). Auditory and visual contributions to the perception of selected English consonants for normally hearing and hearing-impaired listeners. In H. Birk Nielsen & E. Kampp (Eds.), *Visual and audio-visual perception of speech* (*Scandinavian Audiology*, 4[Suppl.], 181-209). Stockholm: Almquist & Wiksell.

Breeuwer, M., & Plomp, R. (1984). Speechreading supplemented with frequency-selective sound-pressure information. *Journal of the Acoustical Society of America*, **76**, 686-691.

Campbell, R., & Dodd, B. (1980). Hearing by eye. *Quarterly Journal of Experimental Psychology*, **32**, 85-99.

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

Cohen, M. M. (1984). *Processing of visual and auditory information in speech perception*. Unpublished doctoral dissertation, University of California, Santa Cruz.

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

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**, 362-381.

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

Green, K. P., & Kuhl, P. K. (1989). The role of visual information in the processing of place and manner features in speech perception. *Perception & Psychophysics*, **45**, 34-42.

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

Lindau, M., & Ladefoged, P. (1986). Variability of feature specifications. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability of speech processes* (pp. 464-478). Hillsdale, NJ: Erlbaum.

MacWhinney, B., & Bates, E. (Eds.) (1989). *The crosslinguistic study of sentence processing*. New York: Cambridge University Press.

Maddieson, I. (1984). *Patterns of sounds*. Cambridge: Cambridge University Press.

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

Massaro, D. W. (1988). Ambiguity in perception and experimentation. *Journal of Experimental Psychology: General*, **117**, 417-421.

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: North-Holland.

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*, **9**, 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. (1993a). The paradigm and the fuzzy logical model of perception are alive and well. *Journal of Experimental Psychology: General*, **122**, 115-124.

Massaro, D. W., & Cohen, M. M. (1993b). Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables. *Speech Communication*, **13**, 127-134.

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, Department of Computer Science.

Parke, F. I. (1975). A model for human faces that allows speech synchronized animation. *Computers & Graphics Journal*, **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. In *Proceedings of Graphics Interface '86* (pp. 136-140).

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

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97-113). Hove, U.K.: Erlbaum.

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

Smeele, P. M. T., & Sittig, A. C. (1991a). The contribution of vision to speech perception. In *Proceedings of the 2nd European Conference on Speech Communication and Technology, Eurospeech 91* (pp. 1495-1497).

Smeele, P. M. T., & Sittig, A. C. (1991b). Effects of desynchronization of vision and speech on the perception of speech: Preliminary results. In *CCITT Brazil Conference Sept. '91* (Stgrp. XII, Wp. XII/2 and XII/3, Contribution D.81).

Smeele, P. M. T., Sittig, A. C., & van Heuven, V. J. (1992). Intelligibility of audio-visually desynchronised speech: Asymmetrical effect of phoneme position. *Proceedings of the International Conference on Spoken Language Processing 92*, **1**, 65-68.

Studdert-Kennedy, M. (1989). Reading gestures by light and sound. In A. W. Young & H. D. Ellis (Eds.), *Handbook of research on face processing* (pp. 217-222). Amsterdam: North-Holland.

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

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.

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

(Manuscript received December 9, 1993;
revision accepted for publication April 4, 1994.)