Perception & Psychophysics
2004, 66 (5), 820-836

Mandarin speech perception by ear and eye follows a universal principle

TREVOR H. CHEN and DOMINIC W. MASSARO
University of California, Santa Cruz, California

In this study, the nature of speech perception of native Mandarin Chinese was compared with that of American English speakers, using synthetic visual and auditory continua (from /ba/ to /da/) in an expanded factorial design. In Experiment 1, speakers identified synthetic unimodal and bimodal speech syllables as either /ba/ or /da/. In Experiment 2, Mandarin speakers were given nine possible response alternatives. Syllable identification was influenced by both visual and auditory sources of information for both Mandarin and English speakers. Performance was better described by the fuzzy logical model of perception than by an auditory dominance model or a weighted-averaging model. Overall, the results are consistent with the idea that although there may be differences in information (which reflect differences in phonemic repertoires, phonetic realizations of the syllables, and the phonotactic constraints of languages), the underlying nature of audiovisual speech processing is similar across languages.

Humans integrate multiple sources of continuous information for pattern recognition (of well-learned patterns), and the nature of integration seems to be fairly invariant across domains. Language is a prototypical case. For example, in speech perception, perceivers integrate auditory (voice) and visual (face) sources of information, and each source is more influential to the extent that the other source is ambiguous (Massaro, 1998).

This behavioral principle has been formalized as the fuzzy logical model of perception (FLMP; Massaro, 1998). Figure 1 illustrates three major operations in the FLMP framework: evaluation, integration, and decision. Features are first independently evaluated (as sources of information) in terms of the degrees to which they match specific memory prototypes. Each feature match is represented by a common metric of fuzzy logic truth values that range from 0 to 1 (Zadeh, 1965). In the second operation, the feature values corresponding to a given prototype are multiplied to yield an overall (absolute) goodness of match for that alternative. Finally, the goodness of match for each alternative is compared against the sum of the support for all the relevant alternatives (the RGR; Massaro, 1998).

Across a range of studies comparing specific mathematical predictions (Massaro, 1988, 1989, 1998; Massaro, Weldon, & Kitzis, 1991), the FLMP has been more successful than other models in accounting for the experimental data (Massaro, 1989, 1998; Massaro & Friedman, 1990). One of the best methods by which to test bimodal speech perception models, as well as to examine the psychological processes involved in speech perception, is to systematically manipulate synthetic auditory and visual speech in an expanded factorial design. This paradigm is especially informative for defining the relationship between bimodal and unimodal conditions and for evaluating a model's specific predictions (Massaro & Friedman, 1990).

Like other theories or models of speech perception, the FLMP claims that its processes are universal across languages. Moreover, given the FLMP framework, we are able to make an important distinction between *information* and *information processing*. The sources of information from the auditory and visual channels make contact with the perceiver at the evaluation stage of processing. The reduction in uncertainty effected by each source is defined as *information*. In the fit of the FLMP, for example, the degree of support for each alternative from each modality corresponds to information. The predicted response probability in the unimodal condition is predicted to be the information given by that stimulus. These values represent how informative each source of information is. *Information processing* refers to how the sources of information are processed. In the FLMP, this processing is described by the evaluation, integration, and decision stages.

The methodology of a set of cross-linguistic experiments allowed us to separate information differences from information-processing differences. Earlier cross-linguistic results had led investigators to conclude that the *processing* of bimodal speech differed for Japanese and English speakers (e.g., Sekiyama & Tohkura, 1993). Although the results of experiments with native English,

The research and writing of this article were supported by Grants CDA-9726363, BCS-9905176, and IIS-0086107 from the National Science Foundation, Public Health Service Grant PHS R01 DC00236, a Cure Autism Now Foundation Innovative Technology Award, and the University of California, Santa Cruz (Cota-Robles Fellowship). The authors thank the reviewers of the article and Michael M. Cohen for offering expert technical assistance, providing computer support, and commenting on a previous draft. Correspondence concerning this article should be addressed to D. W. Massaro, Department of Psychology, University of California, Santa Cruz, CA 95064 (e-mail: massaro@fuzzy.ucsc.edu).

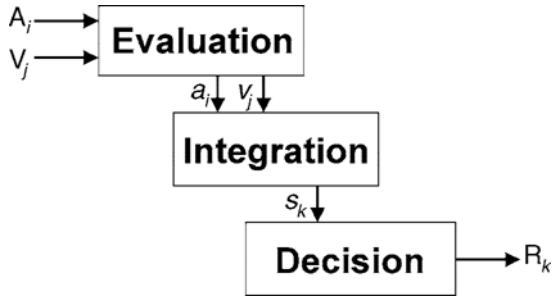

Figure 1. Schematic illustration of three stages of processing in the fuzzy logic model of perception. The three processes involved in perceptual recognition include evaluation, integration, and decision. These processes make use of prototypes stored in long-term memory. The evaluation process transforms sources of information (A_i & V_j) into psychological values (a_i and v_j), which are then integrated to give an overall degree of support (s_k) for each speech alternative. The decision operation maps the outputs of integration into some response alternative (R_k). The response can take the form of a discrete decision or a rating of the degree to which the alternative is likely.

Spanish, Japanese, and Dutch talkers showed substantial differences in performance across the different languages (Massaro, Cohen, & Smeele, 1995; Massaro, Tsuzaki, Cohen, Gesi, & Heredia, 1993), the application of the FLMP indicated that these differences could be completely accounted for by information differences, with no differences in information processing.

The information in a speech segment made available by the evaluation process naturally differs for talkers of different languages, whereas information processing appears to be invariant. The differences that are observed are primarily the different speech categories used by the different linguistic groups, which can be attributed to differences in the phonemic repertoires, phonetic realizations of the syllables, and phonotactic constraints in these different languages. In addition, talkers of different languages are similarly influenced by visible speech, with its contribution being largest to the extent that the other source is ambiguous. The details of these judgments are nicely captured in the predictions of the FLMP.

For example, using synthetic auditory and visible speech in expanded factorial designs that include both unimodal and bimodal stimulus conditions, two studies in which bimodal speech perception of English, Japanese, Spanish, and Dutch speakers was examined (Massaro et al., 1995; Massaro et al., 1993) showed that the FLMP described the experimental data better than did an auditory dominance model (ADM), a categorical model of perception, and an additive model of perception. Together, these studies suggest that the nature of information processing in speech perception is similar across languages, even though the specific response patterns may differ among different language speakers.

On the other hand, some investigators have argued that speech perception processes are not universal across languages. This argument comes from interlanguage differences in the pattern and magnitude of the McGurk effect (i.e., the phenomenon that occurs when auditory speech perception is altered by incongruent visual information) for particular sets of monosyllables. Using sets of recorded video and audio signals, for example, Sekiyama and Tohkura (1989, 1991, 1993; Sekiyama, 1997) presented native speakers of American English, Japanese, and Chinese various combinations of two sets of audiovisual stimuli (/ba/, /pa/, /ma/, /wa/, /da/, /ta/, /na/, /ra/, /ga/, and /ka/) spoken by female native speakers of Japanese and English, respectively. Comparing the McGurk effect magnitude shown by the three groups, English stimuli induced a weaker visual effect in the Chinese speakers than in the American and Japanese speakers, and Japanese stimuli induced a weaker visual effect in the Chinese and Japanese speakers than in the American speakers (Sekiyama, 1997; Sekiyama & Tohkura, 1989, 1991, 1993).

Overall, the visual influence was greatest when the Americans perceived Japanese and was weakest when the Japanese perceived Japanese. Under noise-free settings, the McGurk effects for the Japanese perceiving Japanese almost were limited to stimuli of less than 100% accuracy on auditory-alone trials. The authors proposed that the Japanese use a qualitatively different type of perceptual processing while perceiving Japanese syllables—relying more on auditory information and using visual information only when the auditory speech is not completely identified (Sekiyama, 1997; Sekiyama & Tohkura, 1989, 1991, 1993).

According to this auditory intelligibility hypothesis (Sekiyama & Tohkura, 1993), Japanese participants listening to Japanese syllables are "hardly influenced by visual cues when audition provides enough information and that the size of the visual bias to their responses depends on the intelligibility score of auditory stimuli" (p. 428). They suggested that the results "indicate that Americans automatically integrate visual cues with auditory cues in (vision-dependent processing). . . . Japanese subjects, in contrast, incorporated visual cues much less than Americans when perceiving native syllables (vision-independent processing). . . . the relative rarity of the McGurk effect for Japanese participants must be attributed to the perceptual processing of Japanese listeners" (Sekiyama & Tohkura, 1993, p. 441). Although they posited that other possible factors might also be responsible (such as differences in listeners' perceptual organization for speech [Sekiyama & Tohkura, 1991] and the simpler structure of the Japanese phonological system [Sekiyama & Tohkura, 1993] that enables syllable discrimination without visual information), they proposed that Japanese listeners "tend to separate visual information from auditory information as long as audition provides enough information" (Sekiyama & Tohkura, 1993, p. 442).

These cross-linguistic differences might indicate that cultural and linguistic factors influence the manner of audiovisual speech processing. A possible cultural explanation is that the Japanese (and Chinese) tend to avoid looking directly at talkers' faces (which is considered

impolite) and thus developed a different type of speech processing (Sekiyama, 1997). A possible linguistic explanation is that reliance on auditory information is related to the intrinsic tonal properties of the perceiver's native language (Burnham, Lau, Tam, & Schoknecht, 2001; Sekiyama, 1997).

However, despite Sekiyama (1997; Sekiyama & Tohkura, 1993) and Massaro et al.'s (1995; Massaro et al., 1993) seemingly contradictory observations and interpretations, their data may not necessarily be inconsistent. One possible resolution simply requires a distinction between information and information processing (Massaro, 1998). Different language speakers may be influenced by visual and/or auditory sources of information in different quantities, but the nature of audiovisual processing may still be described by the FLMP. For example, the FLMP has been shown to accurately describe results when visual influences were small (Massaro, 1998; Tiippana, Sams, & Andersen, 2001). This outcome also seems consistent with the observation that although speakers of different languages may differ in terms of the amount of visual influence under certain conditions, their unimodal and bimodal performance appears to be related (e.g., De Gelder & Vroomen, 1992).

Other factors might be responsible for cross-linguistic differences. The study that compared Japanese and Americans (Sekiyama & Tohkura, 1993) was a between-groups design, and the Japanese speakers might have looked and/or visually attended less to the Japanese speech, relative to other groups or conditions. The type of instructions given (report "what they heard, not what they saw") could have increased the influence of the auditory modality, relative to the visual (Massaro, 1998; Tiippana et al., 2001). In addition, there might have been possible cultural differences in interpretation of the instructions. For example, the Americans may have tended to interpret their instructions as "reporting what you thought was said/meant/spoken," whereas the Japanese might have interpreted them as "reporting the exact sound."

The study (Sekiyama & Tohkura, 1993) also required a report of audiovisual discrepancy, which might have influenced the results. For example, reporting audiovisual discrepancy and reporting what was heard are two possibly conflicting tasks. It is possible that between-group differences occurred in the amount of focus on one of these tasks. Indeed, there were fewer reports of audiovisual discrepancies when identification was consistent with the auditory stimulus for the Japanese group who perceived Japanese than for the Japanese group who perceived English. Finally, unimodal and bimodal stimuli were presented in different blocks of trials, which could have promoted other differences across the different conditions. To eliminate these potential confounding factors, participants should be asked to report only what was said and to also use response-paced trials of visual-alone, auditory-alone, and audiovisual stimuli that are all randomly mixed and presented within an experimental session.

To help resolve this controversy, further testing is needed to examine whether speakers of a particular language use a different type of audiovisual speech processing. If the tendency to not look at faces during everyday speech (cultural) and the tonal properties of a native language (linguistic) do indeed influence the nature of audiovisual speech processing, differences should be more clearly observed between native English and native Mandarin Chinese speakers. A face avoidance tendency has putatively been observed in the Chinese culture (Sekiyama, 1997), and Mandarin is considered much more of a multitonal language than English (and Japanese). Because the meaning of Mandarin words almost always change with tone (e.g., /ma/ can mean "mother," "hemp/numb," "horse," or "reproach"), which are more easily distinguished acoustically than visually, one might speculate that native Mandarin speakers use visual information less than do native speakers of monotonal languages. Thus, it is possible that Mandarin speakers process audiovisual speech differently from English speakers.

The present study contrasted the unimodal and bimodal speech perception of native Mandarin Chinese speakers with that of English speakers. In this study, it was asked whether the FLMP would be as successful in predicting and describing bimodal speech perception for Mandarin speakers as it is for English speakers. Similar to our other cross-linguistic experiments, we expected that Mandarin speakers would use visual information and perceive bimodal speech in the same manner as speakers of other languages. In other words, we tested the hypothesis that the information processing of Mandarin speakers is identical to that of English speakers even though the information provided by the auditory and visual speech might differ for these two groups. In addition to testing mathematical models, we examined the specific response patterns of Mandarin speakers and compared them with the response patterns of speakers of other languages studied previously.

EXPERIMENT 1
Two Response Alternatives

Method

Participants. For this 2-h experiment, two groups of participants were recruited from the University of California, Santa Cruz (UCSC). One group consisted of 7 native Mandarin speakers, and the other consisted of 7 native American English speakers. None reported any hearing or vision problems. The Mandarin speakers' ages ranged from 18 to 29 years. On average, their exposure to English (mostly in school) began at the age of 11. Six of them came from mainland China, and they had lived in the United States for 8.6 months on average. One participant came from Taiwan and had lived in the United States for 6 years. The English speakers' ages ranged from 18 to 22. Six of the Mandarin speakers were paid $20 each, and one was rewarded with $10 and 1-h course credit for participating. Each of the English speakers was given a 2-h course credit.

Stimuli. The test stimuli were similar to the audible and visible synthetic speech syllables used by Massaro et al. (1995; Massaro et al., 1993). A five-step continuum from a good /ba/ to a good /da/ was created for both auditory and visible synthetic speech. For syn-

Table 1
Summary of Statistical Comparisons in Experiment 1

Source	Unimodal Visual	Unimodal Auditory	Bimodal
Visual	$F(4,48) = 145.99$**		$F(4,48) = 106.91$**
Auditory		$F(4,48) = 68.73$**	$F(4,48) = 24.69$**
Language	$F(1,12) = 1.93$	$F(1,12) = 2.04$	$F(1,12) = 6.89$*
Language × visual	$F(4,48) = 0.93$		$F(4,48) = 1.67$
Language × auditory		$F(4,48) = 0.42$	$F(4,48) = 0.69$
Visual × auditory			$F(16,192) = 5.29$**
Language × visual × auditory			$F(16,192) = 2.82$**

*Significant at $p < .05$. **Significant at $p < .001$.

thetic audible speech, tokens of /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). A set of five 400-msec consonant–vowel syllables, which covered the range from /ba/ to /da/, was created by altering the parametric information specifying the first 80 msec of the syllable.

For synthetic visible speech, a computer-animated head was created by combining approximately 900 small, parametrically controlled polygons that modeled the three-dimensional surface of the face (Cohen & Massaro, 1990), which was then smooth shaded using Gouraud's (1971) method. The face was controlled by 50 parameters, of which 11 were used for speech animation. Animation and speech synthesis was done in real time on a Silicon Graphics 4D/Crimson VGX workstation (running under the IRIX operating system; for more details, see Massaro 1998; Massaro et al., 1995; Massaro et al., 1993).

Procedure. Synthetic auditory and visual speech stimuli were manipulated in the expanded factorial design. The onsets of the second and third formants were varied to give an auditory continuum between /ba/ and /da/. Analogously, parameters of the facial model were systematically varied to render a visual continuum between /ba/ and /da/. Five levels of audible speech were crossed with five levels of visible speech, and each stimulus was also presented alone as a unimodal condition, for a total of 35 (25 + 5 + 5) stimulus conditions. Six random sequences were determined by sampling the 35 conditions without replacement, yielding six different blocks of 35 trials.

All instructions and interactions were spoken in Mandarin only by native Mandarin speakers and in English only by native English speakers. The participants were instructed to listen and watch the speaker and to identify the spoken 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 participants were given a 5-min break after every 210 trials. Visual stimuli were displayed on individual color video monitors (JVC Model TM-131SU, 13 in.), and auditory stimuli were displayed with speakers (79 dB A).

The participants were tested individually in soundproof rooms. They gave answers by pressing either of the two sticker-labeled buttons, "BA" or "DA." These two buttons corresponded to "B" and "N" (respectively) on standard keyboards (responses were collected on TVI video display terminals, TUI-950). For the Chinese participants, the equivalent pin-yin and BPMF (written Chinese phonemic system for spelling syllables) for /ba/ and /da/ were also on the sticker-labeled buttons (for people from mainland China, these are written as "BA" and "DA," but for people from Taiwan they are not). All the trials were response paced, but the participants were told to respond as soon as the decision had been made. Data analysis was performed on the Silicon Graphics workstation, using Fortran 77 data analysis routines and the SAS statistical package (SAS Institute, Cary, NC).

Results

We carried out two types of data analyses: statistical tests on the response identifications and quantitative model testing. It should be emphasized that there is no necessary relationship between these two types of analyses. The response identifications are a function of both information and information processing, and any significant differences in these do not necessarily mean differences in one or the other. The model tests are needed to determine whether the assumptions of a given model describe performance.

The participants' response identifications (BA or DA) were recorded for each stimulus. The mean observed proportion of /da/ identifications, $P(/da/)$, was computed for each participant under each condition. Separate analyses of variance were carried out on $P(/da/)$ (the dependent variable) for the two unimodal and bimodal conditions, with language group and the auditory and visual continua as independent variables. Table 1 summarizes the statistical comparisons. Figures 2 and 3 show the results for the English and the Mandarin speakers, respectively. As can be seen in the left and right panels of Figures 2 and 3, there were significant effects of the auditory and visual continua in the unimodal conditions. Importantly, there were no significant differences between the language groups, and there were no significant interactions of these variables with language group.

For the bimodal conditions shown in the middle panels of Figures 2 and 3, there were significant effects of the auditory and visual continua and an auditory–visual interaction. There was no significant language–visual interaction and no significant language–auditory interaction. The difference between language groups was significant, as well as the three-way interaction between language groups, visual, and auditory levels. However, these significant differences do not necessarily reflect differences in information processing. As can be seen in Figures 2 and 3, there were significantly more overall /da/ response judgments for the Mandarin ($M = .58$) speakers than for the English ($M = .44$) speakers.

To test models of information processing, three quantitative models, the FLMP, the ADM, and a weighted-averaging model (WTAV), were fit individually for each participant. The ADM is a formalization based on the assumptions of Sekiyama and Tohkura (1993). 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 (Massaro et al., 1993). The ADM is one way to mathematically formalize the idea that auditory intelligibility

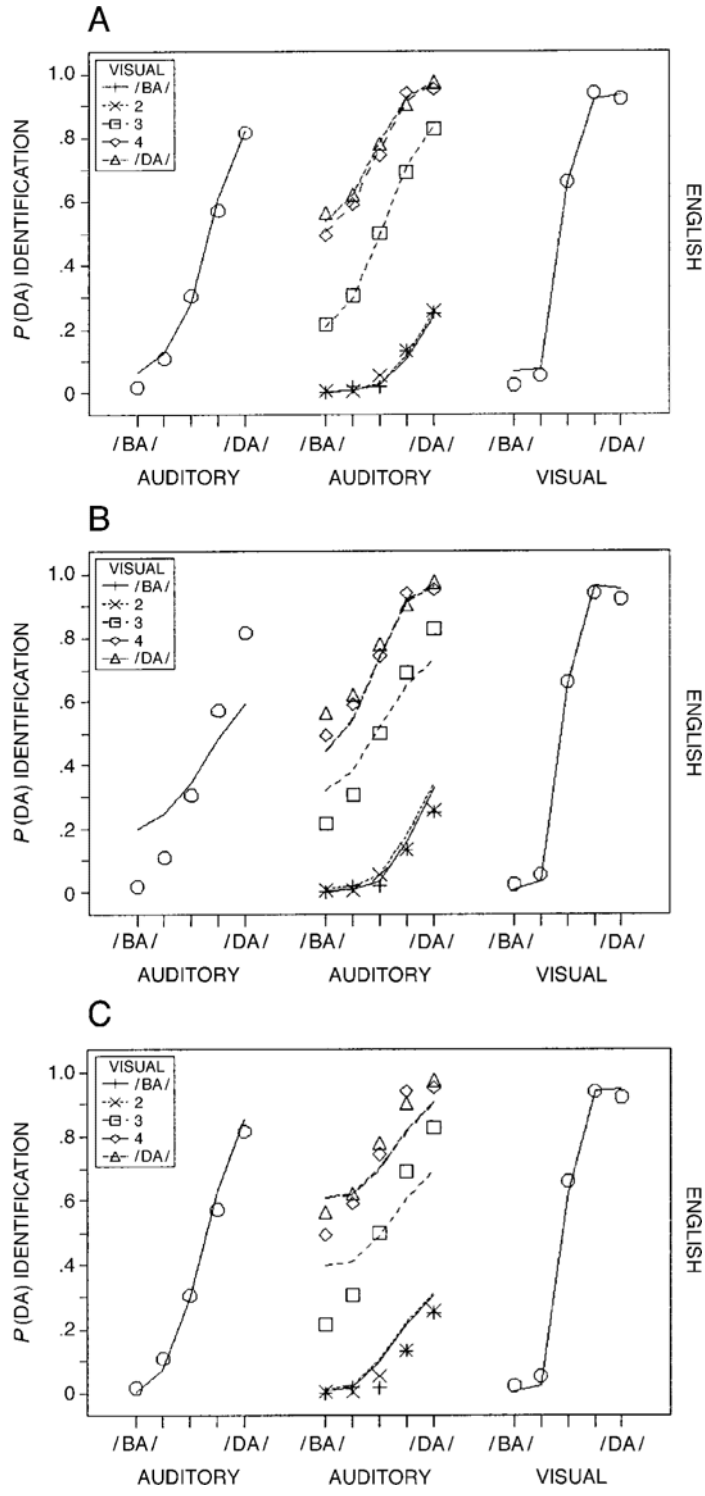

Figure 2. Observed (points) and predicted (lines) probability of a /da/ **response for the auditory-alone (left plot), bimodal (middle plot), and visual-alone (right plot) conditions as a function of the five levels of the synthetic auditory and visual speech varying between** /ba/ **and** /da/ **for English speakers. (A) Predictions for the fuzzy logic model of perception. (B) Predictions for the auditory dominance model. (C) Predictions for the weighted-averaging model.**

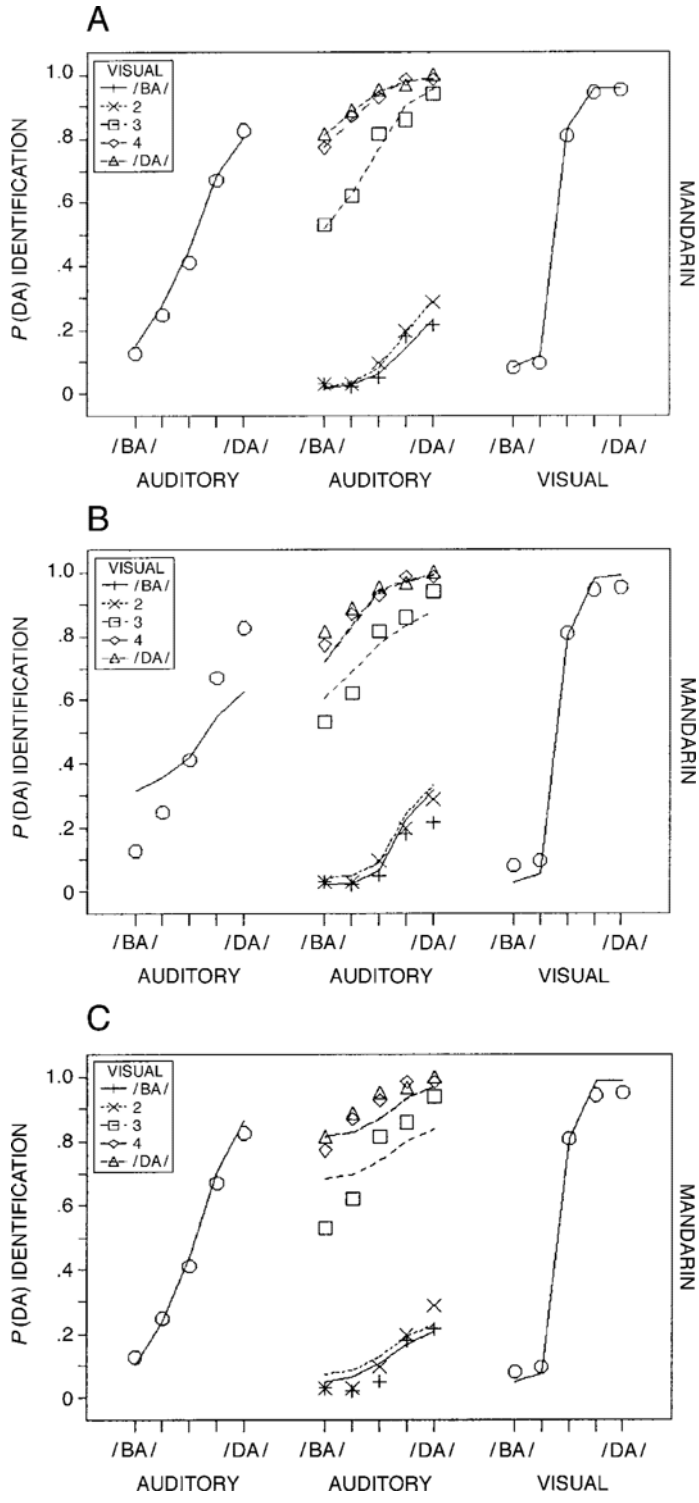

Figure 3. Observed (points) and predicted (lines) probability of a /da/ **response for the auditory-alone (left plot), bimodal (middle plot), and visual-alone (right plot) conditions as a function of the five levels of the synthetic auditory and visual speech varying between** /ba/ **and** /da/ **for Mandarin speakers. (A) Predictions for the fuzzy logic model of perception. (B) Predictions for the auditory dominance model. (C) Predictions for the weighted-averaging model.**

determines whether visible speech will have an effect. The predicted probability of a response (r) for auditory-alone trials is $P(\mathrm{r}\,|\,\mathrm{A}) = a_\mathrm{r} + (1 - \sum a_\mathrm{r})w_\mathrm{r}$, for visual-alone trials it is $P(\mathrm{r}\,|\,\mathrm{V}) = v_\mathrm{r} + (1 - \sum v_\mathrm{r})w_\mathrm{r}$, and for bimodal trials it is $P(\mathrm{r}\,|\,\mathrm{A\ and\ V}) = a_\mathrm{r} + (1 - \sum a_\mathrm{r})[v_\mathrm{r} + (1 - \sum v_\mathrm{r})w_\mathrm{r}]$. This model predicts that either an auditory stimulus is identified or the participant bases the decision on the visual information. The term a_r (or v_r) represents the probability of identifying the auditory (or visual) source as response r, $\sum a_\mathrm{r}$ (or $\sum v_\mathrm{r}$) is the probability of identifying the auditory (or visual) source as any of the response alternatives, and w_r represents some bias toward r in case no identification is made (Massaro, 1998).

On the other hand, the FLMP predicts that both visible and audible speech influence speech perception, and this mathematical relationship is invariant in pattern recognition. For the FLMP, the predicted probability of a response (r) given visible and audible speech is $P(\mathrm{r}\,|\,\mathrm{A}_i\ \mathrm{and}\ \mathrm{V}_j) = (a_i v_j)/[a_i v_j + (1 - a_i)(1 - v_j)]$. The term a_i (or v_j) represents the degree to which auditory stimulus A_i (or visual V_j) supports the alternative r. In unimodal trials, either a_i or v_j is available, so the probability is simply the degree of support for r divided by the sum of support for all the relevant alternatives.

Although the ADM and the FLMP make very different assumptions about the nature of bimodal speech perception, the two models actually make fairly similar predictions. The integration algorithm of the FLMP predicts that the influence of the visual information source is directly related to the ambiguity of the auditory source. The ADM (a nonintegration model) claims a similar outcome, because the visual information is used only when the auditory speech is not identified, which will necessarily occur more often for ambiguous auditory speech. Given this similarity in the predictions, we might expect that not all experimental tests will be sufficiently informative to test between the two models.

For the WTAV, integration is based on a weighted averaging of the support from the two sources of information (see Anderson, 1981). It differs from the FLMP in bimodal trials: $P(\mathrm{r}\,|\,\mathrm{A\ and\ V}) = (w)(a_i) + (1 - w)v_j$; the term $w = w1/(w1 + w2)$, where $w1$ and $w2$ represent the weights given the auditory and the visual sources, respectively, and w represents the relative weight given to the auditory source (for a thorough and complete description of each mathematical model, see Massaro, 1998, chap. 2).

It should be noted that all of the models assume that information processing is invariant, and they allow for differences in information. The ADM, for example, could allow Mandarin speakers to identify auditory speech more often than do English speakers and, therefore, use visual speech less often. There would still be the same information processing. Differences in information processing across languages would mean, for example, that the FLMP would describe the data better for English speakers and the ADM would describe the data better for Mandarin speakers.

The quantitative predictions of the models (in terms of the probability of /da/ identifications) were determined for each participant, using the program STEPIT (Chandler, 1969), which calculates a set of parameters that minimizes the root-mean squared deviations (RMSDs) between the observed and the predicted theoretical points. Each model is represented to the program in terms of a set of prediction equations and a set of parameters, and the RMSD indexes the goodness of fit.

The individual RMSD values of each participant for each model are listed in Appendix A. Table 2 gives the mean RMSD values for each model. Separate analyses of variance on the RMSD values contrasted the FLMP to each of the other two models. Language was also a factor in each analysis. The results showed that the FLMP (10 parameters) gave a significantly better description of the results than did both the ADM [11 parameters; $F(1,6) = 7.19$, $p = .04$] and the WTAV [11 parameters; $F(1,6) = 25.05$, $p < .01$]. There were no significant differences between the language groups for the model fits of the FLMP [$F(1,6) = 0.34$, $p = .58$], the ADM [$F(1,6) = 1.25$, $p = .31$], and the WTAV [$F(1,6) = 1.63$, $p = .25$]. Finally, there were no significant interactions between language groups and models when the FLMP was compared with both the ADM [$F(1,6) = 0.10$, $p = .76$] and the WTAV [$F(1,6) = 1.74$, $p = .24$].

Figures 2 and 3 illustrate that the FLMP is better able to capture the configuration of the results than either the ADM or the WTAV. As can be seen in Figures 2A and 3A, the predictions of the FLMP are very close to the observations. Figures 2B and 3B show how the ADM fails to predict the points. The ADM assumes that visual information is used only when the auditory information is not identified. Thus, given the large effect of visual speech in the bimodal conditions, the model must assume that the auditory speech was not identified very accurately. Thus, the predicted results for the unimodal auditory continuum range between .20 and .63, whereas the actual results ranged between .02 and .84. Figures 2C and 3C show that the WTAV predicts parallel curves for the bimodal conditions and, therefore, cannot describe the much larger change in the identification judgments when the visual speech was relatively ambiguous (Level 3).

Table 2
Average Root-Mean Squared Deviation Values for Each Model and Experiment: Data Set

		Model		
Experiment	Data Set	WTAV	ADM	FLMP
1	Mandarin	.0698	.0821	.0476
1	English	.0951	.0946	.0545
2	9 Responses	.0696	.0655	.0623
2	/ba/, /da/, and "other"	*	.1388	.1080
2	Labials and nonlabials	*	.1369	.1075

Note—WTAV, weighted-averaging model; ADM, auditory dominance model; FLMP, fuzzy logic model of perception. *The WTAV was not tested against these data.

Figure 4 plots the RMSD values from the FLMP and the ADM fits for each of the individual participants. For most participants, this value is smaller for the fit of the FLMP, as compared with the fit of the ADM.

Discussion

It appears that both the Mandarin and the English speakers were influenced by both visual and auditory sources of information, as described by the FLMP. The only observed difference was that the Mandarin speakers were overall more likely to choose /da/ (M = .58) than were the English speakers (M = .44), especially at ambiguous levels of the continua. This overall difference could also be responsible for the significant three-way interaction between language groups, visual level, and auditory level. This could be due to a /da/ decision bias, or it is possible that both the auditory and the visual stimuli were more prototypically /da/-like to the Mandarin speakers. This difference between the two language groups is important because it reveals that the Mandarin speakers did not simply perceive the stimuli as English syllables; otherwise, their responses would have been identical to those of the English speakers on the basis of how prototypical the test stimuli were to /ba/ or /da/ of standard American English. This finding and the fact that the FLMP provided a significantly better description of the results than did both the ADM and the WTAV for both language groups strongly suggest that the nature of information processing underlying bimodal speech perception is invariant despite potential differences in information.

To summarize, this experiment showed that Mandarin speakers use visual information for integrating audiovisual speech, this process is best described by the FLMP, and Mandarin and English language groups do not differ on the type or nature of information processing.

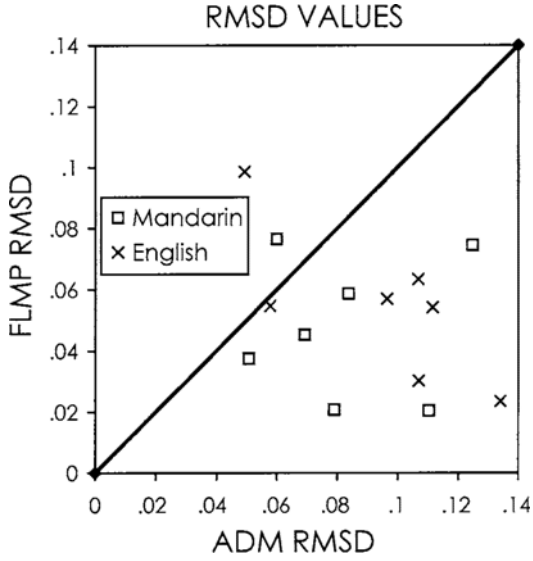

Figure 4. Root-mean squared deviation (RMSD) values from the fits of the fuzzy logic model of perception (FLMP) and the auditory dominance model (ADM) for each of the individual participants.

EXPERIMENT 2
Multiple Response Task

Mandarin speakers indeed use visual information and appear to integrate it with auditory information as described by the FLMP. In the spirit of falsification and broadening the domain of psychological inquiry (Massaro, 1998, chap. 6), it is important to extend the research paradigm. One obvious extension is to allow a larger number of response alternatives. This task should reveal more directly the influence of the phonological repertoire of perceivers. Previous research has suggested that speakers tend to respond with alternatives in their language that best match both the visual and the auditory sources of information (Massaro, 1998; Massaro et al., 1995; Massaro et al., 1993). In addition to /ba/ and /da/ responses, English speakers gave frequent responses of /va/, /ða/ and /bda/ (fewer /za/); Japanese speakers gave frequent responses of /wa/ and /za/ (fewer /ga/); Dutch speakers gave frequent responses of /va/ and /va/ (fewer /za/).

Mandarin Chinese presents an interesting case because, even though /ba/ and /da/ are frequent syllables, there seems to be no clear in-between syllables that are similar psychophysically to both of them (there are no /va/, /ða/, or /va/ segments in Mandarin). If differences in response categories arise from different phonemic repertoires, phonetic realizations of the syllables, and phonotactic linguistic constraints (Massaro, 1998), Mandarin speakers should give fewer responses other than /ba/ and /da/. Potential confusable responses in Mandarin are /ga/, /wa/, and /za/. Even though there are no true Mandarin consonant clusters, Mandarin speakers might also respond /bda/, because it is conceivable that clusters can be produced in Mandarin when two adjacent syllables are said together rapidly.

A pilot study (with 5 additional native Mandarin speakers) showed that /ba/, /da/, and /ga/ responses were by far the most frequent, and all other response choices were relatively infrequent. There were no /va/, /ða/, or /va/ responses given. Consequently, for Experiment 2, we provided nine response categories: /ba/, /da/, /bda/, /dba/, /ga/, /la/, /za/, /wa/, and the category *other* (/bda/ and /dba/ were included to allow direct comparisons to relevant previous cross-linguistic studies). It should be pointed out that /ba/, /da/, /ga/, /la/, and /wa/ are words in Mandarin Chinese, but the other alternatives are not.

If visible speech has an influence only when audible speech is not identified, the ADM should provide a good description of the results for Chinese participants. On the other hand, the FLMP should give the best description if Chinese and English speakers share the same information processing. By examining the response patterns of Mandarin speakers, it is possible to test whether visible speech can still have a significant influence when

multiple alternatives are permitted. These data will also allow a test of the predictions of the quantitative models.

Method

For this 2-h experiment, 7 native Mandarin speakers were recruited from UCSC. None of them had participated in Experiment 1 or in the pilot experiment. None of them reported any hearing or vision problems. All of them came from China, and their ages ranged from 18 to 29 years. On average, they had lived in the United States for 2.3 years, and exposure to English began at the age of 10.8 in school. They were paid $20 each.

The same set of stimuli as that from Experiment 1 was used. All the procedural details were the same as those described before, except that, for this experiment, the participants were instructed to identify the syllables by choosing from nine responses: /ba/, /da/, /bda/, /dba/, /ga/, /la/, /za/, /wa/, and *other*. The sticker-labeled (with equivalent pin-yin, BPMF, and Chinese characters for *other*) buttons corresponded to "X," "C," "V," "B," "N," "M," "<" ">.," and "?" (respectively, from left to right) on a standard keyboard. Instructions were given in Mandarin, and only Mandarin was spoken during the experiment and other interactions.

Results and Discussion

Figure 5 shows the observed results for seven of the most frequent responses. The participants' response identifications were recorded for each stimulus. The mean observed proportion of identifications was computed for each participant under each condition. The probability of response for each choice are /da/ (33.0%), /ba/ (29.5%), /ga/ (18.9%), /bda/ (5.6%), /la/ (4.5%), /za/ (4.5%), /wa/ (1.6%), /dba/ (1.3%), and *other* (1.1%). Separate

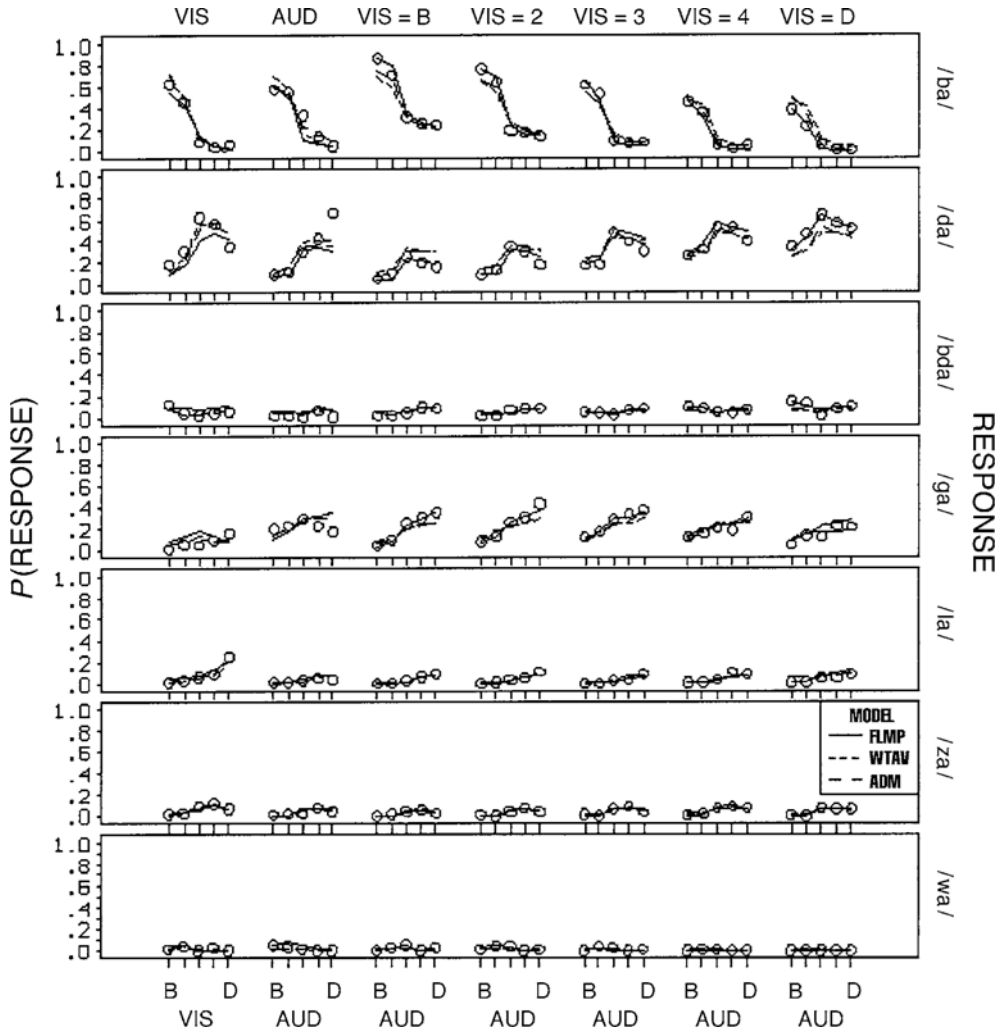

Figure 5. Observed (points) and predicted (lines) proportions of /ba/**,** /da/**,** /bda/**,** /ga/**,** /la/**,** /za/**, and** /wa/ **identifications for the visual-alone (left plot), auditory-alone (second plot), and bimodal (remaining plots) conditions as a function of the five levels of the synthetic auditory (AUD) and visual (VIS) speech varying between** /ba/ **(B) and** /da/ **(D) for Mandarin speakers. The lines give the predictions of the fuzzy logic model of perception (FLMP), auditory dominance model (ADM), and the weighted-averaging model (WTAV).**

analyses of variance were carried out on the probability of responses (the dependent variable) for unimodal and bimodal conditions, with response type and the auditory and visual continua as independent variables.

In the unimodal conditions, there were significant interactions of response type and visual continuum [$F(32,192) = 9.16, p < .001$] and response type and auditory continuum [$F(32,192) = 17.69, p < .001$]. For bimodal conditions, response type again interacted significantly with both the visual continuum [$F(32,192) = 13.01, p < .001$] and the auditory continuum [$F(32,192) = 17.72, p < .001$]. The interaction between auditory, visual, and response type was also significant [$F(128,768) = 1.99, p < .001$].

The relatively strong bias of the Mandarin speakers to respond /da/ in Experiment 1 was not found in Experiment 2. These results are not necessarily inconsistent with one another because of the different response alternatives available in the two cases. For example, it is reasonable that Mandarin speakers in Experiment 1 might have substituted the response /da/ for those cases in which they perceived a /ga/ or a /la/, which would give more /da/ judgments in Experiment 1, but not in Experiment 2.

The individual RMSD values of each participant for each model are listed in Appendix A. Table 2 gives the mean RMSD values for each model. The ADM, the FLMP and the WTAV were fitted individually to the nine responses of each participant. Separate analyses of variance on the RMSD values contrasted the FLMP to each of the other two models. The results showed that the FLMP (80 parameters) gave a significantly better description of the results than did the WTAV [81 parameters; $F(1,6) = 5.52$, $p = .05$]. However, FLMP's .003 RMSD advantage did not reach statistical significance over the ADM [88 parameters; $F(1,6) = 0.63$, $p = .54$].

Although failing to find a significant FLMP advantage over the ADM seemed a bit surprising at first, hindsight suggested that the results were not sufficiently informative to distinguish between the two models. Even though we made available a large set of response alternatives, it appears that there were not enough syllables similar to /ba/ and /da/ for the Mandarin speakers. Many empty cells will make the data less informative in terms of testing different models. In our study, most of the response categories were infrequent response choices, which both models could easily predict. This situation could have diluted the advantage of the FLMP's prediction of the other response categories.

To test this possibility, we carried out two additional model fits by grouping the actual responses into fewer response categories. First, model testing was repeated with responses grouped into three categories: /ba/, /da/, and *other*. As was noted earlier, six of the nine possible responses occurred less than 6% of the time. When grouped in this way, the three response categories are roughly equal in proportion (/ba/, 29.5%; /da/, 33.0%; and *other*, 37.5%). Consistent with our interpretation, the FLMP (20 parameters) provided a better description of the data than did the ADM [22 parameters; $F(1,6) = 43.40$, $p = .001$]. Appendix B (left) lists the individual RMSD values for this

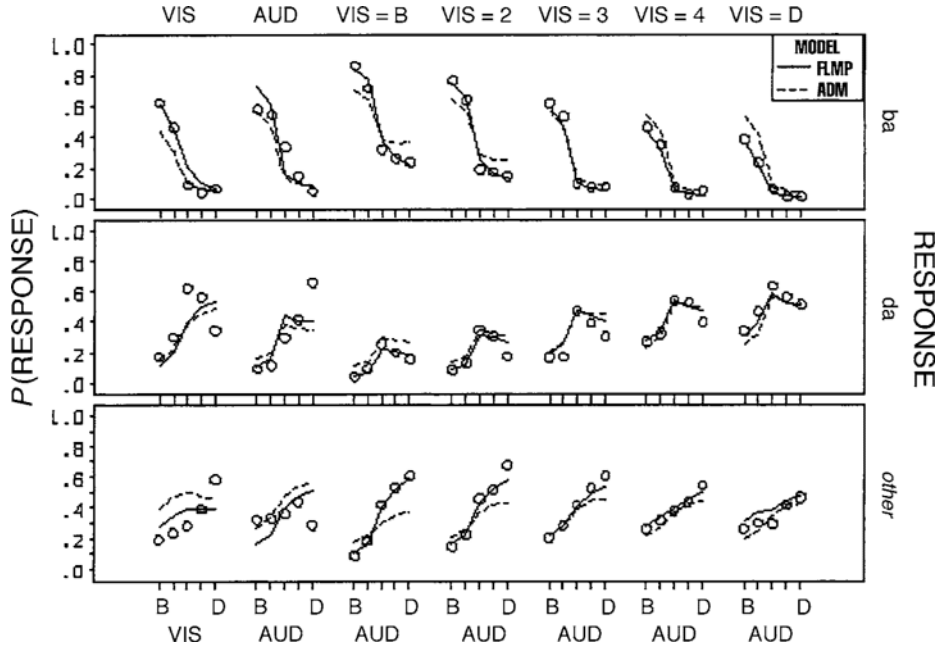

Figure 6. Observed (points) and predicted (lines) proportions of /ba/, /da/, and *other* identifications for the visual-alone (left plot), auditory-alone (second plot), and bimodal (remaining plots) conditions as a function of the five levels of the synthetic auditory (AUD) and visual (VIS) speech varying between /ba/ (B) and /da/ (D) for Mandarin speakers. The lines give the predictions of the fuzzy logic model of perception (FLMP) and the auditory dominance model (ADM).

fit. Table 2 gives the mean RMSD values for each model. Figure 6 shows the observed and the predicted results for these three response categories. As can be seen in the figure, the predictions of the FLMP are closer to the observed results than are the predictions given by the ADM.

Model testing was also repeated with responses grouped into two categories: labials (ba, bda, dba) and nonlabials (everything else). This is a reasonable partition, given the finding that perceivers are highly sensitive to the visual distinction between labials and nonlabials (Sekiyama, 1997; Sekiyama & Tohkura, 1993). The FLMP (10 parameters) also provided a better description of these results than did the ADM [11 parameters; $F(1,6) = 5.62, p = .05$]. Appendix B (right) lists the individual RMSD values for this fit. Table 2 gives the mean RMSD values for each model. Figure 7 gives the observed and the predicted results for these two response categories. As can be seen in the figure, the predictions of the FLMP are closer to the observed results than are the predictions given by the ADM.

Independently of the validity of grouping responses into fewer categories, Experiment 2 added to what was learned in Experiment 1. For example, we found that Mandarin and English speakers do have different phonemic repertoires, since the Mandarin speakers used fewer response alternatives than did the English speakers under similar experimental conditions (Massaro et al., 1995; Massaro et al., 1993). We learned that not all experimental conditions are sufficient to distinguish among mathematical models, and perhaps future research should be sensitive to finding experimental conditions that allow different models to be distinguished.

The size of a visual effect depends on both the auditory and the visual stimuli, as well as the number and types of responses allowed. Therefore, it is not valid to compare the sizes of the visual effects between Mandarin and English speakers or between the Mandarin speakers in Experiments 1 and 2. As an example, the auditory syllables could have been more ambiguous for the Mandarin than for the English speakers, which would necessarily produce a larger visual effect. Furthermore, the /da/ bias in Experiment 1 was eliminated in Experiment 2 when other responses, such as /ga/, were possible. What seems reassuring is the success of FLMP as a fundamental principle that is differentially expressed across conditions and individuals. Therefore, given the advantage of the FLMP in these experiments, it is important that we address the putatively opposing conclusions in previous research. Our analyses will show that all of the research paints a consistent picture that is best understood in the framework of the FLMP.

Analyses of previous literature. As was discussed in the introduction, some experiments in which language differences in the pattern and magnitude of the McGurk effect were examined have led to the proposition that there are different types of audiovisual speech processing. According to Sekiyama and Tohkura (1991), for example, the Japanese McGurk effect is almost only induced with auditory stimuli that do not have 100% intelligibility when heard alone. They proposed that "human beings may depend on eyes in the presence of auditory uncertainty" (p. 1804).

If their actual findings are examined under the FLMP framework, however, the nature of audiovisual information processing need not differ across language conditions. Their results show that the Japanese are influenced by visual information even when Japanese auditory speech was completely intelligible. In one study (Sekiyama &

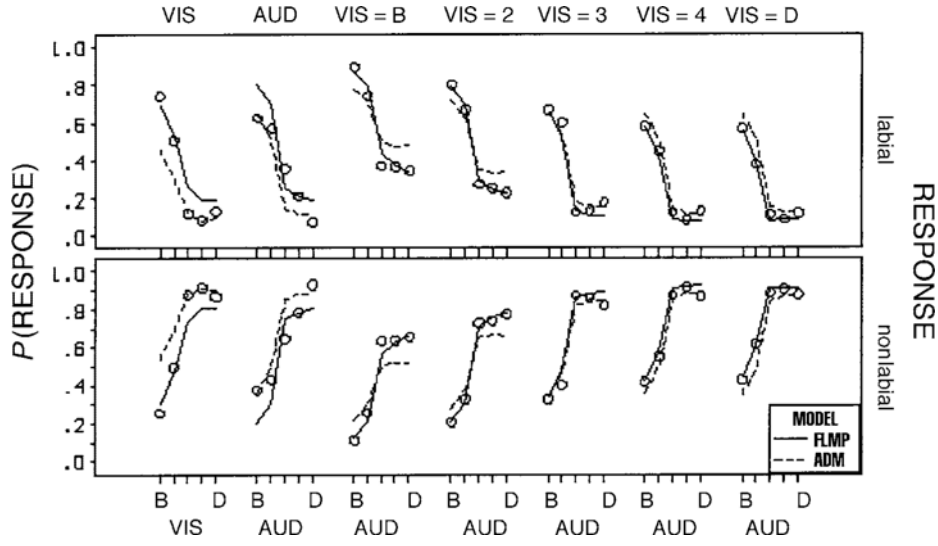

Figure 7. Observed (points) and predicted (lines) proportions of "labial" and "nonlabial" identifications for the visual-alone (left plot), auditory-alone (second plot), and bimodal (remaining plots) conditions as a function of the five levels of the synthetic auditory (AUD) and visual (VIS) speech varying between /ba/ (B) and /da/ (D) for Mandarin speakers. The lines give the predictions of the fuzzy logic model of perception (FLMP) and the auditory dominance model (ADM).

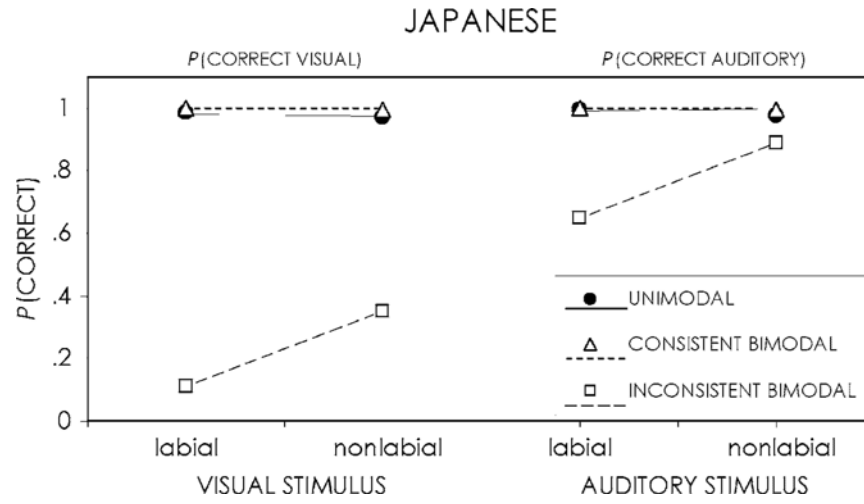

Figure 8. Observed (points) and predicted (lines) probability for correct visual (left panel) and auditory (right panel) responses as a function of visual or auditory level for unimodal, consistent, and inconsistent trials of Mandarin speakers perceiving Japanese stimuli (data calculated from Sekiyama, 1997). The consistent condition is necessarily identical in the two panels because it represents the same results. The lines are the predictions of the fuzzy logic model of perception.

Tohkura, 1991), auditory "ma" was perfectly identified, but 12% of the responses were "na" when it was paired with visual "ka." In another study (Hayashi & Sekiyama, 1998), 40% of the responses were "pa" when visual "pa" was paired with auditory "ta," and 46% of the responses were "na" when visual "na" was paired with auditory "ma"; both auditory stimuli alone were 100% correctly identified. These results appear to directly contradict the hypothesis that visual information is used only when the auditory speech is unintelligible.

If Japanese and Chinese are indeed hardly influenced by vision as long as audition provides enough information, it may be hard to explain findings from a recent study (Hayashi & Sekiyama, 1998). These authors tested Chinese and Japanese with sets of eight syllables (/ba/, /pa/, /ma/, /da/, /ta/, /na/, /ga/, and /ka/), which were pronounced by 2 Japanese and 2 Mandarin speakers. This study produced McGurk effects much greater in magnitude and range than those in previous similar studies (Sekiyama, 1997; Sekiyama & Tohkura, 1991, 1993)

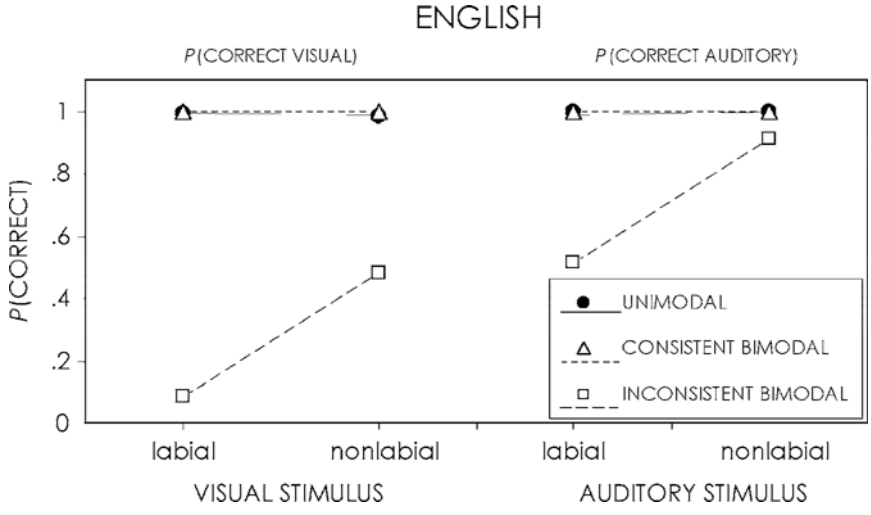

Figure 9. Observed (points) and predicted (lines) probability for correct visual (left panel) and auditory (right panel) responses as a function of visual or auditory level for unimodal, consistent, and inconsistent trials of Mandarin speakers perceiving English stimuli (data calculated from Sekiyama, 1997). The consistent condition is necessarily identical in the two panels because it represents the same results. The lines are the predictions of the fuzzy logic model of perception.

in which stimuli were produced from only a single speaker.

If the size of a perceiver's visual bias depends only on the intelligibility of the auditory speech, it may be hard to explain several findings from this study. For example, one Japanese speaker produced /t/ (44%), which was less intelligible than /p/ (98%) when presented acoustically alone. However, in the bimodal condition, /p/ actually had a bigger visual bias in magnitude (about 97%) than did /t/ (87%). In this case, a putatively more intelligible auditory stimulus was more susceptible to the McGurk effect than was a less intelligible one. This result could be predicted by the FLMP by assuming that the visual information was differentially informative in the two cases.

In an earlier study (Sekiyama & Tohkura, 1993), auditory /pa/ was correctly identified 98% of the time when presented alone. According to the auditory intelligibility hypothesis, it seems that a visual influence should have occurred on only 2% of the trials. However, when this auditory /pa/ was paired with a visual /ga/, only 67% of the responses were /pa/. If a Japanese McGurk effect occurs only when auditory speech is not identified, one might expect close to 98% /pa/ responses in this bimodal condition. It seems that there were cases in which an auditory stimulus was clearly identifiable as a /pa/ but a strong effect of the visual stimulus still occurred.

To explain these results, it is not necessary to assume that Japanese participants incorporate visual information less and use *visual-independent processing*. To illustrate this claim, consider the following case in which auditory support (a_i) is .7 and visual support (v_j) is .5. According to the FLMP equation, $P(\mathrm{r}|\mathrm{A_i}\ \text{and}\ \mathrm{V_j}) = (a_i v_j) / [a_i v_j + (1 - a_i)(1 - v_j)] = (.7*.5) / [.7*.5 + (1 - .7)(1 - .5)] = (.35) / [.35 + (.3)(.5)] = .35/.50 = .7$. In this example, the overall support for r is unchanged by the visual support even though it is integrated with the auditory speech. Of course, $P(\mathrm{r}|\mathrm{A}_i\ \text{and}\ \mathrm{V}_j)$ can give other types of response patterns if the value for the visual support changes. Even though a wide range of response patterns can arise from changing the relative influence of the two sources of information (since there might be interlanguage differences in the quantitative influence of visual cues), the process of integration can be explained by the same mathematical model.

Under the FLMP framework, there is no need or reason to propose, as Sekiyama (1997) did, that for the Japanese "visual information is processed to the extent that the audiovisual discrepancy is detected most of the time . . . . for clear speech, [they] use a type of processing in which visual information is not integrated . . . . for the noise-added condition . . . [they] switch to another type of processing where audiovisual integration occurs" (p. 74).

Given our thesis, an important question is how well the FLMP and the ADM can predict the results found in experiments by Sekiyama (1997; Sekiyama & Tohkura, 1989, 1991, 1993). Because the studies in which Japanese and English speakers were investigated did not include or did not report the unimodal visual responses, model tests are not possible. The study (Sekiyama, 1997) in which Mandarin speakers identifying Japanese and English speech were examined, however, did report unimodal visual performance and, thereby, allows a test between the ADM and the FLMP. Given that the author reported the visual conditions partitioned into labials and nonlabials, these two categories were used for model testing. Because only averaged data were reported, the models were fit to the average results of Mandarin speakers identifying Japanese and English speech. Figures 8 and 9 show the observed and the predicted results of the FLMP for these two response categories. As can be seen in the figures, the FLMP gave a good fit to the results of both language conditions.

To allow a statistical test between the RMSD values of the models, the two stimulus conditions (Japanese and English speech) were treated as a random variable with two levels (analogous to having 2 subjects in an experiment). The RMSDs of the fit of the FLMP (four parameters) were .0096 and .0030 for the Japanese and the English speech, respectively. The corresponding RMSDs of the fit of the ADM (five parameters) were .2363 and .2587. Thus, the FLMP provided a significantly better description of the results than did the ADM [$F(1,1) = 276.71$, $p < .05$]. In a second analysis, the two model tests were treated as a random variable. In this case, there was no significant difference in the fit of the models to the two language conditions [$F(1,1) = 0.30$, $p = .68$]. These statistical comparisons indicate that the FLMP gave a significantly better description than did the ADM for both language conditions.

Several observations about Figures 8 and 9 are worth making. First, the response pattern is comparable to those from experiments examining English speakers with English stimuli. Figure 10 plots data from such an experiment (Massaro, 1998, p. 10). One difference is that Figures 8 and 9 show a near-perfect unimodal performance, and this could be a ceiling effect masking possible improvements in consistent bimodal performance. One similarity is that inconsistent bimodal trials produce more responses consistent with the auditory categories than with the visual categories. This pattern is readily apparent for both Japanese and English speech stimuli, and it shows that both Mandarin and English perceivers are more influenced by auditory than by visual information. Consistent with this pattern, the FLMP accurately predicts the responses of all three conditions: Mandarin speakers perceiving Japanese, Mandarin speakers perceiving English, and English speakers perceiving English.

GENERAL DISCUSSION

We view speech perception as a pattern recognition process that follows a general algorithm described by the FLMP (Massaro, 1998; Oden & Massaro, 1978). Perceivers are influenced by the speaker's face and voice, and visible speech is more influential when auditory

speech is less informative (Massaro, 1998). According to the FLMP, visible and audible speech are sources of information by which perceivers may recognize what is spoken. Information (e.g., a_i and v_j) available for the evaluation operation naturally varies across individuals or groups because of differences in the prototypes related to the perceiver's native language and linguistic experience. However, the information processing instantiated by the evaluation, integration, and decision operations appears to be universal across individuals and languages (Massaro, 1998).

This information processing is assumed to be invariant even though the information might differ. Different speakers may produce visible and audible speech segments that differ in terms of how informative they are. Other potential differences may arise from variations in the linguistic experience of the perceiver and the organizational structure of a language. Tonal languages may have a larger proportion of speech segments that are simply more distinguishable on the basis of sound. These differences will naturally lead to cross-linguistic differences even though the information processing of speech is universally consistent across languages (see Cutler, Demuth, & McQueen, 2002, for an analogous conclusion).

The present results are consistent with this conclusion. In Experiment 1 with two response alternatives, the degree of closeness of the test stimuli to their respective /ba/ and /da/ prototypes would probably differ for Mandarin and English native speakers. And as was expected, Mandarin speakers' response patterns for /ba/ and /da/ differed from those of English speakers (see also Massaro, 1998; Massaro et al., 1995; Massaro et al., 1993). However, both language groups were significantly influenced by visible and audible speech, and their responses were best described by the FLMP. Differences in information do not necessarily reflect interlanguage differences in the nature of audiovisual integration (information processing).

Experiment 2 was designed to explore the particular response patterns of the Mandarin speakers. Its findings are directly comparable to the previous cross-linguistic studies by Massaro et al. (1995; Massaro et al., 1993). Chinese participants in the present study chose /ba/, /da/, and /ga/ about 82% of the time. The English participants, on the other hand, chose /ba/, /da/, /va/, /ða/, and /za/ about 95% of the time (Massaro et al., 1995; Massaro et al., 1993). The results support the idea that the nature of information reflects differences in the phonemic repertoires, the phonetic realizations of the syllables, and the phonotactic constraints of the different languages (Massaro, 1998). For example, Mandarin does not have /va/, /ða/, or /θa/. As was expected, none of the Chinese participants reported perceiving "va," "ða," or "θa" in the pilot study, and in the experiment very few of the responses were *other* (1.1%).

Our research and modeling efforts have some potential implications for theories of speech perception. The FLMP is a formal mathematical or computational model of speech perception. It is a formalization of a set of theoretical principles: temporal extended information processing, continuous information, and pattern recognition processes. Our view is that this type of formalization and model testing offers a deeper scientific understanding than do verbal descriptions (Jacobs & Grainger, 1994). Although the answer to a qualitative question posed by a verbal theory is usually more informative than an answer to a quantitative one, qualitative questions are usually insufficient in scientific inquiry, and there is a need for

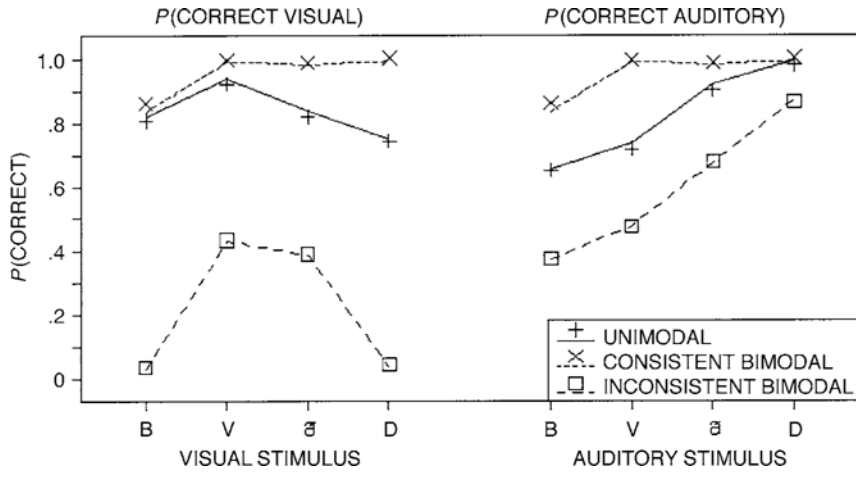

Figure 10. Observed (points) and predicted (lines) probability for correct visual (left panel) and auditory (right panel) responses as a function of visual or auditory level for unimodal, consistent, and inconsistent trials of English speakers perceiving English stimuli (data and graph from Massaro, 1998, p. 10). The consistent condition is necessarily identical in the two panels because it represents the same results. The lines are the predictions of the fuzzy logic model of perception.

quantification. Expressing a theory as a formal predictive model requires that it be fully specified, and its implementation for prediction makes immediately apparent any dimensions that are not completely specified. The experimental tests then provide an indication of how easily the theory can accommodate the findings and, if not, how it can be changed. Perhaps most important, formalization facilitates experimental tests among the theories. As we have shown elsewhere (Massaro, 1987, 1998), very different qualitative theories can make very similar or even identical predictions in certain conditions (as in a single-channel model and a WTAV). Thus, formalization allows the researcher to develop experimental conditions that can indeed distinguish between the theories of interest.

With these observations in mind, we briefly discuss the ramifications of our research for extant theories of speech perception. Psychoacoustic theorists focus on speech as a complex auditory signal without any reference to language-specific processes (e.g., Diehl & Kluender, 1987). Although they accept that visible speech can have an influence, they have not specified in a formal way how it does so. One embellishment of this theory would be to recast speech perception as multimodal and offer a psychophysical account, rather than limit it to a psychoacoustic account. Modeling their assumptions would be beneficial and would allow a more productive contrast with other theories.

The motor theory assumes that the perceiver uses the sensory input to best determine the set of articulatory gestures that produced this input (Liberman & Mattingly, 1985; Mattingly & Studdert-Kennedy, 1991). The functionality of visible speech is, of course, highly compatible with this theory, because visible speech can be considered to be an integral part of the sensory input reflecting the talker's articulatory gestures. The motor theory has not been sufficiently formalized, however, to account for the vast set of empirical findings on the integration of audible and visible speech. Once motor theorists have specified how audible and visible speech together allow the recovery of the articulatory gestures, it can be contrasted with the FLMP.

Speech perception has also been viewed as a module with its own unique set of processes and information. As was stated succinctly by Liberman (1996), "the phonetic module, a distinct system that uses its own kind of signal processing and its own primitives to form a specifically phonetic way of acting and perceiving" (p. 29). Within the context of our distinction between information and information processing, this statement implies that not only information, but also information processing, should be qualitatively different in the speech domain than in other domains of perceptual functioning. We have found, however, that the FLMP provides a good description of performance in domains other than speech perception (Massaro, 1998; Movellan & McClelland, 2001). For example, perceiving emotion from the face and the voice follows the same processing algorithm as that found for speech perception (Massaro, 1998, chaps. 7 and 8).

Consistent with our view, the direct perception theory assumes that speech perception does not follow unique processes (Fowler, 1996). Although gestures are the objects of speech perception, the speech motor system does not play a role in speech perception. In direct perception, persons directly perceive the causes of sensory input. The cause of an audible–visible speech percept is the vocal tract activity of the talker, and it is reasonable that both audible and visible speech should influence speech perception. Speech perceivers therefore obtain direct information from integrated perceptual systems from the flow of stimulation provided by the talker (Best, 1993). Formalizing this theory would allow it to be tested against the FLMP and other models, but until then it remains an open question whether the objects of speech perception are best considered to be the vocal activity of the speaker or relatively abstract symbols (Nearey, 1992) or an amodal motor representation (Robert-Ribes, Schwartz, & Escudier, 1995).

On the basis of just this short review of extant theories of speech perception, it is apparent that they are stated in verbal, rather than quantitative, form. Although no one can deny that a qualitative fact is more informative than a quantitative one, qualitative theories do not seem to be sufficiently precise to be distinguishable from one another. Very different theories make very similar predictions. Some quantitative refinement of the theories is usually necessary to create a chance for falsification and strong inference (Platt, 1964; Popper, 1959). Therefore, our strategy has been to quantify and test a family of specific models that represent the extant theories and also other reasonable alternatives.

In conclusion, the present study seems to suggest one central idea: There is no evidence to support the hypothesis of different types of audiovisual processing for speakers of different languages. A substantial body of research in second-language processing documents that perceivers process their first and second languages in a very similar manner (Flege, 2003). In a similar fashion, we do not believe that participants switch among different types of audiovisual processing depending on the physical properties of the stimulus and whether it is native or foreign. Differences in the pattern and magnitude of the McGurk effect may be attributed to idiosyncrasies in audiovisual stimuli in terms of the relative similarity to the perceivers' ideal prototypes in their language.

The underlying nature of audiovisual speech processing seems invariant across languages: Perceivers can use both sources of information, as is described by the FLMP. In a broader perspective that views perceiving speech as an instance of perception more generally (e.g., Massaro, 1998), just as the fundamental nature of visual and auditory processing is not expected to differ across cultural or linguistic groups (e.g., Gallistel, 2002), there is no reason to expect that the nature of audiovisual speech processing should differ.

Our results and conclusions reflect a more general principle of behavior. Although no two living entities behave exactly alike, there are certain essentials that we all share as humans. These principles are assumed to be universal, although individuals, groups, or cultures may exert a unique twist or idiosyncratic spin on a general fundamental process. For example, developmental research suggests that attachment types and patterns are found similarly across cultures, even though the specific behavioral expressions between children and their caregivers may differ, such as hugging in Western societies or handshaking in the Gusii of Kenya (e.g., Van Ijzendoorn & Sagi, 1999). At the social level, it has been claimed that societies of patriarchy encourage signals of female vulnerability, despite different specific manifestations across cultures (e.g., Donovan, 2001). Signs of restriction, whether Western high-heel shoes or Chinese foot-binding, and of female fertility, whether waist-narrowing in traditional Western costumes or hip-stuffing by some women in Africa, may be thought of as particular cultural expressions of certain underlying psychosocial ideologies.

Cognitive psychology is not new to this idea: Many theories at least implicitly are assumed to describe processes universal to humans. However, people obviously differ, making it important to understand how or why the fundamental nature of a process can be differentially expressed. Because visual and auditory information can influence speech perception, theories or models should describe not only audiovisual processing, but also exactly how different perceivers might be differentially influenced by the two sources of information. The FLMP is capable of describing the fundamental process and how it can be manifested differently in the behavior of individuals or groups.

REFERENCES

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

Best, C. (1993). Emergence of language-specific constraints in perception of non-native speech: A window on early phonological development. In B. de Boysson-Bardies & S. de Schonen (Eds.), *Developmental neurocognition: Speech and face processing in the first year of life* (pp. 289-304). Norwell, MA: Kluwer.

Burnham, D., Lau, S., Tam, H., & Schoknecht, C. (2001). Visual discrimination of Cantonese tone by tonal but non-Cantonese speakers, and by non-tonal language speakers. In *Proceedings of International Conference on Auditory–Visual Speech Processing* (pp. 155-160), Sydney, Australia. Available from http://www.isca-speech.org.archive/avsp98.

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

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

Cutler, A., Demuth, K., & McQueen, J. M. (2002). Universality versus language-specificity in listening to running speech. *Psychological Science*, **13**, 258-262.

De Gelder, B., & Vroomen, J. (1992). Auditory and visual speech perception in alphabetic and non-alphabetic Chinese–Dutch bilinguals. In R. J. Harris (Ed.), *Cognitive processing in bilinguals* (pp. 413-426). Amsterdam: Elsevier.

Diehl, R. L., & Kluender, K. R. (1987). On the categorization of speech sounds. In S. Harnad (Ed.), *Categorical perception* (pp. 226-253). Cambridge: Cambridge University Press.

Donovan, J. (2001). *Feminist theory: The intellectual traditions* (3rd ed.). New York: Continuum.

Flege, J. E. (2003). Assessing constraints on second-language segmental production and perception. In A. Meyer & N. Schiller (Eds.), *Phonetics and phonology in language: Comprehension and Production: Differences and Similarities*. Berlin: Mouton de Gruyter.

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America*, **99**, 1730-1741.

Gallistel, C. R. (2002). Language and spatial frames of reference in mind and brain. *Trends in Cognitive Sciences*, **6**, 321-322.

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

Hayashi, Y., & Sekiyama, K. (1998). Native–foreign language effect in the McGurk effect: A test with Chinese and Japanese. In *Proceedings of Auditory–Visual Speech Processing 1998* (pp. 61-66), Sydney, Australia. Available from http://www.isca-speech.org/archive/avsp98.

Jacobs, A. M., & Grainger, J. (1994). Models of visual word recognition: Sampling the state of the art. *Journal of Experimental Psychology: Human Perception & Performance*, **20**, 1311-1334.

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

Liberman, A. M. (1996). *Speech: A special code*. Cambridge, MA: MIT Press.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, **21**, 1-36.

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

Massaro, D. W. (1988). Some criticisms of connectionist models of human performance. *Journal of Memory & Language*, **27**, 213-234.

Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology*, **21**, 398-421.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

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.

Massaro, D. W., Weldon, M. S., & Kitzis, S. N. (1991). Integration of orthographic and semantic information in memory retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **17**, 277-287.

Mattingly. I. G., & Studdert-Kennedy, M. (Eds.) (1991). *Modularity and the motor theory of speech perception.* Hillsdale, NJ: Erlbaum.

Movellan, J. R, & McClelland, J. L. (2001). The Morton–Massaro law of information integration: Implications for models of perception. *Psychological Review*, **108**, 113-148.

Nearey, T. M. (1992). Context effects in a double-weak theory of speech perception. *Language & Speech*, **35**, 153-171.

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

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

Popper, K. R. (1959). *The logic of scientific discovery*. New York: Basic Books.

Robert-Ribes, J., Schwartz, J.-L., & Escudier, P. (1995). A comparison of models for fusion of the auditory and visual sensors in speech perception. *Artificial Intelligence Review*, **9**, 323-346.

Sekiyama, K. (1997). Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. *Perception & Psychophysics*, **59**, 73-80.

Sekiyama, K., & Tohkura, Y. (1989). Effects of lip-read information on auditory perception of Japanese syllables [Abstract]. *Journal of the Acoustical Society of America*, **85**(1, Suppl.), 138.

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.

Tiippana, K., Sams, M., & Andersen, T. S. (2001). Visual attention influences audiovisual speech perception. In D. W. Massaro, J. Light, & K. Geraci (Eds.), *Proceedings of Auditory–Visual Speech Processing* (pp. 167-171). Aalborg. Available from http://www.isca_speech.org/archive/avsp01.

Van Ijzendoorn, M. H., & Sagi, A. (1999). Cross-cultural patterns of attachment: Universal and contextual dimensions. In J. Cassidy & P. R. Shaver (Eds.), *Handbook of attachment: Theory, research, and clinical applications* (pp. 713-734). New York: Guilford.

Zadeh, L. A. (1965). Fuzzy sets. *Information & Control*, **8**, 338-353.

APPENDIX A
Root-Mean Squared Deviation Values of Each Participant for Each Model (Experiments 1 and 2)

	Model		
Speakers	WTAV	ADM	FLMP
	Experiment 1		
Mandarin 1	.1065	.0837	.0586
Mandarin 2	.0791	.0692	.0452
Mandarin 3	.0605	.0507	.0376
Mandarin 4	.0226	.0760	.0207
Mandarin 5	.0321	.1103	.0204
Mandarin 6	.1180	.1248	.0745
Mandarin 7	.0700	.0599	.0765
English 1	.1021	.0492	.0986
English 2	.1120	.0577	.0547
English 3	.1489	.1067	.0633
English 4	.1063	.0964	.0570
English 5	.0718	.1340	.0235
English 6	.0592	.1069	.0304
English 7	.0656	.1115	.0542
	Experiment 2		
Mandarin 1	.0566	.0569	.0505
Mandarin 2	.0803	.0794	.0617
Mandarin 3	.0898	.0823	.0788
Mandarin 4	.0683	.0574	.0639
Mandarin 5	.0641	.0589	.0718
Mandarin 6	.0626	.0577	.0562
Mandarin 7	.0652	.0657	.0534

Note—WTAV, weighted-averaging model; ADM, auditory dominance model; FLMP, fuzzy logic model of perception.

APPENDIX B
Individual Root-Mean Squared Deviation Values From Experiment 2: Responses Grouped Into Three (/ba/, /da/, & *Other*) and Two (Labials and Nonlabials) Categories

	/ba/, /da/, and *Other*		Labials and Nonlabials	
Participant	ADM	FLMP	ADM	FLMP
Mandarin 1	.1115	.0880	.1556	.0695
Mandarin 2	.1202	.1070	.1511	.1475
Mandarin 3	.1901	.1396	.1498	.1505
Mandarin 4	.1560	.1183	.1199	.1012
Mandarin 5	.1432	.1215	.1357	.1080
Mandarin 6	.1293	.0970	.1574	.0946
Mandarin 7	.1210	.0845	.0890	.0811

Note—ADM, auditory dominance model; FLMP, fuzzy logic model of perception.

(Manuscript received January 1, 2003;
revision accepted for publication October 14, 2003.)