Journal of Phonetics (1993) **21**, 445–478

Bimodal speech perception: an examination across languages

Dominic W. Massaro, Michael M. Cohen, Antoinette Gesi and Roberto Heredia

Program in Experimental Psychology, University of California, Santa Cruz, Santa Cruz, CA 95064, U.S.A.

Minoru Tsuzaki

ATR Auditory & Visual Perception Research Laboratories, Seika-cho, Kyoto, 619-02, Japan

Received 28th October 1991, and in revised form 11th November 1992

We examined whether language and culture influence speech perception in face-to-face communication. Native speakers of Japanese, Spanish and English identified the same synthetic unimodal and bimodal speech syllables. Five-step /ba/–/da/ continua were synthesized along auditory and visual dimensions, by varying properties of the syllable at its onset. In the first experiment, the three language groups identified the test syllables as /ba/ or /da/; in the second, Japanese and English speakers were given an open-ended set of response alternatives. For all language groups, identification of the speech segments was influenced by both auditory and visual sources of information. Given the results, we were able to reject an auditory dominance model (ADM) which assumes that the contribution of visible speech is dependent on poor-quality audible speech. The results also falsified a categorical model of perception (CMP) in which the auditory and visual sources are categorized before they are 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. No differences in the nature of processing across language groups suggests that the underlying mechanisms for speech perception are similar across language and culture.

1. Introduction

Speech perception has been studied extensively in the last decade. We have learned that people use many sources of information in perceiving and understanding speech. One interesting observation is that people manage to communicate under the most adverse conditions imaginable. In one series of investigations, researchers have examined the important contribution of visible information in face-to-face

Please address all correspondence to: Dr Dominic W. Massaro, Program in Experimental Psychology, Clark Kerr Hall, University of California, Santa Cruz, Santa Cruz, CA 95064, U.S.A.

0095-4470/93/040445 + 34 $08.00/0

communication. These experiments have shown that visible speech is particularly helpful when the auditory speech is degraded due to noise, bandwidth filtering or hearing-impairment (Summerfield, 1979, 1983; Breeuwer & Plomp, 1984; Massaro, 1987). Although the influence of visible speech is substantial when auditory speech is degraded, visible speech also contributes to performance even when paired with intelligible speech sounds. The importance of visible speech is most directly observed when conflicting visible speech is presented with intelligible auditory speech. As an example, the auditory syllable /ba/ might be dubbed onto a videotape of a talker saying /da/ (McGurk & MacDonald, 1976). A strong effect of the visible speech is observed because the subject will often report perceiving (or even hearing) the syllable /ða/. Thus, the strong influence of visible speech is not limited to situations with degraded auditory input, but it also appears to have an important influence even when paired with perfectly intelligible speech sounds.

Although the study of bimodal speech perception has been primarily carried out with English talkers, it offers a valuable domain for the study of cross-linguistic and cross-cultural differences and similarities. It is important to know to what extent the results to date are dependent on language and culture. In addition, cross-linguistic and cross-cultural differences offer a powerful paradigm for broadening the domain for inquiry (Massaro, 1992). Our empirical findings, theories and models often tend to be limited to highly specific situations. Cross-linguistic studies allow us to determine the degree to which we can generalize our conclusions across language and culture.

Our task manipulates synthetic auditory and visual speech in an expanded factorial design, as shown in Fig. 1. Five levels of audible speech varying between

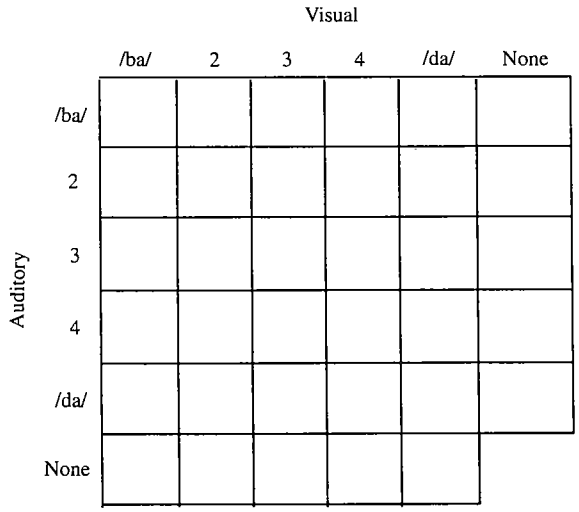

Figure 1. Expanded factorial design used in the current experiments to include both bimodal speech and auditory and visual conditions presented alone. The five levels along the adutory and visible continua represent auditory and visible speech syllables varying in equal physical steps between /ba/ and /da/. For the auditory continuum, /ba/ corresponds to rising F_2 and F_3 transitions and /da/ corresponds to falling F_2 and F_3 transitions. For the visual continuum, /ba/ corresponds to closed lips at the onset of the syllable and /da/ corresponds to open lips at onset.

/ba/ and /da/ are crossed with five levels of visible speech varying between the same alternatives. The onsets of the second and third formants are varied to give an auditory continuum between the syllables /ba/ and /da/. In analogous fashion, parameters of an animated face are varied to give a continuum between visual /ba/ and /da/. This design allows us 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).

2. Models of bimodal speech perception

We adhere to a falsification and strong-inference strategy of inquiry (Platt, 1964; Massaro, 1987, 1989*a*). Results are informative only to the degree that they distinguish among alternative theories. Thus, the experimental task, data analysis and model testing are devised specifically to reject some theoretical alternatives. A fuzzy logical model of perception (FLMP), an auditory dominance model (ADM), and a categorical model of speech perception (CMP) are formalized and tested against the results. The FLMP has been the most successful model to date (Massaro, 1987, 1989*b*, 1990; Massaro & Friedman, 1990) and we begin with the description of this model.

2.1. *Fuzzy logical model of perception*

The results from a wide variety of experiments have been described within the framework of the FLMP. Within the present framework, speech perception is robust because there are usually multiple sources of information that the perceiver evaluates and integrates to achieve perceptual recognition. The assumptions central to the model are: (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 Fig. 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.

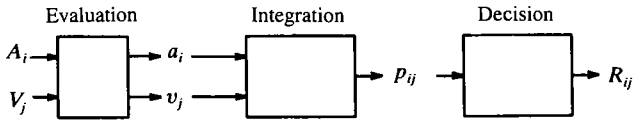

Figure 2. Schematic representation of the three operations involved in perceptual recognition. The evaluation of an auditory source of information A_i produces a truth value a_i, indicating the degree of support for alternative R. The visual source V_j is evaluated similarly to give v_j. Integration 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.

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. Defining the onsets of the second (F_2) and third (F_3) formants as the important auditory feature and the degree of initial opening of the lips as the important visual feature, the prototype for /da/ might be something like:

/da/ Slightly falling F_2–F_3 and Open lips.

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

/ba/ Rising F_2–F_3 and Closed lips

and so on for the other prototypes.

Given a prototype's independent specifications for the auditory and visual sources, the value of one source cannot change the value of the other source. The integration of the features defining each prototype is evaluated according to the product of the feature values. We let a_{Di} represent the degree to which the auditory stimulus A_i supports the alternative /da/, that is, has Slightly falling F_2–F_3. Similarly, v_{Dj} represents the degree to which the visual stimulus V_j supports the alternative /da/, that is, has Open lips. It is assumed that the outcome of prototype matching for /da/ would be a multiplicative contribution of the auditory and visual support:

$$S(/da/ \mid A_i \text{ and } V_j) = a_{Di} \times v_{Dj} \tag{1}$$

where $S(/da/ \mid A_i \text{ and } V_j)$ 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 modalites, respectively. Analogously, if a_{Bi} represents the degree to which the auditory stimulus A_i has Rising F_2–F_3 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:

$$S(/ba/ \mid A_i \text{ and } V_j) = a_{Bi} \times v_{Bj} \tag{2}$$

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, $P(/da/)$, is predicted to be:

$$P(/da/ \mid A_i) = \frac{a_{Di}}{\sum_k a_{ki}}. \tag{3}$$

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

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

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

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

As can be seen in Equations (1) and (2), the absolute support for a given prototype will be less for two sources of information than just one. However, the identification judgement 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, Chapter 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 five free parameters for the visual feature values and five for the auditory feature values for each response alternative. (The procedure for estimating the free parameters for the fit of the models is given in Section 4.2.3.)

2.2. *Auditory dominance model*

A second potential explanation of bimodal speech perception is that an effect of visible speech occurs only when the auditory speech is not completely intelligible (Sekiyama & Tohkura, 1991). Sekiyama & Tohkura tested four labial and six non-labial consonants in the context /a/, under auditory and auditory–visual conditions. The auditory speech was presented either in quiet or in noise. As expected, identification of the auditory speech was very good in quiet and poor in noise. The influence of visible speech in the bimodal condition depended on the quality of the auditory speech. There was very little visual influence with good-quality auditory stimuli and substantial influence with poor-quality auditory speech. For many alternatives, visible speech had an influence for only those auditory stimuli that were not perfectly identified in the auditory condition. However, there were exceptions to this general trend. The auditory syllable /ma/ was perfectly identified in the auditory condition, but was identified as non-labial about 6% of the time when it was paired with a non-labial visible articulation.

The hypothesis that auditory intelligibility determines whether or not visible speech will have an effect is difficult to test, primarily because intelligibility is not easily defined. Perfect identification in one test might not mean perfect intelligibility. Even given these limitations in the measure of intelligibility, we can formulate one version of an intelligibility model, called the auditory dominance model (ADM). The central assumption of the ADM is that the influence of visible speech given a bimodal stimulus is solely a function of whether or not the auditory speech is identified correctly. This model appears to capture Sekiyama & Tohkura's (1991, p. 1804) conclusion that "human beings may depend on eyes in the presence of auditory uncertainty". Similarly, Vroomen (1992) describes (but does not defend) the possibility of lip-reading as a backup device. In this case, the visual information "is relied on whenever the auditory signal is ambiguous". (Vroomen, 1992, p. 9). These views lend themselves to the current instantiation of the ADM in which

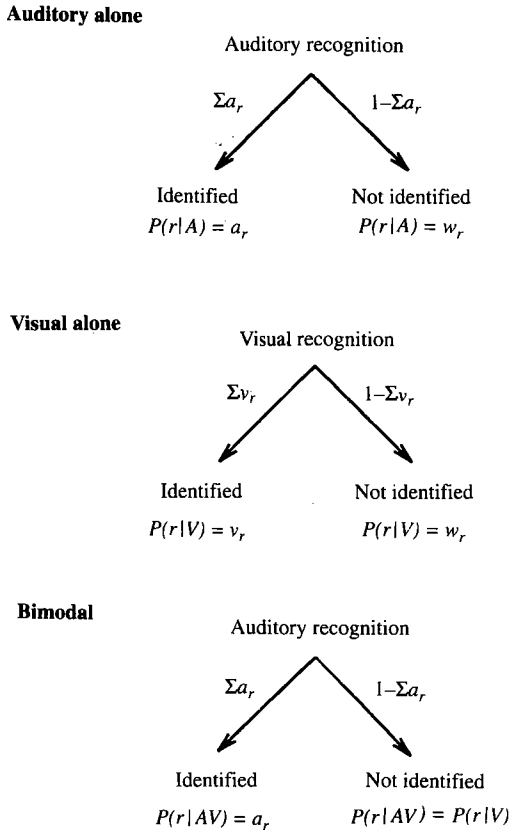

Figure 3. Decision trees for ADM for auditory alone, visual alone, and bimodal trials. See text for explanation.

visible speech has a possible influence *only* when the auditory speech is not identified. 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 shown in the top panel of Fig. 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 or 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

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

where a_r is the probability of identifying the auditory source as response r, $\sum_r a_r$ is

the probability of identifying the auditory source as any of the response alternatives, and the term $\left(1 - \sum_r a_r\right)$ is the probability of not identifying the auditory source.

For visual alone trails the situation is analogous. As shown in the middle panel of Fig. 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 or she responds with that alternative. In the case that no identification is made the subject responds with a given alternative with the bias probability w_r. Therefore, the predicted probability of a response on visual alone trials is equal to

$$P(r \mid V) = v_r + \left(1 - \sum_r v_r\right) w_r, \tag{7}$$

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

Finally, we consider the bimodal case, shown in the bottom panel of Fig. 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 or 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

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

Equation (8) represents the theory that the auditory stimulus is either identified or else the subject bases his or 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, then this model should give a better description of the results than 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. In the FLMP, the influence of visible speech in bimodal speech perception is a direct function of its influence in the identification of visible speech in isolation. A good lip-reader will necessarily show some effect of visible speech in bimodal perception. In the ADM, a subject might be a good lip-reader given just visible speech and show very little influence of visible speech in bimodal perception.

2.3. *Categorical model of perception*

In the categorical model of perception (CMP), it is assumed that only categorical information is available from the auditory and visual sources and that the response is

TABLE I. The probabilities of the four possible outcomes of the two unimodal categorizations of a bimodal speech stimulus for the CMP

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

based on separate categorizations of the auditory and visual sources. The four possible cases are shown in Table I. If the two categorizations to a given speech event agree, the single possible identification response can be based on either source. When the two categorizations disagree, it is assumed that the subject will respond with the categorization to the auditory source on some proportion p of the trials, and with the categorization to the visual source on the remainder $(1-p)$ of the trials. The weight p reflects the relative dominance of the auditory source. Considering a /ba/ response, the visual and auditory categorizations could be /ba/–/ba/, /ba/–not /ba/, not /ba/–/ba/ or not /ba/–not /ba/.

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

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

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

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

For each response alternative, the CMP 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 CMP is mathematically equivalent to both a single channel model in which the subject attends to just one modality on bimodal trials (Thompson & Massaro, 1989) and a weighted averaging model in which the subject simply performs a weighted averaging of the two modalities (Massaro, 1987).

3. Previous results and extension to other languages

Experiments using synthetic auditory and visual speech have been carried out with native English-speaking Americans as subjects (Massaro & Cohen, 1990). These subjects give a variety of responses when they are given a range of response alternatives. Both audible and visible speech has a strong influence on performance. In addition, the contribution of one source is larger to the extent the other source is

ambiguous. The details of these judgements were nicely captured in the predictions of the FLMP.

The goal of the present research is to determine if bimodal speech is processed in the same manner across three languages. The FLMP has provided the best description of previous results with English speakers (Massaro, 1989*a,b,* 1990). The question, thus, reduces to asking whether it will also give a superior description of the results from Japanese and Spanish speakers. We can thus assess how language and culture influence unimodal and bimodal speech perception.

We can speculate about what results might be expected from Japanese and Spanish speakers relative to English. All three languages have /b/ and /d/ segments (Maddieson, 1984). Bilingual speakers of any pair of these languages usually claim that these segments are roughly equivalent across the two languages. However, we can be sure that the ideal auditory and visual speech will not be equivalent for these segments across the three languages. The /d/ is more dental for Spanish speakers, for example. The vowel /a/ is shorter in Japanese than in English and Spanish, but the Japanese might interpret the vowel as their long-vowel /ba:/ and /da:/. Differences in phonetic realizations across the languages should have some influence on performance in our task.

The phonological inventories of these three languages also differ from one another. Unlike English, Japanese does not have the phonemes /ð/ or /v/, and American Spanish also does not have the phoneme /v/. These differences have important consequences for the outcome of bimodal speech perception. The syllables /va/ and /ða/ are frequent response alternatives when auditory and visual speech are varied along a /ba/ to /da/ continuum. These alternatives are reasonable because of the auditory and visible properties of these segments. Our research has shown that perceivers respond with alternatives that have the best fit with both the auditory and visual information. Presented with an auditory /ba/ paired with a visible /da/, we might expect the perceiver to respond with one of these two alternatives. However, this is not the case because there is a complete mismatch on one of the two sources of information. On the other hand, the response alternative /ða/ is reasonable. Auditory /ba/ is more similar to auditory /ða/ than /da/, and visible /da/ is also more similar to visible /ða/ than /ba/. Thus, with open-ended response alternatives we would expect that English speakers would respond /ða/ given an auditory /ba/ paired with a visible /da/.

Following this logic, subjects whose language does not have the segment /ð/ should behave differently in this task with open-ended alternatives. We might even expect somewhat different results from subjects who have only learned English as a second language. Mills & Theim (1980) tested native German speakers who had learned English as a foreign language. These subjects identified English bimodal CV syllables consisting of conflicting auditory and visual information. The 15 syllables represent distinctive phonetic categories. The phoneme /ð/ does not occur in German but was considered to be familiar enough to the subjects, who were native speakers of German but had learned English as a foreign language. Both the auditory and visual components had strong effects on identifying what the speaker had said. With respect to identification of the phoneme /ð/, a visual /v/ paired with auditory /ð/ never produced the identification of /ð/. A visual /ð/ paired with auditory /v/ gave 33% /ð/ responses. This result contrasts with the result for English subjects in which a visual /v/ plus auditory /ð/ gave 17% /ð/ responses and

a visual /ð/ plus auditory /v/ gave over 80% /ð/ responses (Massaro, 1987, Chapter 2, Fig. 6). These differences illustrate how the influence of both audible and visible speech is modulated by the perceiver's native language. Analogously, with open-ended response alternatives we would expect fewer, if any, /ða/ responses for the Japanese and Spanish perceivers. Given different phoneme inventories, we would expect different performance in unimodal and bimodal speech perception. Relative to English, the Japanese and Spanish phoneme space can be considered to be less cluttered around /ba/ and /da/. Given an auditory /ba/ paired with a visual /da/, Japanese and Spanish speakers would be more likely to respond with one of these two alternatives.

Another frequent response for English speakers is the consonant cluster /bda/ when the stimulus is a visible /ba/ paired with an auditory /da/. This perceptual judgement is reasonable if subjects choose a response that is the most consistent with both the auditory and visual information. Visible /bda/ is similar to visible /ba/ because both alternatives begin with closure of the lips followed by a mouth opening. Audible /da/ is also similar to audible /bda/ because both segments begin and end with the same formant transitions. Other alternatives given a visual /ba/ and auditory /da/ are less plausible. The alternative /dba/, for example, is not reasonable because of the huge mismatch of visible /dba/ with visible /ba/. The syllable /ba/ begins with a closing of the lips whereas the alternative /dba/ begins with a mouth opening followed by closure of the lips.

In addition to the different phoneme inventories, we would expect cross-lingusitic differences because of the differences in the sequential occurrence of consonant segments in these languages. Consonant clusters, such as /bl/, occur in initial position in English. The /bda/ cluster occurs in simple and compound words and across word boundaries. English has words like *abdicate* and *subdue* and word sequences like *crab dish*. Considering the general case of a cluster consisting of labial followed by an alveolar or velar consonant, we found in a word list composed of the Oxford Unabridged Dictionary plus the standard UNIX word list 4 648 cases in 296 666 words or 1.567%. In most of these cases, the clusters occur at syllable boundaries, i.e., a syllable-final consonant followed by a syllable-initial consonant, with some exceptions occurring for the /pt/ cluster, e.g., "apt" and "crypt". In our task, however, the cluster responses are given as syllable initial. Given that other consonant clusters do occur initially in English, /bda/ might be perceived initially even though it only occurs non-initially.

Consonant clusters do not occur initially in the Kansai dialect of Japanese spoken by our subjects. In addition, stop consonant clusters do not occur across word boundaries in Japanese because all words end in vowels or nasals. In a word list of 115 600 Japanese words there were only three occurrences of labial followed by alveolar or velars (0.003%). Compared to English, there are fewer consonant clusters in Spanish. This is confirmed with a count from a word list of 86 061 Spanish words in which there were 557 occurrences of labial followed by alveolar or velar consonant (0.647%). In addition, consonant clusters are also less likely to occur across word boundaries in Spanish than English because a greater proportion of Spanish words end in a vowel. Thus, even if subjects are given an open-ended set of response alternatives, we might expect that Japanese and Spanish speakers would be less likely to respond /bda/ given a visible /ba/ and an auditory /da/. In this case, they might show less of an influence from the visible speech and be more likely to respond /da/.

4. Experiment 1: two alternatives

We therefore face the problem of comparing the processing of speech across languages when these languages differ in their phoneme inventories. It can be argued that the phoneme inventories should play less of a role if subjects are limited to just two responses, /ba/ and /da/. In this case, there should be no difference across languages with respect to the number of prototypes that are functional in the task. English, Spanish and Japanese languages have /ba/ and /da/ syllables. In this case, the additional /va/, /ða/, and /bda/ prototypes for the English speakers should have no influence.

According to the FLMP, performance according to the relative goodness rule (RGR) at decision should be a function of only the relevant alternatives in the task (in this case, only /ba/ and /da/). Consider a hypothetical situation of 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 for all speakers is predicted to be

$$P(/da/ \mid A_i \text{ and } V_j) = \frac{a_{Di} \times v_{Dj}}{a_{Di} \times v_{Dj} + (1 - a_{Di}) \times (1 - v_{Dj})}, \tag{11}$$

where a_{Di} is the auditory support for /da/ and v_{Di} 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 Equations (1)–(5) it can be shown that the model makes equivalent predictions if a_{Bi} is assumed to be equal to $(1 - a_{Di})$ and v_{Bj} is assumed to be equal to $(1 - v_{Dj})$ (Massaro, 1989*a*). Given the success of the FLMP with English speakers, the hypothesis that all speakers process speech in the same manner predicts that the equation will give an equally good description of Japanese and Spanish speakers.

In our previous research, it has been important to distinguish between information and information processing. Information refers to just the output of the evaluation operation in the FLMP (see Fig. 2). Information processing refers to the nature of the evaluation, integration and decision operations, not the input to or output from these operations. Our study primarily addresses differences in information processing across the three 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 describe the results for speakers of all three languages.

An alternative hypothesis predicts that the FLMP will fail to describe differences across the three language groups. For example, suppose that one language group is less influenced by visible speech in bimodal speech perception—as has been proposed for Japanese perceivers relative to English speakers (Sekiyama & Tohkura, 1993). If this hypothesis is correct, then Equation (11) should fail. Japanese should be about as accurate as English speakers in identifying visible speech when it is presented alone (without the auditory signal), but should use this information less in the identification of bimodal speech. If the Japanese process speech in this manner, the FLMP should give a poor description of their results

because the FLMP cannot predict this type of selective weighting of one of the two sources of information.

Although we have concentrated on potential linguistic differences, cultural differences might also contribute to performance differences among the three language groups. Speakers within one culture might interact differently with one another in face-to-face communication, relative to speakers within another culture. Speakers of one language might tend to avoid face-to-face contact and therefore, not be influenced by visible speech. As a consequence, these speakers would be expected to learn less about visible speech and to be poorer lip-readers. In this case, the FLMP should still predict the results even though there would be less of an influence of visible speech. Although cultural differences are confounded with language differences in the present study, it is still important to acknowledge both of these potential contributions.

The discussion concerning the FLMP and linguistic differences also applies equally to the ADM and CMP. 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.

4.1. *Method*

4.1.1. *Subjects*

Three different subject populations were sampled for this 2 h experiment. The subjects spoke American English, Japanese or American Spanish as their first language. The English speakers were 21 students from the University of California, Santa Cruz. The Japanese subjects were 21 students from Dohshisha University. Their ages ranged from 18 to 20-years-old. These subjects spoke the Kansai dialect and their study of English began at age 13 with the "standard" English lessons at junior high school. The average duration of English instruction was eight years. The focus of their English classes was on reading and writing, but not on listening or speaking. None of the students had lived abroad nor taken any lessons in English outside of school. The 20 native Spanish speakers were also from the University of California, Santa Cruz community. The majority of the subjects were Mexican-Americans and three were Puerto Rican-Americans. The Spanish speakers spoke English to various degrees of proficiency. Seventy-two percent of the subjects reported learning English in elementary school (K–5th grade), and 16% reported learning English in ESL (English as a second language) classes in secondary school. Twelve percent reported learning both Spanish and English simultaneously. The Spanish speakers had an average of 14.5 (SD = 4.9) years speaking English.

Some of the English speakers were recruited from the subject pool who were then given 2 h credit for participating. The rest of the English speakers and the Spanish-speaking subjects were recruited by posted advertisements on the campus, through mutual friends, and word of mouth. The English and Spanish speakers were paid $10.00. The Japanese were paid 4000 yen or the equivalent of roughly $30 for this experiment and another 1 h experiment having to do with discrimination of auditory speech. Each of the three groups had a different experimenter who was a native speaker of the subject's native language, and only the native language was used during the experiment.

4.1.2. *Apparatus and materials*

The stimuli were presented by the means of auditory and visible synthetic speech. Using an auditory speech synthesizer, we created a continuum of five sounds that varied between a good /ba/ and a good /da/. The first sound was a good /ba/. The fifth sound was a good /da/. The middle sound was halfway between /ba/ and /da/. The second sound was somewhat more /ba/-like and the fourth sound was somewhat more /da/-like. In an exactly analogous manner using computer animation, we synthesized a face saying /ba/ and /da/ and also saying three syllables intermediate between them. Thus, a five-step continuum going from /ba/ to /da/ was created.

Synthetic audible speech. Tokens of the first author's /ba/ and /da/ were analyzed 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 ms of the consonant–vowel syllable, a set of five 400 ms syllables covering the range from /ba/ to /da/ was created. Figures 4(b) and 4(c) 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 ms, the first formant (F_1) went from 250 Hz to 700 Hz following a negatively accelerated path. The F_2 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 F_3 followed a linear transition to 2729 Hz from one of five values equally spaced between 2387 and 2637 Hz. All other stimulus charac-

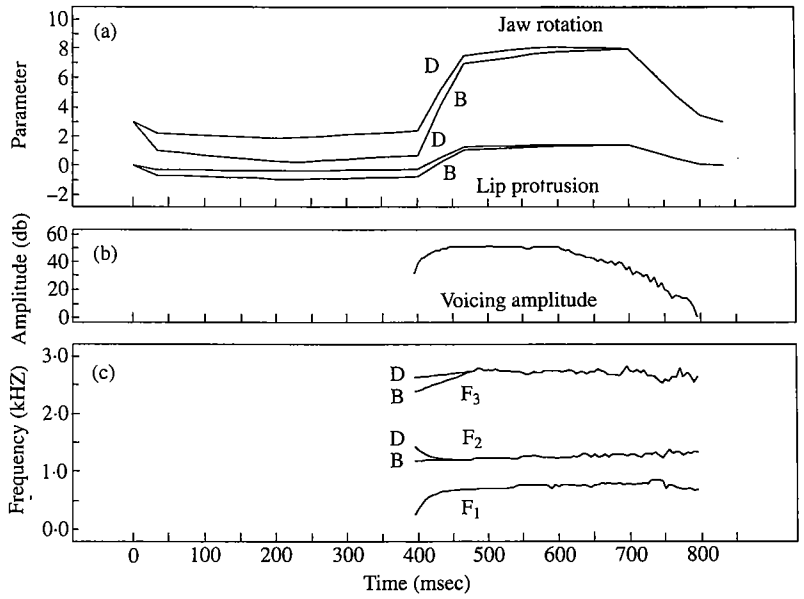

Figure 4. Visual and auditory parameter values over time for visual /ba/ and /da/ stimuli and auditory /ba/ and /da/ stimuli. (a) Jaw rotation and lip protrusion; (b) voicing amplitude; (c) formants F_1, F_2 and F_3. See text for details.

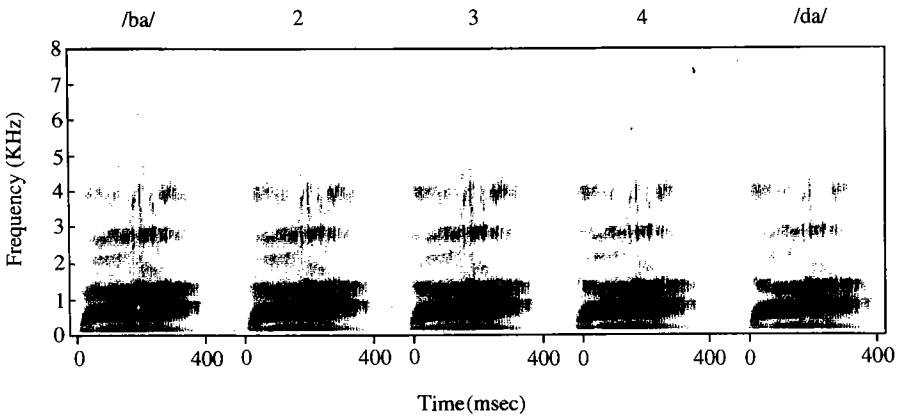

Figure 5. Spectrograms for the five levels of auditory speech between /ba/ and /da/.

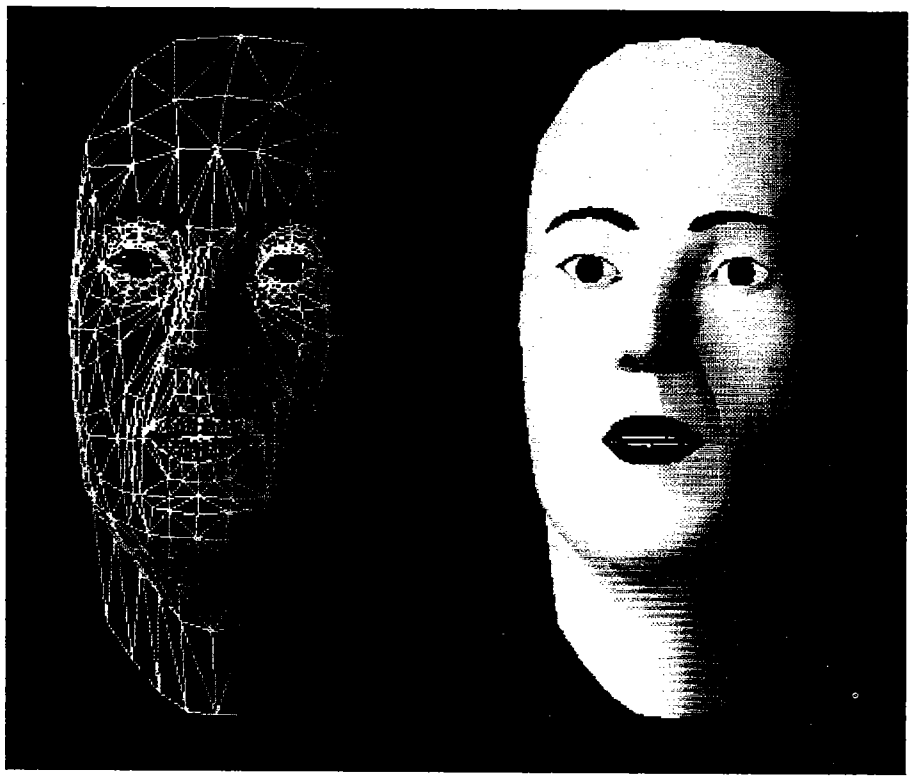

Figure 6. Framework (left) and Gouraud shaded (right) renderings of polygon facial model.

teristics were identical for the five auditory syllables. Figure 5 gives the spectrograms of the five syllables along the continuum.

Synthetic visible speech. As employed in 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 3D, joined together at the edges (Parke, 1974, 1975, 1982). The left panel of Fig. 6 shows a framework rendering of this model. To achieve a natural appearance, the surface was smooth shaded using Gouraud's (1971) method (shown in the right panel of Fig. 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 used for several demonstration sentences were selected and refined by the investigator by studying his own articulation frame by frame and estimating the control parameters values (Parke, 1974). Each phoneme is defined in a table according to target values for segment 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

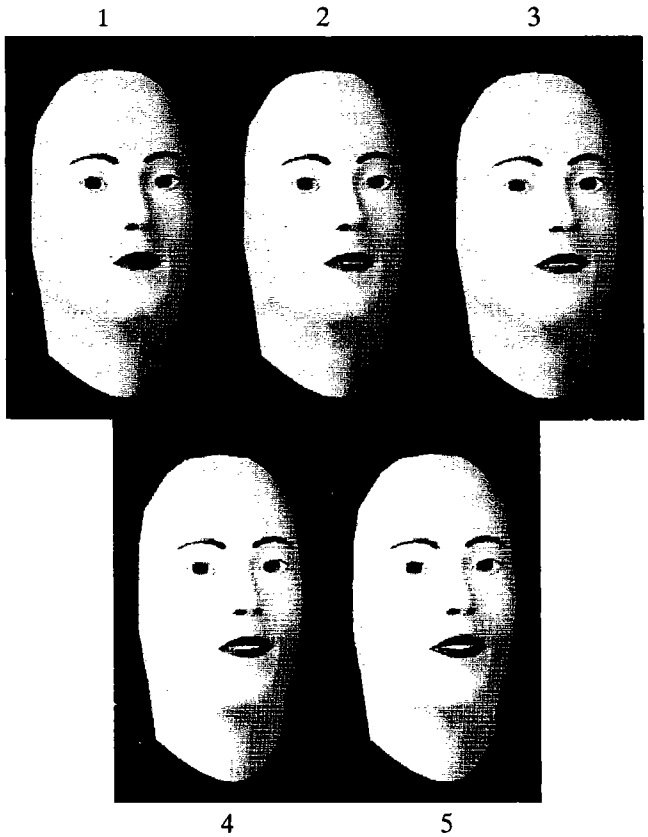

Figure 7. The facial model at the onset of the syllable for each of the five levels of visible speech between /ba/ and /da/.

TABLE II. 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

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 & Hill (1986) and ourselves (Cohen & Massaro, 1990) was implemented on a Silicon Graphics Inc. IRIS 3030 computer. We adapted the software to allow new intermediate test phonemes. To create an animation sequence, each frame was recorded using 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 II 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/. Figure 4(a) shows how the visual synthesis parameters changed over time for the first (/ba/) and last (/da/) visual levels. For clarity, only two of the visual parameters are shown—jaw rotation (larger parameter value means more open), and lip protrusion ("mouth z offset" in Table II, smaller value means more protrusion). Not shown in the figure, the face with the default parameter values was recorded for 2000 ms preceding and 2000 ms following the time shown for a total visual stimulus of 4866 ms. A dark screen was presented for the auditory alone trials.

Following the synthesis a Betacam tape was dubbed to 3/4″ U-Matic for editing. Only the final 4766 ms 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 ms 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.

4.1.3. *Design and procedure*

In this experiment, synthetic auditory and visual speech were manipulated in an expanded factorial design previously illustrated in Fig. 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.

Subjects were instructed to listen and to watch the speaker, and to identify the syllable as /ba/ or /da/. The native English speakers and the native Spanish speakers were tested at the University of California, Santa Cruz. The native Japanese speakers were tested at ATR in Japan. The same procedure was used for the three different groups with a few minor differences listed below. All subjects were tested on the same videotape. They all received 10 practice trials and the number of test trials were 840 ($35 \times 6 \times 4$). Thus there were 24 observations at each of the 35 unique experimental conditions. Subjects were given a short break after every 210 trials. The display monitor subtended a visual angle of 39 degrees. For the English speakers, the experimental tape was played to the subjects over individual NEC model C12-202A 12-inch color monitors. For the Japanese subjects, the monitor was a National (Panasonic) 14″ diagonal screen TH-15B1 and the video tapes were played on a National (Panasonic) AG-6500 VHS recorder. For all three groups, the loudness level of the auditory stimuli was 79 dB (A). The measurement was done with the sound level meter (B&K 2231, with the Microphone Type 4133; Time Weighting, "Fast"; Frequency Weighting, "A"; Display Parameter, "SPL"). The background noise level was 47.5 dB (A).

For the English and Spanish speakers, up to four subjects could be tested simultaneously in individual sound-attenuated rooms. These rooms were each illuminated by two 60 watt incandescent bulbs in a frosted glass ceiling fixture. These subjects made their responses by pressing a key labeled "ba" or "da" on the terminal keyboard. The reaction times (RTs) of these key presses were measured. The Japanese subjects wrote their responses in kana on 3×5″ note cards. The experimenter entered these responses into the terminal after the experiment was over. In all cases, the experimenter was a native speaker of the subject's native language and all instructions and interactions were in the native language.

4.2. *Results*

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 judgements. As illustrated in Fig. 8, the proportion of responses changed systematically across the visual continuum, both for the unimodal, $F(4, 236) = 244.10$, $p < 0.001$, and the bimodal, $F(4, 236) = 91.03$, $p < 0.001$ conditions. Note that the four degrees of freedom for the numerator comes from the five levels of the stimulus while the 236 degrees of freedom for the denominator comes from the 62 subjects minus the three groups, times the numerator degrees of freedom. Similarly, the pattern of responses changed in an orderly fashion across the auditory continuum, for both the unimodal, $F(4, 236) = 1038.48$, $p < 0.001$, and bimodal, $F(4, 236) = 671.23$, $p < 0.001$,

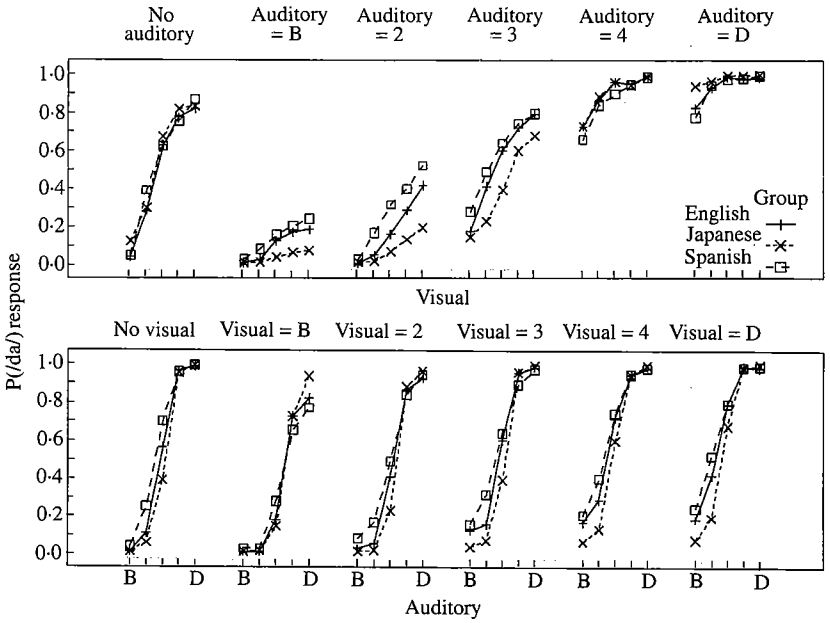

Figure 8. Probability of a /da/ response as a function of the visual and auditory levels of the speech stimulus for the visual alone (top left plot), auditory alone (bottom left plot) and bimodal (remaining plots) conditions as a function of the five levels of the synthetic auditory and visual speech varying between /ba/ (B) and /da/ (D) for the English, Japanese and Spanish native speakers.

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, $F(16, 944) = 45.37$, $p < 0.001$.

For the visual alone condition illustrated in the left-most plot of Fig. 8(a), there was no significant difference across the three languages groups, $F(2, 59) = 0.335$, $p = 0.72$. However, the bottom left plot shows a significant difference between language groups in the unimodal auditory condition, both as a simple effect, $F(2, 59) = 10.437$, $p < 0.001$, and as an interaction of auditory level and groups, $F(8, 236) = 8.216$, $p < 0.001$. Not surprisingly, the synthetic auditory speech did not match the natural language categories equivalently across the three languages. The second and third levels along the /ba/–/da/ continuum were identified as /da/ more frequently by the Spanish speakers. The Japanese speakers identified the third level as /da/ less frequently than either the Spanish and English speakers.

The five plots on the top right-hand side of Fig. 8(a) show $P(/da/)$ as a function of the five levels along the visual continuum. The five plots correspond to the five levels of the auditory continuum. Language group is the curve parameter within each plot. There was a significant difference for the auditory source of information in the bimodal condition, $F(2, 59) = 3.954$, $p < 0.05$. The differences are primarily due to the Spanish speakers identifying the three most /ba/-like auditory stimuli as more /da/-like and Japanese subjects identifying these same stimuli as more /ba/-like than the English speakers, as reflected in a significant group by auditory level interaction, $F(8, 236) = 5.120$, $p < 0.001$.

The three groups did not show any significant difference with respect to identification of the syllables along the visual continuum in either the visual [$F(8, 236) = 1.050$, $p = 0.399$]) or bimodal [$F(8, 236) = 1.818$, $p = 0.074$] conditions.

4.2.1. *Relative influence of visible and audible speech*

One question of interest is the relative contribution of visible and audible speech in the bimodal condition. An index of the magnitude of the effect of one modality can be calculated by taking the difference in response probabilities to the two endpoint stimuli 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 0.9 probability of /da/ given the visual /da/ endpoint stimulus and an overall 0.2 probability of /da/ given the visual /ba/ endpoint stimulus would give a visual effect of 0.7. Analyses of variance were carried out on these visual and auditory effect scores. The left plot of Fig. 9 shows the visual effect as a function of the auditory level for the subjects in the bimodal condition for the English, Japanese and Spanish speakers. As can be seen, the visual effect was higher for the more ambiguous central auditory levels than for the end points, $F(4, 236) = 77.476$, $p < 0.001$. The overall magnitude of the visual effect (0.322, 0.219 and 0.353 for the English, Japanese and Spanish speakers, respectively) did not differ across the three groups, $F(2, 59) = 2.412$, $p = 0.097$, although there was an interaction between group and auditory level, $F(8, 236) = 4.029$), $p < 0.001$. We believe that this interaction occurred because the subjects in the three groups perceived the auditory levels somewhat differently. The Japanese speakers, for example had a wider range of marginal $P(/da/)$ judgments. It should be stressed that we do not expect the size of the effect of a given modality to be equivalent across the three languages. The synthesized continuum between /ba/ and /da/ will not match the prototypical values

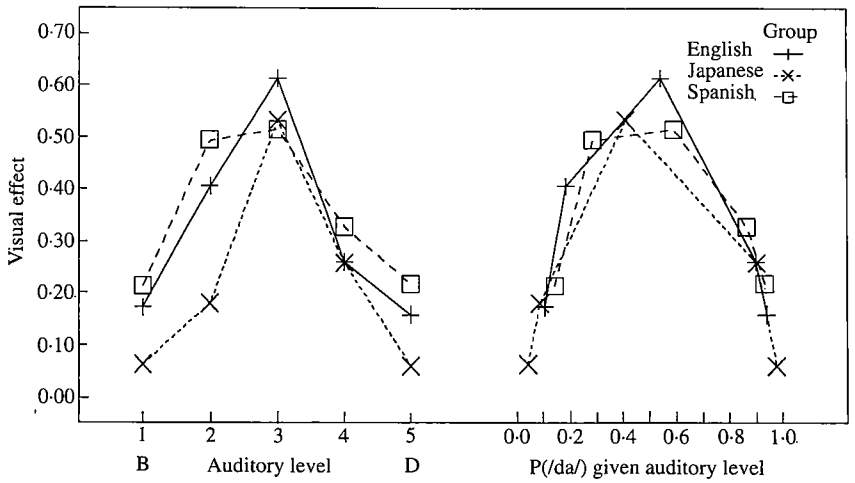

Figure 9. Visual effect as a function of the auditory level (left plot) and as a function of the probability of a /da/ response given the auditory level (right plot) over subjects in the bimodal condition for the English, Japanese and Spanish speakers.