

# **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.

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 /ɔ̃a/. 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

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

**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 auditory 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).

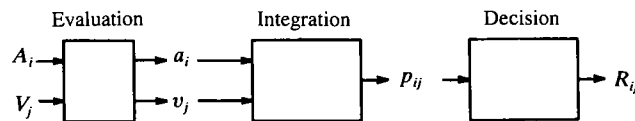
## 2. Models of bimodal speech perception

We adhere to a falsification and strong-inference strategy of inquiry (Platt, 1964; Massaro, 1987, 1989a). 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, 1989b, 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:

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

where  $S(/da/ | A_i \text{ and } V_j)$  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 modalities, 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:

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

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:

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

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

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

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

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

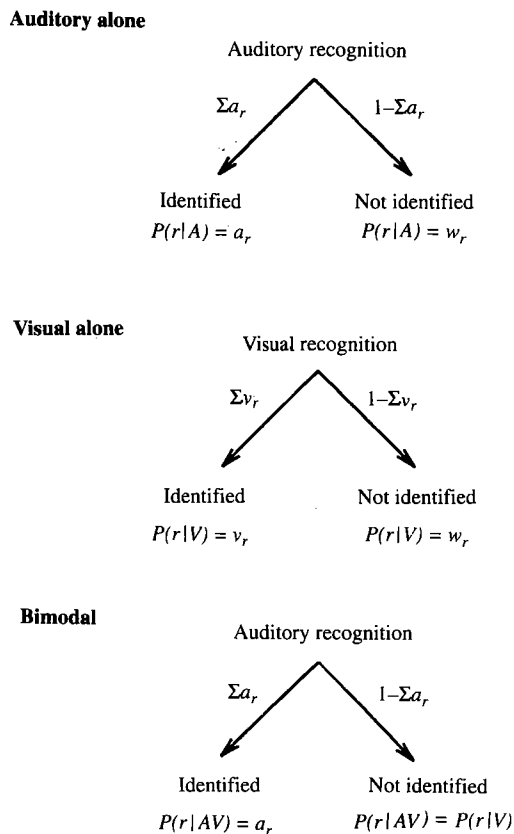
As can be seen in Equations (1) and (2), the absolute support for a given prototype will be less for two sources of information than just one. However, the identification judgement is a function of the relative degree of support as shown in Equations (3), (4) and (5). Thus, it is possible that a given 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

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

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 trials 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

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

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

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

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 visible 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:

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

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:

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

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

