# A FUZZY LOGICAL MODEL OF SPEECH PERCEPTION

DOMINIC W. MASSARO

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

## ABSTRACT

Speech perception is viewed as having available multiple sources of information supporting the identification and interpretation of the language input. The results from a wide variety of experiments can be described within a framework of a fuzzy logical model of perception. The assumptions central to the model are 1) each source of information is evaluated to give the degree to which that source specifies various 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 and interpretation follows the relative degree of support among the alternatives. A formalization of these assumptions is applied to results of an experiment manipulating audible and visible characteristics of the syllables /ba/ and /da/. In addition, the results are used to test an alternative categorical model of speech perception. The good description of the results by the fuzzy logical models indicate that the sources of support provide continuous rather than categorical information. The integration of the multiple sources results in the least ambiguous sources having the most impact on processing. These results provides major constraints to be met by theories of speech perception and language processing.

## INTRODUCTION

Speech perception is a human skill that rivals our other impressive achievements. Even after decades of intense effort, speech recognition by machine remains far inferior to human performance. The central thesis of the present proposal is that there are multiple sources of information supporting speech perception, and the perceiver evaluates and integrates all of these sources to achieve perceptual recognition. Consider recognition of the word *performance* in the spoken sentence

*The actress was praised for her outstanding performance.*

Recognition of the critical word is achieved via a variety of bottom-up and top-down sources of information. Top-down sources include semantic, syntactic, and phonological constraints and bottom-up sources include audible and visible features of the spoken word.

## A THEORETICAL FRAMEWORK FOR PATTERN RECOGNITION

According to the present framework, well-learned patterns are recognized in accordance with a general algorithm, regardless of the modality or particular nature of the patterns [1, 2, 3]. The model has received support in a wide variety of domains and consists of three operations in perceptual (primary) recognition: feature evaluation, feature integration, and pattern classification. 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. The model is called a fuzzy logical model of perception (abbreviated FLMP).

Central to the FLMP are summary descriptions of the perceptual units of the language. These summary descriptions are called prototypes and they contain a conjunction of various properties called features. A prototype is a category and the features of the prototype correspond to the ideal values that an exemplar should have if it is a member of that category. The exact form of the representation of these properties is not known and may never be known. However, the memory representation must be compatible with the sensory representation resulting from the transduction of the audible and visible speech. Compatibility is necessary because the two representations must be related to one another. To recognize the syllable /ba/, the perceiver must be able to relate the information provided by the syllable itself to some memory of the category /ba/.

Prototypes are generated for the task at hand. In speech perception, for example, we might envision activation of all prototypes corresponding to the perceptual units of the language being spoken. For ease of exposition, consider a speech signal representing a single perceptual unit, such as the syllable /ba/. The sensory systems transduce the physical event and make available various sources of information called features. During the first operation in the model, the features are evaluated in terms of the prototypes in memory. For each feature and for each prototype, featural evaluation provides information about the degree to which the feature in the speech signal matches the featural value of the prototype.

Given the necessarily large variety of features, it is necessary to have a common metric representing the degree of match of each feature. The syllable /ba/, for example, might have visible featural information related to the closing of the lips and audible information corresponding to the second and third formant transitions. These two features must share a common metric if they eventually are going to be related to one another. To serve this purpose, fuzzy truth values [4] are used because they provide a natural representation of the degree of match. Fuzzy truth values lie between zero and one, corresponding to a proposition being completely false and completely true. The value .5 corresponds to a completely ambiguous situation whereas .7 would be more true than false and so on. Fuzzy truth values, therefore, not only can represent continuous rather than just categorical information, they also can represent different kinds of information. Another advantage of fuzzy truth values is that they couch information in mathematical terms (or at least in a quantitative form). This allows the natural development of a quantitative description of the phenomenon of interest.

Feature evaluation provides the degree to which each feature in the syllable matches the corresponding feature in each prototype in memory. The goal, of course, is to determine the overall goodness of match of each prototype with the syllable. All of the features are capable of contributing to this process and the second operation of the model is called feature integration. That is, the features (actually the degrees of matches) corresponding to each prototype are combined (or conjoined in logical terms). The outcome of feature integration consists of the degree to which each prototype matches the syllable. In the model, all features contribute to the final value, but with the property that the least ambiguous features have the most impact on the outcome.

The third operation during recognition processing is pattern classification. During this stage, the merit of each relevant prototype is evaluated relative to the sum of the merits of the other relevant prototypes. This relative goodness of match gives the proportion of times the syllable is identified as an instance of the prototype. The relative goodness of match could also be determined from a rating judgment indicating the degree to which the syllable matches the category. The pattern classification operation is modeled after Luce's [5] choice rule. In pandemonium-like terms [6], we might say that it is not how loud some demon is shouting but rather the relative loudness of that demon in the crowd of relevant demons. An important prediction of the model is that one feature has its greatest effect when a second feature is at its most ambiguous level. Thus, the most informative feature has the greatest impact on the judgment.

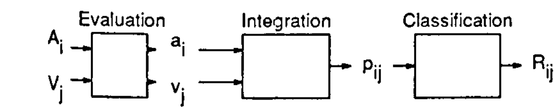


**Figure 1.** Schematic representation of the three operations involved in perceptual recognition.

Figure 1 illustrates the three stages involved in pattern recognition. Auditory and visual sources of information are represented by uppercase letters. The evaluation process transforms these into psychological values (indicated by lowercase letters) that are then integrated to give an overall value. The classification operation maps this value into some response, such as a discrete decision or a rating. The model confronts several important issues in describing speech perception. One issue has to do with whether multiple sources of information are evaluated in speech perception. Two other issues have to do with the evaluation of the sources in that we ask whether continuous information is available from each source and whether the output of evaluation of one source is contaminated by the other source. The issue of categorical versus continuous perception can also be asked with respect to the output of the integration process. Questions about integration assess whether the components passed on by evaluation are integrated into some higher-order representation and how the two sources of information are integrated.

The theoretical framework of the FLMP has proven to be a valuable framework for the study of speech perception. Experiments designed in this framework have provided important information concerning the sources of information in speech perception, and how these sources of information are processed to support speech perception. The experiments have studied a broad range of information sources, including bottom-up sources such as audible and visible characteristics of speech and top-down sources, including phonological, lexical, syntactic, and semantic constraints. As examples, experiments have assessed the contributions of formant structure and duration of vowels in vowel identification [7], the role of vowel duration and consonant duration in the identification of post-vocalic stop consonants [8, 9] and fricatives [10], the integration of voice onset time and formant structure of segment-initial stop consonants [11, 12] and fricatives [13]. These results are not limited to western languages; experiments have shown that both pitch height and pitch contour contribute to the perception of Mandarin Chinese lexical tone [14]. Experiments have also revealed the integration of nonauditory sources of information, such as pointing gestures, with auditory sources [15]. Several experiments have also addressed the relative contributions of acoustic information and higher-order constraints in the pattern. These experiments have included formant structure and phonological constraints in the identification of glides [16], the formant structure and lexical constraints in the identification of stop consonants [17], segmental information and syntactic constraints in the identification of words [18], semantic constraints in word identification [19], and word order, animacy, and noun-verb agreement in sentence interpretation [20].

## EXPANDED FACTORIAL DESIGN

An expanded factorial design with open-ended response alternatives offers the potential of addressing important issues in speech perception. I will describe an experiment manipulating auditory and visual information in a speech perception task. The novel design illustrated in Figure 2, along with open-ended response alternatives, has not been used previously in speech perception research and it provides a unique method to address the issues of evaluation and integration of audible and visible information in speech perception.

Eight college students from the University of California, Santa Cruz, participated for one hour in the experiment. All test stimuli were recorded on videotape. On each trial the speaker said either /ba/ or /da/ or nothing, as cued by a video terminal under computer control. When the speaker was cued to say nothing, a

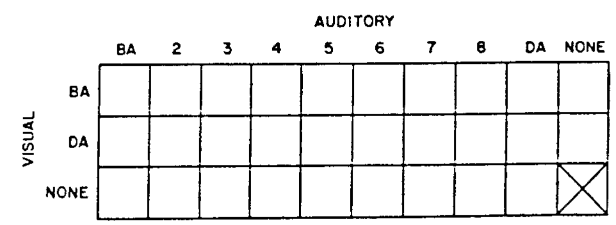


**Figure 2.** Expansion of a typical factorial design to include auditory and visual conditions presented alone. The nine levels along the auditory continuum represent speech sounds varying in equal steps between /ba/ and /da/.

computer-controlled tone was recorded on the audio channel of the videotape 400 msec after the onset of the neutral cue. The original audio track of the videotape was replaced with synthetic speech. A nine-step /ba/ to /da/ auditory continuum was used to replace the original audio. By altering the parametric information specifying the first 80 msec of the consonant-vowel syllable, a set of nine 400 msec syllables covering the range from /ba/ to /da/ was created. The experimental videotapes were made by copying the original tape and replacing the original sound track with the synthetic speech. The presentation of the synthetic speech was synchronized with the original audio track on the videotape.

The 29 speech events illustrated in Figure 2 were presented to each subject in a randomized order. Each subject made about 600 identifications, which were converted to probabilities of responding with each of the eight alternatives. Figure 3 presents the observed probability of each of the eight responses for the 29 unique speech events.

## FUZZY LOGICAL MODEL OF PERCEPTION (FLMP)

Applying the model to the present task using auditory and visual speech, both sources are assumed to provide continuous and independent evidence for the alternatives /ba/ and /da/. Defining the onsets of the second (F2) and third (F3) formants as the important auditory feature and the degree of initial opening of the lips as the important visual feature, the prototype for /da/ would be:

/da/ : Slightly falling F2-F3 & Open lips.

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

/ba/ : Rising F2-F3 & Closed lips,

and so on for the other response alternatives. 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 at the prototype matching stage. The integration of the features defining each prototype is evaluated according to the product of the feature values. If $aD_i$ represents the degree to which the auditory stimulus $A_i$ supports the alternative /da/, that is, has Slightly falling F2-F3; and $vD_j$ represents the degree to which the visual stimulus $V_j$ supports the alternative /ba/, that is, has Open lips, then the outcome of prototype matching for /da/ would be:

/da/ : $aD_i\ vD_j$

where the subscripts i and j index the levels of the auditory and visual modalities, respectively. Analogously, if $aB_i$ represents the degree to which the auditory stimulus $A_i$ has Rising F2-F3 and $vB_j$ represents the degree to which the visual stimulus $V_j$ has Closed lips, the outcome of prototype matching for /ba/ would be:

/ba/ : $aB_i\ vBj$

and so on for the other alternatives.

The pattern classification operation would determine their relative merit leading to the prediction that

$$P(/da/|A_iV_j) = \frac{aD_i\ vD_j}{\sum} \qquad (1)$$

where $\sum$ is equal to the sum of the merit of all eight alternatives, derived in the manner illustrated for /da/ and /ba/.

The 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. Each level of a source supports each alternative to differing degrees represented by feature values. The feature values representing the degree of support from the auditory and visual information for a given alternative are integrated following the multiplicative rule given by the FLMP. The model requires 2 parameters for the visual feature values and 9 parameters for the auditory feature values, for each of the 8 response alternatives, for a total of 88 parameters.

## CATEGORICAL MODEL OF PERCEPTION (CMP)

It is essential to contrast one model with other models that make alternative assumptions. One alternative is a categorical model of perception (CMP). It assumes that only categorical

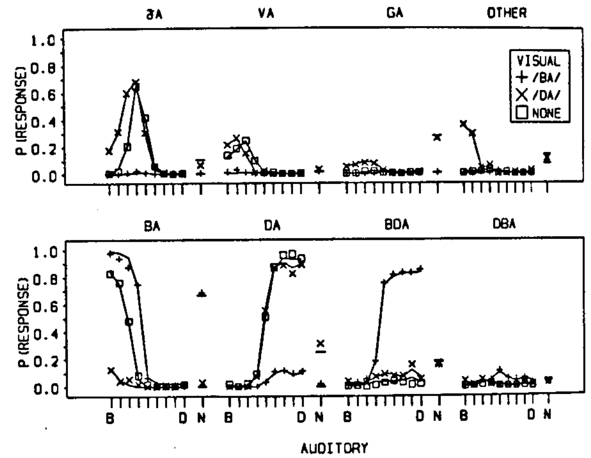


**Figure 3.** Probability of responding with each of the eight alternatives as a function of the auditory and visual sources under the bimodal and unimodal conditions. The nine levels between B and D along the auditory continuum represent speech sounds varying in equal steps between /ba/ and /da/. The level N refers to no auditory information. The curve parameter corresponds to a visual /ba/, a visual /da/, and no visual information. The lines give the predictions for the FLMP.

information is available from the auditory and visual sources and that the identification judgment is based on separate decisions to the auditory and visual sources. Given eight response alternatives, there are eight possible outcomes for a particular combination of auditory and visual information. Considering the /ba/ identification, the visual and auditory decisions could be /ba/-/ba/, /ba/-not /ba/, not /ba/-/ba/, or not /ba/-not /ba/. If the two decisions to a given speech event agree, the identification response can follow either source. When the two decisions disagree, it is assumed that the subject will respond with the decision to the auditory source on some proportion $p$ of the trials, and with the decision to the visual source on the remainder $(1-p)$ of the trials. The weight $p$ reflects the relative dominance of the auditory source.

The probability of a /ba/ identification response, P(/ba/), given a particular auditory/visual speech event, $A_iV_j$, would be:

$$P(/ba/|A_iV_j) = (1)\,aB_i\,vB_j + (p)\,aB_i\,(1-vB_j) + (1-p)(1-aB_i)vB_j + (0)(1-aB_i)(1-vB_j) \qquad (2)$$

where $i$ and $j$ index the levels of the auditory and visual modalities, respectively. The $aB_i$ value represents the probability of a /ba/ decision given the auditory level $i$, and $vB_j$ is the probability of a /ba/ decision given the visual level $j$. The value $p$ reflects the bias to follow the auditory source. Each of the four terms in the equation represents the likelihood of one of the four possible outcomes multiplied by the probability of a /ba/ identification response given that outcome. To fit this model to the results, each unique level of the auditory stimulus requires a unique parameter $aB_i$, and analogously for $vB_j$. The modeling of /ba/ responses thus requires 9 auditory parameters plus 2 visual parameters. Each of the other seven response alternatives needs an analogous equation to Equation 2 and an additional 11 parameters, thus requiring a total of 88 visual and auditory parameters. For any particular auditory-visual combination, the sum of the eight decision probabilities to a given source also has to be constrained to be less than or equal to one; the assumption is that a given





source is categorized as only a single category on any given presentation. An additional $p$ value would be fixed across all conditions for a total of 89 parameters. Thus, we have a fair comparison to the FLMP which requires 88 parameters.

## MODEL TESTS

Figures 3 and 4 give the average observed results and the average predicted results of the FLMP and CMP. As can be seen in the Figure 4, the CMP gave a poor description of the observed results. The predictions of the FLMP shown in Figure 3, on the other hand, provide a very good description. The FLMP gave a mean root mean square deviation (RMSD) of .030 averaged across the individual subject fits of the 8 subjects compared to an average RMSD of .148 for the CMP.

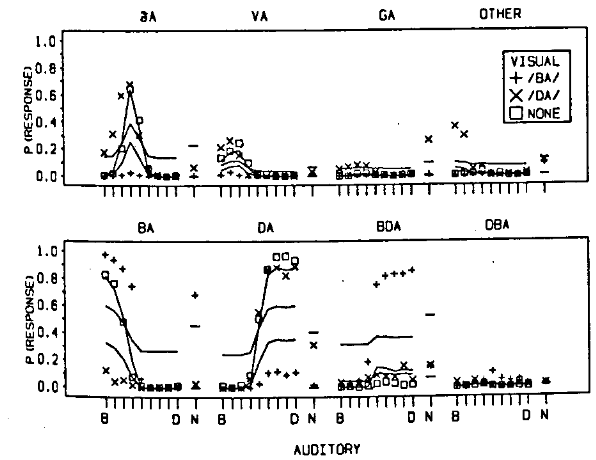


**Figure 4.** Probability of responding with each of the eight alternatives as a function of the auditory and visual sources under the bimodal and unimodal conditions. The nine levels between B and D along the auditory continuum represent speech sounds varying in equal steps between /ba/ and /da/. The level N refers to no auditory information. The curve parameter corresponds to visual /ba/, a visual /da/, and no visual information. The lines give the predictions for the CMP.

## CONCLUSION

The present framework provides a valuable approach to the study of speech perception. We have learned about some of the fundamental stages of processing involved in speech perception by ear and eye, and how multiple sources of information are used in speech perception. Given the potential for evaluating and integrating multiple sources of information in speech perception and understanding, no single source should be considered necessary. There is now good evidence that perceivers have continuous information about the various sources of information, each source is evaluated, and all sources are integrated in speech perception. Future work should address the nature of the variety of sources of information, and how they function in recovering the speaker's message.

## REFERENCES


[1] Massaro, D. W. (1979). Reading and listening (Tutorial paper). In P. A. Kolers, M. Wrolstad, & H. Bouma (Eds.), *Processing of visible language: Vol. 1* (pp. 331-354). New York: Plenum.

[2] Massaro, D. W., & Oden, G. C. (1980b). Speech perception: A framework for research and theory. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice: Vol. 3* (pp. 129-165). New York: Academic Press.

[3] Massaro, D. W. (in press). *Speech perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale, NJ: Lawrence Erlbaum Associates.

[4] Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8,* 338-353.

[5] Luce, R. D. (1959). *Individual choice behavior.* New York: Wiley.

[6] Selfridge, O. G. (1959). Pandemonium: A paradigm for learning. In *Mechanization of thought processes* (pp. 511-526). London: Her Majesty's Stationery Office.

[7] Massaro, D. W. (1984). Time's role for information, processing, and normalization. *Annals of the New York Academy of Sciences, Timing and Time Perception, 423,* 372-384.

[8] Denes, P. (1955). Effects of duration on the perception of voicing. *Journal of the Acoustical Society of America, 27,* 761-764.

[9] Massaro, D. W., & Cohen, M. M. (1983). Consonant/vowel ratio: An improbable cue in speech. *Perception & Psychophysics, 33,* 502-505.

[10] Massaro, D. W., & Cohen, M. M. (1977). The contribution of voice-onset time and fundamental frequency as cues to the /zi/-/si/ distinction. *Perception & Psychophysics, 22,* 373-382.

[11] Massaro, D. W., & Oden, G. C. (1980). Evaluation and integration of acoustic features in speech perception. *Journal of the Acoustical Society of America, 67,* 996-1013.

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

[13] Massaro, D. W., & Cohen, M. M. (1976). The contribution of fundamental frequency and voice onset time to the /zi/-/si/ distinction, *Journal of the Acoustical Society of America, 60,* 704-717.

[14] Massaro, D. W., Cohen, M. M., & Tseng, C-Y. (1985). The evaluation and integration of pitch height and pitch contour in lexical tone perception in Mandarin Chinese. *Journal of Chinese Linguistics, 13,* 267-289.

[15] Thompson, L. A., & Massaro, D. W. (1986). Evaluation and integration of speech and pointing gestures during referential understanding. *Journal of Experimental Child Psychology, 42,* 144-168.

[16] Massaro, D. W., & Cohen, M. M. (1983). Phonological context in speech perception. *Perception & Psychophysics, 34,* 338-348.

[17] Ganong, W. F. III. (1980) Phonetic categorization in auditory word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 6,* 110-125.

[18] Tyler, L. K., & Wessels, J. (1983). Quantifying contextual contributions to word-recognition processes. *Perception & Psychophysics, 34,* 409-420.

[19] Oden, G. C. (1978). Semantic constraints and judged preference for interpretations of ambiguous sentences. *Memory & Cognition, 6,* 26-37.