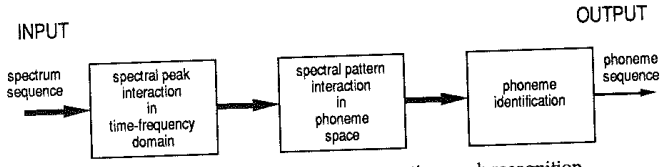

Fig. 9. Contextual effect model for automatic speech recognition.

(7) The difference can be represented as a function of the distance between the preceding vowel anchor and the perceived vowel in the phoneme space.

These results show that it is possible to formulate a contextual effect model to apply to automatic speech recognition. This model is shown in Fig. 9. First, the input spectrum sequence goes through a model that predicts target spectral peaks in reduced vowels, based on the interaction between spectral peak pairs in the time-frequency domain. Next, phoneme boundaries are shifted to pull reduced spectral patterns back into their correct categories, based on the influence of preceding spectral patterns in the phoneme space.

Acknowledgment

A portion of this study was carried out during the author's stay at MIT. He wishes to thank Dr. Victor Zue at LCS, MIT for his helpful suggestions and guidance.

References

Akagi, M. (1990). Evaluation of a spectrum target prediction model in speech perception. *J. Acoust. Soc. Am.*, **87**, 2, 858-865.

Akagi, M. and Tohkura, Y. (1990). Spectrum target prediction model and its application to speech recognition. *Computer Speech and Language (1990)* **4**, 325-344.

Hirahara, T. (1988). On the role of the fundamental frequency in vowel perception. *J. Acoust. Soc. Am., Suppl. 1*, **84**, WW11.

Kuwabara, H. (1985). An approach to normalization of coarticulation effects for vowels in connected speech. *J. Acoust. Soc. Am.*, 77, 2, 686-694.

Lindblom, B. E. F. and Studdert-Kennedy, M. (1967). On the role of formant transition in vowel recognition. *J. Acoust. Soc. Am.*, **42**, *4*, 686-694.

Zwicker, E. and Terhardt, E. (1980). Analytic expressions for critical-band rate and critical bandwidth as a function of frequency. *J. Acoust. Soc. Am.*, **68**, 5, 1523-1525.

The Fuzzy Logical Model of Speech Perception: A Framework for Research and Theory

Dominic W. Massaro

Program in Experimental Psychology
University of California, Santa Cruz
Santa Cruz, California, USA

The papers by Huang, Shigeno, Akagi, and Fox are representative of the many systematic, thorough, and intricate research programs in speech perception. A nonexpert in any of these four areas must necessarily have the impression of a complex and impenetrable world of esoteric science, with little potential of some unifying understanding of these studies. From a more practical and threatening perspective, one might hazard the question of whether this research is worth the time, effort, and financial resources that are required. My charge is not to challenge the authors, but perhaps to remind them that they should also keep the big picture in mind in their day-to-day struggle with the wonders of speech perception.

The goal of this commentary is to describe the Fuzzy Logical Model of Perception (FLMP) and illustrate how the research findings from the four disparate lines of research are consistent with the basic assumptions of the model. The model is formulated within the context of the information-processing approach to the study of human performance. Psychological models must account for the information available to the participants in perception and cognition, as well as the information processing the information undergoes. Perception and cognition are viewed as a sequence of information-processing operations, beginning with the sensory transduction of the environment and ending with some knowledge representation in the mind of the perceiver. Three stages of information processing are assumed by the model: evaluation, integration, and decision. Building on the idea of fuzzy sets and fundamental properties of human performance, the FLMP

has been developed and applied to several domains of perception, pattern recognition, and memory. Various characteristics of the model, such as its optimality properties and its relation to probabilistic models and neural network models, are discussed in Massaro (1987, 1989).

The assumptions central to the model are 1) there are multiple sources of information supporting the identification and interpretation of the language input, 2) each source of information is evaluated to give the degree to which that source specifies various alternatives, 3) the sources of information are evaluated independently of one another, 4) the sources are integrated to provide an overall degree of support for each alternative, and 5) perceptual identification and interpretation follows the relative degree of support among the alternatives. Thus, well-learned patterns are recognized in accordance with a general algorithm, regardless of the modality or particular nature of the patterns (Massaro, 1987). 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.

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, there is activation of all prototypes corresponding to the perceptual units of the language being spoken. 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 (Zadeh, 1965) 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 decision. The decision operation is a relative decision rule. During this stage, the merit of each relevant prototype is evaluated relative to the sum of the merits of all 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. 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.

Within the context of the FLMP, Huang's finding of the context dependency of vowel identification is of interest. The FLMP has been grounded in V, CV, and VC syllables as the fundamental perceptual units of speech perception. Hence, the model correctly predicts the importance of stop versus liquid-glide contexts in the performance of the Gaussian classifier. Her results also provide evidence for the importance of multiple sources of information in speech. The classifier performed better given an averaged F_1 value and an extrapolated F_2 value than given either of these presented alone or than given just the midpoint value. In like manner, Akagi's research gives evidence for the importance of the syllable as the perceptual

unit. Co-articulation has a large impact on the characteristics of the consonant and vowel segments and speech recognition is best carried out with respect to the syllabic segment. Akagi's findings also show two influences from the contextual stimulus rather than just one, showing that a single stimulus can have multiple influences.

Shigeno's research also illustrates the influence of multiple sources of information on speech perception. Her studies substantiate the importance of fundamental auditory processes—assimilation and contrast—in speech perception. Her studies uncover three factors influencing these contributions: the temporal position of stimulus relative to other stimuli, the acoustic differences among the stimuli, and the speech categories of the stimuli.

In contrast to the other three papers showing multiple bottom-up influences, Fox's research documents a top-down influence of vowel perception. Consistent with previous research, vowel identification is influenced by the perceiver's knowledge of phonological constraints. Finally, Fox's research uncovers another instance of an important property of perception in general and speech perception particularly. The perceiver's intentions are relatively impotent when faced with several sources of stimulus information. His subjects were unable to ignore the influence of the postvocalic consonant on the identification of the preceding vowel—even given detailed instructions to do so and a small number of test stimuli.

As a final comment, I would like to question how much we can expect a theory of speech perception to predict or explain. It is reasonable that complexity of the natural setting far exceeds what any scientific theory could hope to accommodate. We expect theories to have predictive power only in the laboratory in which the complexity of everyday life can be simplified, measured, and controlled. However, the complexity of the prototypical speech perception experiment might exceed any theory's predictive power. Typical laboratory data might exceed the constraints demanded by theory (even reasonably correct theory). My contention is that our experiments too infrequently have the constraints to inform theoretical alternatives. To overcome these shortcomings, we have delineated a paradigm for speech perception research (Massaro, 1987).

References

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

Massaro, D. W. (1989). Multiple book review of speech perception by ear and eye: A paradigm for psychological inquiry. *Behavioral and Brain Sciences*, **12**, 741-794.

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

Comment

Sieb G. Nooteboom

Research Institute for Language and Speech (OTS)
Utrecht University , Utrecht, THE NETHERLANDS

I will first propose a qualitative interpretation of the main findings of Drs. Shigeno and Akagi, and then see how some of the results reported by Fox and Huang fit in.

Drs. Shigeno and Akagi both describe ingenious, rather complex experiments with rather complex results. I feel an urgent need here for reducing this complexity in terms of more simple underlying causes. I think that, at least qualitatively, such reduction is possible in the following way:

Results of both studies can be understood from an interaction between two perceptual effects. One is the well-known effect that in the perception of (speech) sounds two sounds that belong to the same (phonemic) category are perceived as more similar than one would expect on the basis of their physical differences, and two sounds that belong to different (phonemic) categories are perceived as more dissimilar than one would expect on the basis of their physical differences. The second effect is the effect of temporal proximity. Two sounds, at least if they have about the same pitch, are perceived as belonging together, as belonging to the same stream of speech, when they follow each other closely in time. This togetherness of necessity weakens when the temporal separation becomes stronger.

The interaction in the data is clear: When two similar sounds (same phonemic category) follow each other more and more closely in time, they will be perceived as belonging more and more to the same stream of speech, and as a result tend to be more and more perceived as together preserving categorial continuity over time. Because of the increasing categorial sameness, the two sounds will be perceived as becoming more and more similar.

When two dissimilar sounds (different phonemic category) follow each other more and more closely in time, they also will be perceived as belonging more and more to the same stream of speech, but as a result their