CONNECTIONIST MODELS OF SPEECH PERCEPTION

Dominic W. Massaro

University of California, Santa Cruz, CA 95064 U.S.A.

ABSTRACT

Interactive-activation and feed-forward connectionist models are evaluated, tested, and compared to a process model, the Fuzzy Logical Model of Perception (FLMP). Empirical results indicate that while several sources of information simultaneously influence speech perception, the representation of each source remains independent of other sources. This independence is strong evidence against interactive activation in speech perception. Although some feed-forward models with input and output layers bear some similarity to the FLMP, there is evidence against the additive integration that is assumed by feed-forward models.

1. INTRODUCTION

At the Eleventh Congress of Phonetic Sciences, I described the Fuzzy Logical Model of Perception (FLMP), an information processing model of speech perception [2]. The FLMP has been shown to provide a good description of speech perception in a variety of different experiments. The model accounts for the evaluation and integration of multiple sources of information in speech perception. These sources of information include acoustic, visible, and electrotactile sources of bottom-up stimulus input, as well as top-down sources of phonological, syntactic, and semantic context. In the present paper, several classes of connectionist models are compared to the FLMP. The FLMP is used as a standard for judgment because it has been shown to provide a good description of speech perception in a variety of different experiments.

Evaluating different classes of models and testing among them is a highly involved and complex endeavor [6]. Each class has models that give a reasonable description of the results of interest. Distinguishing among models, therefore, requires a fine-grained analysis of the predictions and observations to determine quantitative differences in the accuracy of the models. Preference for one class of models is also influenced by factors other than just goodness of fit between experiment and theory. Some models are too powerful and thus not falsifiable. With enough hidden units, for example, connectionist models can predict too many different results [3]. Models should also help us understand the phenomena of interest. For example, parameters of a model might provide illuminating dependent measures of the information available in speech perception and the processing of that information. Finally, one should take into account the parsimony of a model. Certainly, a model should contain fewer parameters than the number of data points that it predicts. Models which can provide a good fit to the data with relatively few parameters should be preferred.

2. INTERACTIVE ACTIVATION

In interactive activation models, layers of units are connected in hierarchical fashion with two-way connections among units both within a layer and between layers. For example, the TRACE model of speech perception has feature, phoneme, and word layers. There are excitatory two-way connections between pairs of units from different layers and inhibitory two-way connections between pairs of units within the same layer. Thus, interactive activation is based on the assumption that the activation of a higher layer eventually modifies the activation and information representation at a lower layer [7].

How might interactive activation and the TRACE model be formulated to predict the results of bimodal speech perception? Given audible and visible speech, for example, separate sets of feature units would be associated with the two different information sources. Figure 1 gives a schematic representation of the auditory feature, visual feature, and phoneme layers and the connections between units within and between these layers. The two layers of feature units would both be connected to the phoneme layer. Following the logic of interactive activation, there would be two-way excitatory connections between the feature and phoneme layers (as in the TRACE model). Presentation of auditory speech would activate some units within the auditory feature layer. These activated units in turn would activate certain phoneme units, which would in turn activate units at both feature layers, and so on during the period of interactive activation. Activated units would also inhibit other units within the same layer.

If auditory and visual units interact, as assumed by interactive activation, then presentation of a syllable in one modality should influence processing of

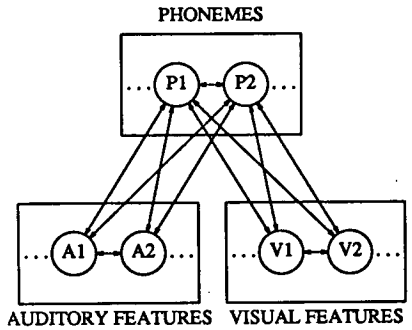

Figure 1. Illustration of the TRACE model applied to bimodal speech perception. Two input layers contain auditory and visual feature units, respectively. The third layer contains phoneme units. There are positive connections between two units from different layers and negative connection between two units within the same layer.

the syllable in the other modality. If interactive activation does not occur, on the other hand, the contribution of visible speech should be independent of the contribution of audible speech. Independence means that the representation of the visible speech should not be modified by the representation of the audible speech. The results from several different experiments in several different tasks indicate that interactive activation does not occur in bimodal speech perception [2, 8]. More generally, there is now a substantial body of evidence against interactive activation in speech perception [4, 5].

3. FEED-FORWARD MODELS

In contrast to interactive activation, feed-forward models assume that activation feeds only forward. Two-layer models have an input layer connected to an output layer, as illustrated in Figure 2.

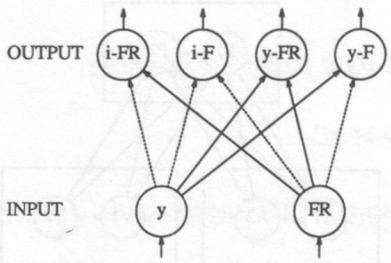

Figure 2. Illustration of a connectionist model (CMP) of speech perception of four words in Mandarin Chinese. The formant structure y and (F_0) contour FR input units are connected to the four response units corresponding to the four word alternatives. Solid arrows indicate connections with weight 1, and dashed arrows indicate connections with weight -1.

The feed-forward model illustrated in Figure 2 is tested against the results of an experimental study of the identification of Mandarin Chinese words [6]. There were four possible responses in the experiment. The experimental task was a graded factorial design with seven levels of each of two factors. The factors were the formant structure of the vowel in the monosyllabic words and the fundamental frequency (F_0) contour (tone) during the vowel. Mandarin Chinese is a tone language and both of these sources of information function to distinguish among different words. The formant structure was varied to make a continuum of vowel sounds between /i/ and /y/. (The phoneme /y/ is articulated in the same manner as /i/, except with the lips rounded.) The F_0 contour varied between falling-rising to falling during the vowel. Six native Chinese speakers participated for four days, giving a total number of 48 responses to each of the 49 test stimuli. The subjects identified each of the 49 test stimuli as one of the four words.

Figure 3 gives the observed results and the predictions of the FLMP and the connectionist model (CMP). As can be seen in the figure, the CMP fails catastrophically primarily because it cannot predict a probability of a response greater than .5. The FLMP, on the other hand, captures the results reasonably well. The success of the FLMP is due to the multiplicative integration of the two sources of information. A perfect match of a stimulus with a given response alternative on just one source does not necessarily mean that this alternative should qualify as a reasonably good alternative. The linear integration in the CMP, however, guarantees that a perfect match of a response alternative with just a single source of information will be significantly activated even if the other source of information mismatches the response alternative completely.

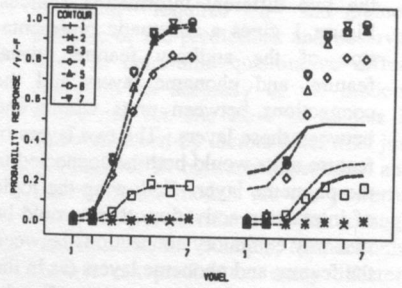

Figure 3. Observed (points) and predicted (lines) probability of /y/-falling responses for the Chinese word identification study [6]. The left panel gives the predictions for the FLMP and the right panel gives the predictions for the CMP.

Admittedly, we have falsified a very restricted implementation of the class of feed-forward connectionist models. However, we are only willing to test models that are falsifiable. Three-layer models, for example, assume that the input units are connected to a layer of "hidden" units that are connected to an output layer of units. In a theoretical and analytical report, I have shown that models with hidden units are superpowerful—that is, they can predict many types of results and even results that do not occur [3]. Because these models can predict many results—not just those that are empirically observed, this superpower might be better described as flabbyness. Therefore, one cannot reasonably propose feed-forward models with hidden units as testable models of speech perception. These models are not reasonable because they are not falsifiable. In one case, for example, the model is essentially assuming more than it is predicting [1], and the good performance by the model in this situation should not be surprising.

In summary, there is evidence against interactive activation models, while feed-forward models with hidden units are not falsifiable. Feed-forward models with input and output units can be shown to be mathematically equivalent to the FLMP in situations with just two responses [6]. With a larger number of responses, the FLMP provides a more adequate description of the results than does this feed-forward model.

4. REFERENCES

[1] Landauer, T. K.; Kamm, C. A.; & Singhal, S. (1987). Network to recognize speech sounds. *Proceedings of the Cognitive Science Society*, 531-536. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

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

[5] Massaro, D. W., & Cohen, M. M. (in press). Integration versus interactive activation: The joint influence of stimulus and context in perception. *Cognitive Psychology*.

[6] Massaro, D. W., & Friedman, D. (1990). Models of integration given multiple sources of information. *Psychological Review, 97*, 225-252.

[7] McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*, 1-86.

[8] Roberts, M., & Summerfield, Q. U. (1981). Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory. *Perception & Psychophysics, 30*, 309-314.

V. ACKNOWLEDGMENT

The research reported in this paper and the writing of the paper were supported, in part, by grants from the Public Health Service (PHS R01 NS 20314), the National Science Foundation (BNS 8812728), a James McKeen Cattell Fellowship, and the graduate division of the University of California, Santa Cruz. The author would like to thank Michael M. Cohen for ecletic assistance.