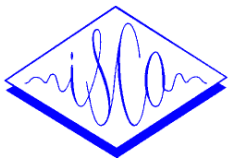

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

THE JOINT INFLUENCE OF STIMULUS INFORMATION AND CONTEXT IN SPEECH PERCEPTION

Dominic W. Massaro and Michael M. Cohen

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

ABSTRACT

Empirical results in speech perception indicate that the joint influence of stimulus and contextual information is consistent with the independence of these sources of information and inconsistent with interactive processing. Thus, the results appear to be inconsistent with an interactive activation and competition (IAC) model [1], and consistent with the fuzzy logical model of perception (FLMP) [2][3]. In order to overcome its empirical shortcomings, McClelland (in press) modified the interactive activation to be stochastic rather than deterministic and to use a best one wins (BOW) decision rule. When tested against real data and contrasted with the FLMP, however, the new stochastic IAC (SIAC) model gives a poorer description of the joint influence of stimulus information and context in perception. Interactive activation is both inconsistent with empirical results and not necessary to describe the joint influence of stimulus information and context in language perception.

I. INTRODUCTION

Psychologists have long been intrigued with the finding that context appears to influence perception. The same stimulus information in different contexts can produce different perceptual events. At the turn of the century, Bagley [4] showed that a sentence context facilitated recognition of a spoken word. A recent example of a context effect in psycholinguistic research is the influence of phonological constraints in speech perception [3]. Subjects were asked to identify a glide consonant in different phonological contexts. Each speech sound was a consonant cluster syllable beginning with one of the three consonants /p/, /t/, or /s/ followed by a glide consonant ranging (in five levels) from /l/ to /r/, followed by the vowel /i/. There were 15 test stimuli created from the factorial combination of five levels of the glide consonant combined with three initial-consonant contexts. Subjects, instructed to listen to each test syllable and to respond whether they heard /l/ or /r/, were influenced by both the glide consonant and the context.

Two models of these phonological context effects are the fuzzy logical model of perception (FLMP) of Massaro [3] and the TRACE model [1]. Both models provide a detailed description of the integration of top-down and bottom-up sources of information in speech perception. These two models share a variety of processing assumptions and make highly similar predictions. Both are information processing models and assume some perceptual processing followed by decision. Continuous rather than categorical information is available during processing. Both the original IAC models and the FLMP assumed decision rule based on the relative goodness of match. These similarities and others [2][5][6][7] lead to very similar predictions in most situations. Thus, differentiating between the models requires a fine-grained analysis of experiments specifically aimed at testing between the models.

Using a signal detection framework, Massaro [3] demonstrated that the TRACE model predicts sensitivity differences in the phonological constraints experiment—rather than just bias differences. That is, context influences the discriminability of the stimulus information specifying or representing the glide consonant. The discrimination of two adjacent levels along the /li/-/ri/ continuum differs for different contexts. In Massaro's experiment, the effect of phonological context turned out to be only a biasing effect rather than an effect on sensitivity, thus contradicting the predictions of the TRACE model. On the other hand, the results were well-described by the FLMP—whose most distinguishing feature (relative to TRACE) is independence of stimulus information and context at the evaluation stage of processing. When analyzed in the signal detection framework, the FLMP correctly predicts that context in the phonological constraints experiment should influence only bias and not sensitivity.

II. REVISED IAC MODELS

McClelland [7] argued that the decision stage of the TRACE model was responsible for its failure to predict Massaro's [3] results rather than interactive activation during the evaluation stage. By adding noise to the input or to its processing, and by assuming a decision rule of

choosing the response alternative corresponding to the most active phoneme unit, the predictions of a new stochastic IAC (SIAC) model and TRACE were brought into line with a biasing effect of context. Thus, the new TRACE appeared to be consistent with the observations (and the predictions of the FLMP).

Massaro and Cohen [8] tested the SIAC model against several different data sets. Its asymptotic behavior was compared with that of the FLMP in predicting the results of a phonological experiment by Massaro and Cohen [6]. The two models were tested against results showing top-down effects of phonological constraints. Massaro and Cohen [6] tested a larger number of experimental conditions and recorded more observations per condition than Massaro [3]. Subjects were presented with CCV syllables with the first consonant being /p/, /t/, /s/, or /v/, the second consonant being one of seven glides equally spaced on a continuum between /l/ and /r/, and with the vowel /i/. The glide was changed from /l/ to /r/ by changing its initial F_3 frequency from high to low. Seven subjects from an introductory psychology class were each presented with each of the 28 possible experimental conditions (4 context times 7 glide) 40 times in 4 sessions run over a two day period. Subjects made their responses by pressing one of eight buttons combining context and glide identifications, but we will concern ourselves mainly with the data pertaining to glide identification, except to note that subjects were 95% correct in context identification. Readers are referred to the original paper for further details of the stimuli and procedures. Figure 1 shows the proportion of observed /r/ identifications for the 28 experimental conditions averaged over the seven subjects. The effects of context and glide level were highly significant, with each independent variable having its largest effect when the other was most ambiguous.

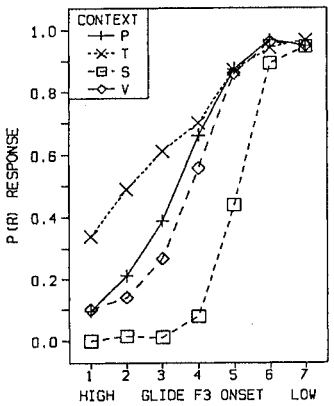

Figure 1. Observed probability of an *r* response as a function of the glide F_3 onset level and context (after Massaro & Cohen, Experiment 2, 1983).

2.1 SIAC Model with Input Noise

The network we used, shown in Figure 2, assumes three layers of units: Target, Context, and word.

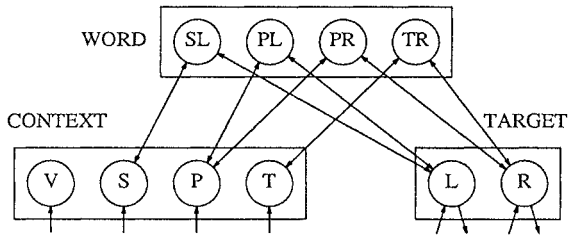

Figure 2. Network used in the simulation of the SIAC model applied to the phonological constraints experiment of Massaro & Cohen (1983). The inhibitory connections between units within the word, context, and target levels are not shown in the network.

The units within the Context layer (except V) are bidirectionally connected to units within the Word layer. Similarly, the units within the Target layer are bidirectionally connected to units within the Word layer. Within each layer, each unit sends inhibitory connections to all other units. Given a stimulus presentation, external inputs are applied to the Context and Target units. These units pass on activation to units in the Word layer, which in turn pass on activation back to the Context and Target units. Processing continues in this manner for a number of cycles. In the SIAC model, it is assumed that noise is added to the Target unit inputs.

The formal algorithm of the SIAC model is given in Massaro and Cohen [8]. In the network, the effects of stimulus and context are combined via the units in the word layer. The activations of Word units are fed back to the Target and Context units, changing their activations in a manner that reflects the activations of both Target and Context units. In this manner, the joint effect of Target and Context are represented in the activations of units in both the Target and Context layers. The activations of the /r/ and /l/ input units after 60 cycles of processing were used as inputs to the BOW decision rule.

The SIAC model was fit to the observed data using the program STEPIT [9]. Several hundred adjustments of the set of parameter values were needed to maximize the goodness-of-fit of the IAC model. The root mean squared deviation (RMSD) averaged .111 across the fits of the 7 subjects.

2.2 SIAC Model with Intrinsic Noise

In the SIAC model, it is assumed that variability is added to the inputs. Processing itself is deterministic. A second type of SIAC model proposed by McClelland assumes variability added at each processing cycle. This model is called the stochastic IAC - Intrinsic Noise

(SIAC-INT) model. Given the possibility that this type of model would give a better description of actual results, this model was tested against the results in the same manner. This model required 16 free parameters relative to the 11 free parameters of the SIAC model and gave an average RMSD equal to .079.

III. FUZZY LOGICAL MODEL

A critical assumption of the application of the FLMP in the phonological constraints study is that the featural information from the glide and the phonological context provide *independent* sources of information. The model consists of three operations in perceptual (primary) recognition: feature evaluation, feature integration, and decision (see Figure 3). 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.

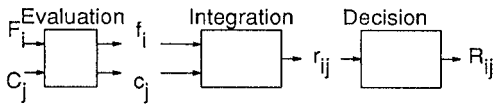

Figure 3. Schematic representation of the three operations involved in perceptual recognition.

Central to the FLMP are summary descriptions of the perceptual units of the language. These summary descriptions are called prototypes and 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. Prototypes can be 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. 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 invented by Zadeh [10] 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. The second operation of the model is called feature integration. Here, the features (actually the degrees of matches) corresponding to each prototype are combined (or conjoined in logical terms). Feature integration provides the degree to which each prototype matches the syllable. Thus all features contribute to the final degree of match.

The third operation during recognition processing is decision. 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. Given the integration and decision operations, 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.

Consider our identification task in which a set of seven syllables along a /li/-/ri/ continuum were factorially combined with four different initial consonant contexts /p/, /t/, /s/, or /v/. It is assumed that subjects adopt the prototypes R and L in the task, and evaluate and integrate the two sources of information with respect to these prototypes. The stimulus featural information in level i of the glide supporting the R prototype can be represented by the truth value f_i, and $(1-f_i)$ specifies the stimulus support for L. The value c_j represents how much level j of the context supports the prototype R, and the degree to which the phonological context supports the prototype L is indexed by $(1-c_j)$. Truth values index the degree of support of each source of information for each alternative. These values range between zero and one reflecting no support to complete

support, with .5 corresponding to a neutral support in a two-alternative task.

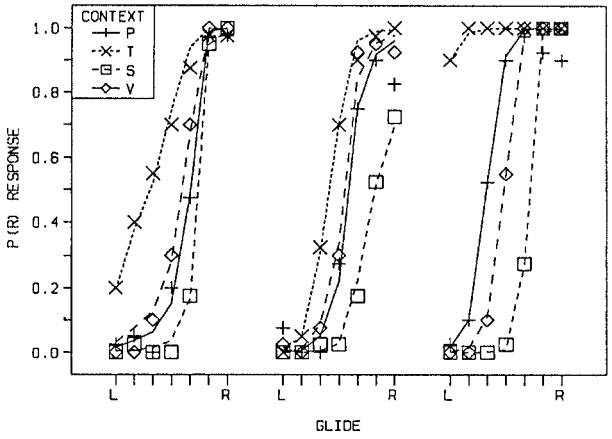

Figure 4. Observed (points) and predicted (lines) probability of an *r* response for three typical subjects as a function of the glide F_3 onset level and context (after Massaro & Cohen, Experiment 2, 1983). Predictions are for the FLMP.

Given two independent sources of information, the total degree of match with the prototypes R and L is determined by integrating these two sources. Feature integration involves a multiplicative combination of the two truth values. Therefore, the degree of match to R and L for a given syllable can be represented by

$$R = f_i \times c_j \quad (1)$$

$$L = (1-f_i) \times (1-c_j) \quad (2)$$

The decision operation maps these outcomes of integration into responses by way of a relative goodness rule (RGR). The probability of an /r/ response given test stimulus S_{ij} is predicted to be

$$P(r \mid S_{ij}) = \frac{f_i\, c_j}{f_i\, c_j + (1-f_i)\,(1-c_j)} \quad (3)$$

Fitting the experimental data requires 11 parameters. These include 4 parameters giving the r-ness of the context c_j and 7 parameters giving the r-ness of the glide f_i. Figure 4 shows the close agreement of the observed and predicted results of the FLMP model for three typical subjects. The average of the RMSDs of fits of the seven subjects was .055, a significantly better fit than the SIAC and SIAC-INT models [8].

Although the SIAC models supposedly predict independence of stimulus information and context, their fit of the actual results falls short relative to the FLMP. Of course, it is possible that another architecture might achieve as good a fit as that given by the FLMP. Given our experience with developing and testing various SIAC models, we expect that achieving a fit of this accuracy will require several more parameters than the FLMP.

IV. REFERENCES

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

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

[4] Bagley, W. C. (1900). The apperception of the spoken sentence: A study in the psychology of language. *American Journal of Psychology, 12*,80-130.

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

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

[7] McClelland, J. L. (in press). Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology.*

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

[9] Chandler, J. P. (1969). Subroutine STEPIT - Finds local minima of a smooth function of several parameters. *Behavioral Science, 14*, 81-82.

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

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.