CATEGORICAL, PROTOTYPICAL AND GRADIENT THEORIES OF SPEECH: REACTION TIME DATA

D. H. Whalen

Haskins Laboratories

ABSTRACT

How do we make phonetic decisions? Categorical, prototypical, and gradient theories were tested using the times to identify a /sa/-/sta/ continuum created by inserting varying amounts of silence into a /sa/ syllable or deleting silence from a /sta/. The gradient model requires 6-8 times as many parameters as the others, and so is difficult to compare. Two variants of a prototypical model and a simple categorical one accounted for some of the variance in the reaction times, but a modified categorical model with the same number of parameters accounts for more. In identification, it seems that all unambiguous syllables elicit identical reaction times, but syllables farther from that range elicit increasingly longer times.

1. INTRODUCTION

When we listen to speech, we are exposed to a great deal of variation in the acoustic waveform, much of which we accept with ease. How is it that we can hear these unique acoustic events and yet extract a few categories from them? Early studies of categorical perception (e.g., [1]) proposed that acoustic variation was not even perceived. Reaction time data from Pisoni and Tash [3] seemed to confirm this notion for plain identification. For those stimuli within a phonetic category, identification times were the same. However, for same/different judgments, physically identical tokens were judged the same faster than ones that differed within the category. They interpreted this finding as evidence that different levels of processing are available to different tasks.

Another theory assumes that phonetic continua are evaluated in relation to phonetic prototypes [4]. In Samuel's account, phonetic decisions should be easiest when the prototypical value is used, and increasingly less easy as the acoustic distance between the stimulus and the prototype increases.

Other explanations of phonetic perception depend on the combination of gradient acoustic parameters. Massaro and Cohen [2], for example, compute phonetic decisions from interactions of two acoustic parameters. They have little to say about experiments with only one factor, however, so their theory will not be elaborated on here.

The present study will test the prototypical model against an extended categorical model in explaining identification times. The extension to the categorical model is that of an ambiguous region, rather than just a single boundary between categories. Such an extension is necessary to account for the fact that there are ambiguous stimuli that subjects can report as being ambiguous, rather than hearing the stimuli first as one category and then as another. Such a modification reduces but does not eliminate the differences between the models.

2. EXPERIMENTAL METHOD

2.1 Stimuli.

A male native speaker of American English recorded several tokens of the nonsense syllables /sa/ and /sta/. These were low-pass filtered at 10 kHz and digitized at 20 kHz on the Haskins PCM system [5]. One token of each syllable was selected, with each having the same duration in the fricative noise and in the vocalic segment. (160 and 240 ms respectively). A continuum of gap closures was made by inserting silence between the noise and the vocalic segment for /sa/. The original silence and the burst were removed from /sta/ and replaced as in /sa/. The values ranged from 0 to 78 in 3 ms steps, yielding 27 values; with two sources, there were 54 unique tokens.

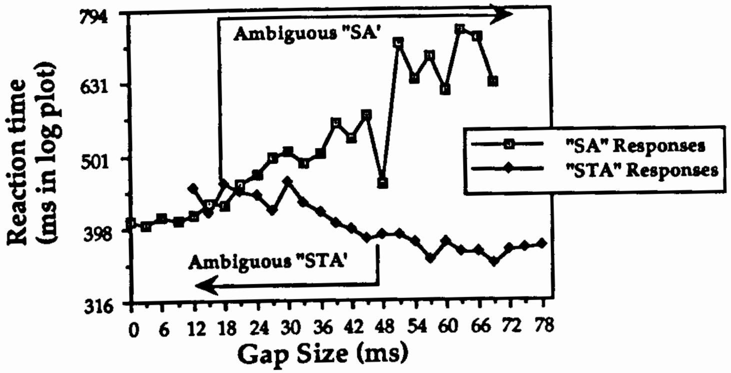

Figure 1: Identification times averaged across 10 subjects.

2.2 Subjects

The subjects were 10 Yale undergraduates who were paid for their participation.

2.3 Apparatus

The stimuli were recorded onto audio tape and played to the subjects over headphones. Their judgments as to whether the syllable was "SA" or "STA" were made by pressing a button, which generated a signal that stopped a clock on an Atari computer, giving the reaction time. Times were assessed from the onset of the vocalic segment, not from the onset of the syllable so that the times would not directly vary with stimulus duration.

2.4 Procedure.

A tape containing twenty exemplars of the stimuli was played to familiarize the subjects with the kinds of judgments they would have to make. Then four blocks, each containing five repetitions of each of the 54 stimuli, were presented. Each block, which had a different randomization of the stimuli, began with four "warm-up" stimuli which were not included in the analysis. A brief rest period was given between blocks.

3. RESULTS AND DISCUSSION

An analysis of the reaction times showed that the subject variances increased as the mean time increased, suggesting a log transform. All further times, though reported in ms, are means of the log values. An analysis that included block and source (original /sa/ or original /sta/) as factors revealed no effect of block, and an effect of source that was the same for both "s" and "st" judgments (the /sta/ source gave slightly faster times). Therefore, further analyses collapsed across these two factors.

It was desirable to eliminate mistaken responses, but the subjects had no way of indicating whether a response was the one intended or not. Instead, "isolates" were excluded. These were responses that were separated from a region of judgments by one or more gaps with no responses of that category. Thus one subject might have "s" responses at the 48 ms gap that would be included in the analysis (since gaps 45 and lower also had "s" responses), while another might have such a response excluded (since at least the 45 ms gap received no "s" responses). Isolates accounted for 1.1% of the data. Figure 1 shows the reaction times averaged across the 10 subjects.

The models were tested by examining how much of the possible variance they could account for. The variance of the individual times in relation to the overall mean established the minimal level for a

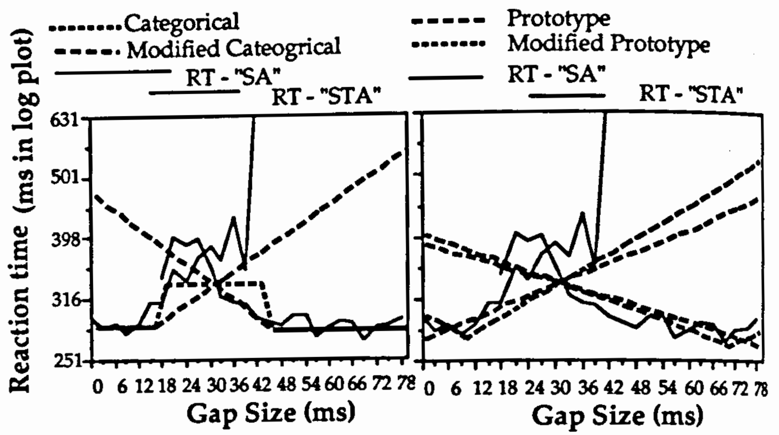

Figure 2: Models generated for the first subject's data (also presented).

model to attain, while using the mean for each judgment for each gap duration established the maximal level. (Since only one acoustic parameter was varied, this maximal description is essentially what would be proposed by a gradient theory, such as Massaro's fuzzy logic model.) The minimal model thus had two parameters (the overall mean for each response), while the maximal model would have between 27 and 54 (since the ambiguous regions could overlap), though the average was 39.9.

Figure 2 shows the four models that were generated for the first subject. (All the modelling was done for each subject individually.) The categorical model was generated with the five parameters: s-boundary (that is, the upper limit of gap values at which 95% of the responses were "s"), the st-boundary (the lower limit of gap values at which 95% of the responses were "st"), the mean times for the "s" region and for the "st" region, and the mean time for the ambiguous region (which included non-isolate "s" responses in the "st" region and "st" responses in the "s" region). In the modified categorical model, the time for the ambiguous responses was calculated from two parameters, a linear interpolation from the edge of the unambiguous region through the mean time for the ambiguous stimuli, temporally located in the center of the ambiguous region.

The prototypical model was generated by taking the fastest time for stimuli of any gap duration as the interpolation value for the continuum endpoints, with the other value being the mean of the responses to ambiguous stimuli (temporally located in the middle of the ambiguous region as in the modified categorical model). The modified prototypical model (which more closely resembles Samuel's) used the location of the fastest time as the definition of the subjects prototype. Values were then interpolated through the ambiguous region as before, and values toward the endpoints were interpolated with a mirror image of the pattern.

Figure 3 shows the percentage of possible variance accounted for by the four models. Since the range of "possible variance" was defined by two more models, it was possible to do worse than the minimum. For two subjects, this was in fact the case for all four models. Subject 5 had very little variation across the gap durations, and had very long times in general (about two standard deviations above the group mean). Subject 7 had a very small "s" range (i.e., only the 0 ms gap), and actually had faster times for ambiguous "st" judgments than for unambiguous ones. Still, for 9 of the 10 subjects, the modified categorical model performed better than the modified prototypical one. An analysis of variance was run on the percentages shown in the figure, with the factors of type (categorical or prototypical) and modification (modified or not). While type was not significant as a main effect (F(1,9) 1.49, n.s.), modification was (F(1,9) 9.38, p <.05), as was the interaction (F(1,9) 8.86, p <.05). As is apparent, only the modified categorical stands out from the others (by a Newman-Keuls post-hoc test).

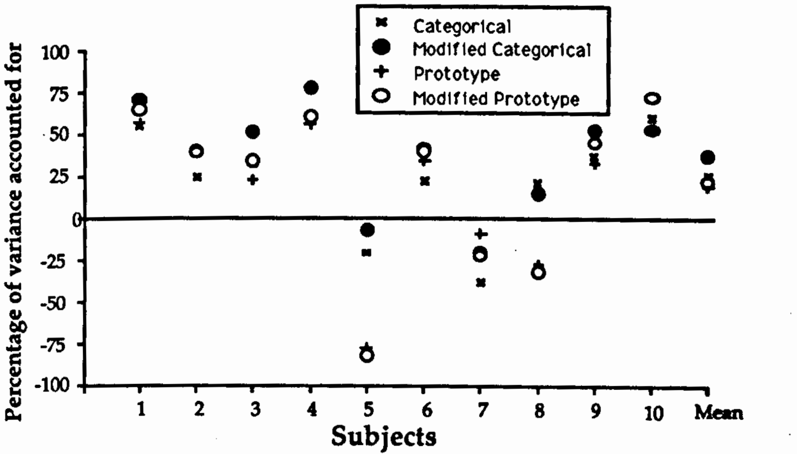

Figure 3: Performance of the four models for the ten subjects. Some of the symbols for the unmodified models are hidden by symbols for the modified ones.

Since the simple categorical model had one more parameter than the simple prototypical one, comparisons between those two models are somewhat problematic. Both modified models, however, required six para-meters, putting them on an equal footing.

This does not exhaust the possibilities for modelling the data, of course. One further modification of the prototypical models would be to allow the interpolation to be parabolic rather than linear. Though initially appealing, such a modification would make it very difficult to tell the prototypical model from the categorical--perhaps giving us a benign ambiguity. It may also be that there is a floor effect on the reaction times. Perhaps the times in the unambiguous regions were subject to, say, a mechanical limitation, so we might have found a more prototypical pattern if the limitation were circumvented. It is possible that a fast repetition (shadowing) paradigm might be useful here.

For the present results, however, it appears that the best model is the one that assumes that all unambiguous judgments are equally easy, while more difficult (due to ambiguity) ones become increasingly so the greater the distance from the category region.

Acknowledgments:

Work supported by NICHD grant HD-01994. Thanks to A. G. Samuel and A. G. Levitt for helpful comments.

References:

[1] Liberman, A.M., Harris, K.S., Hoffman, H.S., and Griffith, B.C. (1957), "The discrimination of speech sounds within and across phoneme boundaries", *Journal of Experimental Psychology, 54*, 358-368.

[2] Massaro, D. W., and Cohen, M. M. 1983 Phonological context in speech perception. *Perception and Psychophysics, 34*,338-348.

[3] Pisoni, D.B., and Tash, J. (1974), "Reaction times to comparisons within and across phonetic categories", *Perc. and Psychophysics, 15*, 285-290.

[4] Samuel, A. G. (1982), "Phonetic prototypes", *Perc. and Psychophysics, 31*, 307-314.

[5] Whalen, D. H., Wiley, E. R., Rubin, P. E. and Cooper, F. S. (1990), "The Haskins Laboratories pulse code modulation (PCM) system", *Behavior Research Methods, Instruments and Computers, 22*, 550-559.