MODELING RECOGNITION OF SPEECH SOUNDS WITH MINERVA2

Travis Wade, Deborah K. Eakin†, Russell Webb‡, Arvin Agah‡, Frank Brown‡, Allard Jongman*, John Gauch‡, Thomas A. Schreiber†, and Joan Sereno**

* Department of Linguistics
†Department of Psychology
‡ Department of Electrical Engineering and Computer Science
University of Kansas, Lawrence, Kansas, USA
twade@ukans.edu

ABSTRACT[1]

This study investigates the extent to which a localist-distributive hybrid formal model of human memory replicates observed behavioral patterns in perception and recognition of appropriately coded language data. Extending previous research that considered for modeled memorization only items with uniform, undefined randomly generated featural specifications, a MINERVA2 simulation was trained to recognize linguistic events and categories at both acoustic-phonetic and phonological-featural processing levels. Results of both test conditions parallel two important effects observed in behavioral data and are discussed with respect to speech perception as well as human memory research.

1. INTRODUCTION

Formal models have been used to emulate various aspects of human speech perception and memory. The purpose of the present study is to model the encoding, storage, and retrieval of speech sounds using the MINERVA2 [1, 2] model. Specifics of this model's simulation of behavioral data regarding recognition memory are discussed below. Furthermore, a series of new experiments designed to study its ability to learn and recognize language data at both the acoustic and phonological levels is presented.

MINERVA2 is a unique memory model in that it combines critical aspects of both localist (in which learned events are assumed to be stored at discrete nodes) and distributed (in which events are represented as ordered patterns of features) model types. In the model, events are themselves comprised of numerically coded feature patterns, but upon learning they are represented as discrete locations in memory. Memory, then, consists of a large collection of observed event items, each represented by a vector of features with values from the set {1 (present), -1 (absent), 0 (irrelevant or unknown)}. Learning an event involves application of a learning rate to each feature, such that not all values are accurately stored, and then copying the vector into *secondary* (long-term) memory. The stored vector is called a memory *trace*, and its retrieval is achieved by presenting a new event *probe*. This probe activates stored traces by simultaneous comparison with each one, resulting in a composite *echo* vector representing the content and intensity of similarity between the probe and any corresponding previously learned vector(s).

[1] This research was supported by grants from the National Science Foundation and The University of Kansas.

MINERVA2 has been shown to simulate findings from behavioral studies involving, among other things, effects of category frequency on recognition of related and unrelated words. In a behavioral study [2], human subjects studied large word lists divided into semantic categories represented by 1, 2, 3, 4, or 5 closely related words, so that occurrence frequency of categories during learning was varied. At test, participants were presented with two items, one old (learned) and one new, and asked to identify the old item. In a *related* condition, the two items were taken from the same semantic category, and in an *unrelated* condition they came from different categories of the same occurrence frequency. As shown in Figure 1, ability to recognize old items decreased with increasing category frequency, due to within-category interference. Additionally, performance was better for related item pairs than for unrelated pairs.

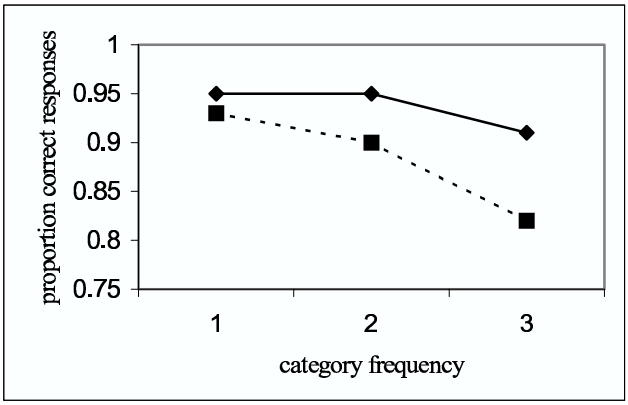

Figure 1: Behavioral data from Experiment 2 of [2]. The solid line represents related word pairs; the dotted line represents unrelated word pairs.

As outlined in [4], MINERVA2 performance on a parallel, simulated recognition task shows both of these important effects.

A commonly observed shortcoming of such studies claiming that formal models such as MINERVA2 are therefore successful in modeling perception of events as complicated and varying as those occurring in natural language, however, involves their formulation of event-defining features for learning. In previous research, feature values are typically generated randomly, with no attempt to replicate or even represent the content or redundancy associated with real-world information. Thus, the present study seeks to investigate the extent to which previously observed (frequency and relatedness) patterns in the model's recognition memory maintain when learned data constitute tenable

representations of phonetic and phonological information. In one experiment, feature values for item formulation denote specifications of language sounds as defined by human phonological patterning behavior; in another, they represent spectral information known to be used in identifying vowel sounds perceptually.

2. EXPERIMENTS

2.1 Experiment 1

The first experiment replicated Simulation 2 from [2], in which simulated MINERVA2 subjects demonstrated recognition memory by performing forced-choice decisions on pairs of test probes. Learning consisted of storing 60 original items, grouped into equal numbers of categories (20 total) containing 1, 2, 3, 4, and 5 members, in a secondary memory with learning rate L=.70. Items each contained 20 features with values from the set {-1, 0, 1}; items within a category were defined by applying a .30 distortion rule to a randomly generated prototype. In the testing phase, recognition of old items by subjects was accomplished by comparing echo intensities between probe pairs of an old item and a new item from either the same category (related condition) or another category of the same frequency (unrelated condition).

2.1.1 Results

Results for Experiment 1 are shown in Figure 2. As in [2], results were in good agreement with human data with regard to pair relatedness and category frequency effects: old items were better recognized when paired with items of the same category compared to items of another category, and accuracy decreased with increasing category membership.

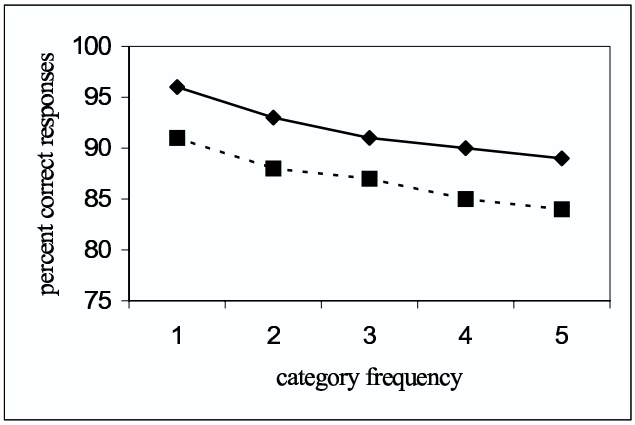

Figure 2: Averaged recognition accuracy of items defined by randomly generated features. The solid line represents related word pairs; the dotted line represents unrelated word pairs.

2.2 Experiment 2

The second experiment involved the same MINERVA2 simulation described above and a similar recognition task, except that items for memorization were representations of possible English CVC words rather than uniform distortions of randomly generated feature lists. Each phoneme in a word was represented by values of +1 (present), -1 (absent), or 0 (unspecified[2]) for each of 18 features considered linguistically relevant based on their phonological behavior, following representations from [3] and [4], for a total of 54 features for each item. These features, with their values for a sample set of seven phonemes, are shown in Table 1. Word categories were defined prosodically, with members within a category sharing (in separate tests) either an onset consonant (C) or a rhyme (VC).

FEATURE	p	d	s	tS	m	I	o
consonantal	1	1	1	1	1	-1	-1
sonorant	-1	-1	-1	-1	1	1	1
continuant	-1	-1	1	0	1	1	1
strident	-1	-1	1	1	-1	-1	-1
nasal	-1	-1	-1	-1	1	-1	-1
lateral	-1	-1	-1	-1	-1	-1	-1
voice	-1	1	-1	-1	1	1	1
labial	1	-1	-1	-1	1	-1	1
round	-1	0	0	0	-1	0	1
coronal	-1	1	1	1	-1	-1	-1
anterior	0	1	1	-1	0	0	0
distributed	0	0	-1	1	0	0	0
dorsal	-1	-1	-1	-1	-1	1	1
high	0	0	0	0	0	1	-1
low	0	0	0	0	0	-1	-1
back	0	0	0	0	0	-1	1
radical	-1	-1	-1	-1	-1	1	1
ATR/tense	0	0	0	0	0	-1	1

Table 1: Features used to create Experiment 2 items with sample specified phonemes.

As in Experiment 1, 20 categories of 1 to 5 members in equal proportions were stored in memory, and simulated subjects compared echo intensities for probe pairs from the same or same-frequency categories. Preliminary testing indicated that non-random, linguistically determinate asymmetry in phoneme-to-phoneme featural similarity was sufficient to obscure any categorically defined relatedness or frequency effects in such small individual lists. Therefore, to uniformly distribute this asymmetry over appropriately large numbers of categories and members, the simulation was extended to generate multiple randomly specified 60-member (20-category) sets and to average recognition accuracy over the separate subjects created for each set. In this manner, item set structure in accordance with previous experiments and satisfying size limitations imposed by the phonemic inventory of English could be maintained while removing effects specific to individual wordlists.

2.2.1 Results

As shown in Figure 3, recognition results for sets of both onset-defined and rhyme-defined categories parallel both human behavioral data and simulated results from [2], showing both pair relatedness and category frequency effects. However, accuracy for each condition is substantially lower here than in the case of randomly generated features, owing to the overall greater-than-

[2]Formally, viewed as a phonological feature-tree node not present due to the absence or (-) value of a dominating node.

chance similarity inherently present between all pairs of CVC words (e.g. all vowels share certain features, as do all consonants).

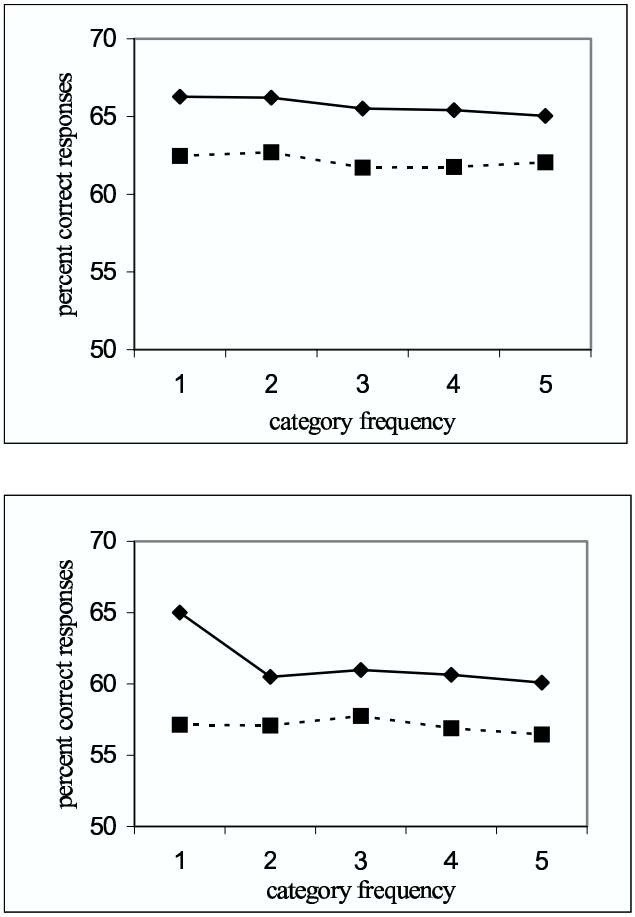

Figure 3: Averaged recognition accuracy of items representing CVC words in onset-defined (top) and rhyme-defined (bottom) categories. Solid lines represent related word pairs; dotted lines represent unrelated word pairs.

2.3 Experiment 3

A final experiment extended these results by modeling the recognition of real acoustic data, where variation within and across categories resulted from human inconsistency in production rather than uniform random distortion of any kind. Experiment, again employing the forced-choice recognition task from Experiment 3, used as trace/probe items acoustic representations of actual vowel productions across a speaker condition. Three tokens each of the English vowel /a/, following the four voiceless fricatives and preceding the phoneme /p/ in a CVC syllable, were produced by 20 speakers. Productions used were previously recorded by 10 male and 10 female speakers as part of [5]. Tokens were digitized at 22.05kHz, and spectral properties of each vowel were analyzed and adapted to feature lists automatically using the sview 1.0 package from the CSLU Speech Toolkit [6]. Spectral intensity values corresponding to 64 equally spaced frequency regions from 0 to 8kHz were first averaged over the 50ms preceding the vowel midpoint. Then, a convolution of this representation along the frequency axis was created by averaging intensity values for the five regions centered around each of the 60 central points, as shown in Figure 4.

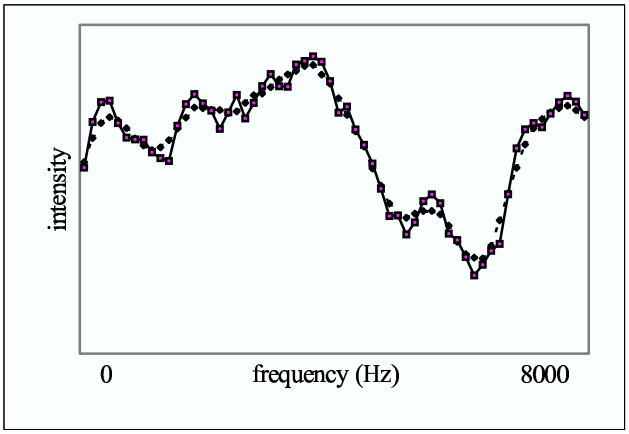

Figure 4: Observed and convoluted versions of typical vowel spectra for feature formulation, with sample feature designations noted; dotted line represents the convolution.

Each vowel was then assigned a 60-member feature representation based on the position of each point on the actual spectrum with respect to its corresponding point on the convolution line: if a given spectral point was an (experimentally determined) large distance above the average of its surrounding points, this point was considered relatively likely to be at or near a local peak in spectral intensity (representing a fundamental or formant frequency), and was assigned a value of (+)1. If a point was similarly below its convoluted correspondent, it was assumed very unlikely to represent such a location and assigned a –1 feature value. If the two representations were nearly equal, a value of 0, denoting an intermediate or irrelevant intensity, was assigned. Thus, (semi-) continuous values for both frequency and intensity of each vowel's spectrum were collapsed into a single one-dimensional array of discrete values.

Items were then grouped into categories with members sharing speaker subjects. As in Experiment 2, category frequencies were randomly chosen from among the 20 speakers, and appropriate numbers of individual tokens were chosen for each category to create lists for memorization. Large numbers of lists were generated as in Experiment 2, and forced-choice recognition accuracy values for the task described above were averaged over the lists.

2.3.1 Results

Results for Experiment 3 are shown in Figure 4. Once again, relatedness and frequency effects were clearly maintained in agreement with previous studies. In this case, overall recognition rates were in general higher in all cases than in previous conditions due to the larger (60 features) item size compared to Experiment 1 and less within-category interference caused by overall somewhat greater dissimilarity between items both within and between categories compared to Experiment 2.

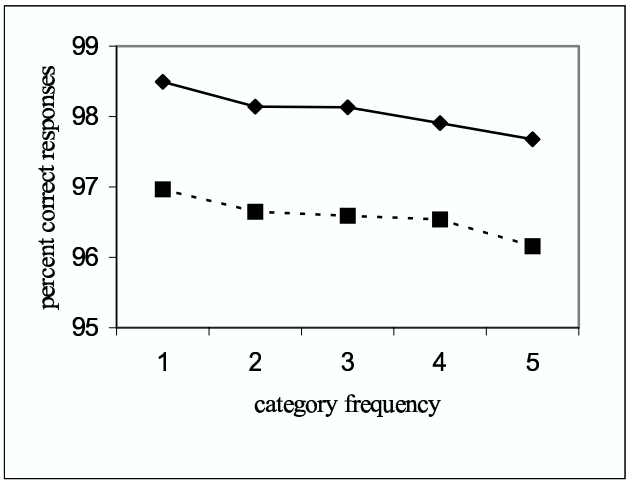

Figure 5: Averaged recognition accuracy of items representing vowel spectra. The solid line represents related word pairs; the dotted line represents unrelated item pairs.

3. CONCLUSION

Results are consistent with the notion that MINERVA2 in fact successfully models human memory for perceived language events. The present experiments demonstrate that relatedness and category frequency effects observed in human word recognition can be replicated by the model in processing not only randomly specified feature lists but also realistic linguistic information. After replication of previous findings in Experiment 1, Experiment 2 showed effects of relatedness and category frequency in recognition of phonological specifications of possible English CVC words. Experiment 3 further extended the model's claims in successfully utilizing acoustic information from actual production data. This latter finding is additionally consistent with behavioral data from research regarding Transfer-Appropriate Processing and Context-Dependent Learning (see [7, 8, 9]), where learning task and/or context (*cf.* speaker category) correlate to recognition accuracy with respect to these same factors at testing. Further research could reveal the extent to which these trends apply to simulated memory at other levels of linguistic processing, such as the semantic and syntactic levels.

4. REFERENCES

[1] Hintzman, D.L. MINERVA2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers,* 16: 96-101, 1984.

[2] Hintzman, D. L. Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review,* 95: 528-551, 1988.

[3] Halle, M. "Phonological features." *International Encyclopedia of Linguistics*, 8: 149-76, 1992.

[4] Kenstowicz, M. Phonology in Generative Grammar. Blackwell, Oxford, 1992.

[5] Jongman, A., R. Wayland, and S. Wong. "Acoustic characteristics of English fricatives", *Journal of the Acoustical Society of America*, 108 (3): 1252-63, 2000.

[6] http://cslu.ogi.edu/toolkit/

[7] McGeoch, J.A. "Forgetting and the law of disuse", *Psychological Review,* 39: 352-370, 1932.

[8] Carr, H.A. Psychology: A study of mental activity. Longmans, Green, 1925.

[9] Morris, C.D., Bransford, J.D. and Franks, J.J. 1977. "Levels of processing versus transfer appropriate processing", *Journal of Verbal Learning and Verbal Behavior,* 16: 519-533.