DO LISTENERS LEARN FOREIGN VOWEL CATEGORIES AS DISJUNCTIVE SETS, THROUGH SELECTION ATTENTION, OR AS PROTOTYPES?

John Kingston
University of Massachusetts, Amherst, U.S.A.
Elliott Moreton
The Johns Hopkins University, U.S.A.

ABSTRACT

In exemplar models, category learning is a secondary process that depends on which dimensions the observers' attention is selectively focused [2,7]. We tested adult American English speakers' ability to categorize the nonlow rounded vowels of German, which contrast for the features [high], [tense], and [back]. Earlier work [3] showed that high speaker or context variability impaired learning, but didn't test variability across the different phonemes belonging to the natural class defined by a distinctive feature value. Here, the number of vowels in each natural class was manipulated in the training stimuli. If each class is a disjunctive set of exemplars, more training vowels should, like more speakers or contexts, impair generalization to a new vowel pair. If learning a natural class is instead discovering and attending selectively to the class's defining properties, more training vowels should improve generalization. Finally, if learners abstract a feature-based prototype, learning the class is discovering the relevant feature, and more training vowels should improve learning only so long as they narrow down the choice between the class's defining feature value vs values that vary within the class. Results show that listeners learned feature-based prototypes, for [high] and [back], and did not distinguish between feature-based prototypes and attention-weighted exemplars for [tense].

1. INTRODUCTION

In an exemplar model, a natural class is the set of all tokens of that category which the listener has ever experienced. The evidence for this is that listeners retain and consult information about individual stimulus tokens that is irrelevant to a phonetic task [2]. Most pertinent here is Logan et al.'s [5] demonstration that speaker and phonetic context influenced the ability of Japanese listeners to learn to identify American English /l/ vs /r/. Our own studies of American English listeners' learning to identify German vowels also show that greater speaker or context variability slows learning and impairs generalization to new tokens [3]. Such evidence contradicts traditional linguistic models that represent a phonemic category as an abstract prototype composed only of linguistic features. Moreover, Marcus, et al. [6] showed that infants can learn and generalize patterns that conform to algebraic rules as early as they can learn patterns based on transitional probabilities [8].

In previous research [5], listeners learned to categorize tokens as individual phonemes, e.g. /l/ vs. /r/. In our experiments, they learn to categorize tokens of several vowel phonemes into the larger natural classes defined by linguistic features, e.g., [+high] vs. [-high]. Other linguistic information (e.g., [back]) thus becomes task-irrelevant, like speaker and phonetic context.

2. METHODS

2.1. Design and Conditions

The experiment had two phases: training and training-testing. American English listeners were first trained with feedback to categorize one, two, or three pairs of German rounded vowels according to their values for a single distinctive feature. They were then tested with a novel pair of German rounded vowels.

Listeners were assigned to one of 18 conditions produced by orthogonal combination of three variables. First, were training stimuli drawn from the ends of diagonals through the body of the cube in Figure 1 or from the ends of parallel edges of the cube?

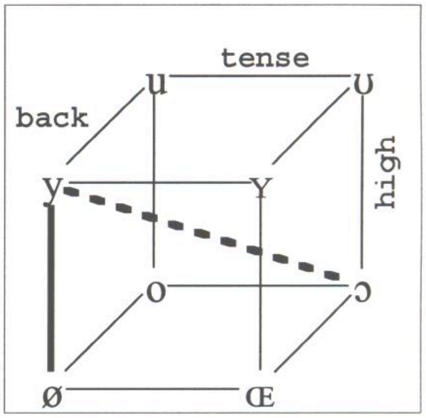

Figure 1. The rounded German vowels, by their values for [back], [tense], and [high]. The heavy line connects a pair of vowels differing just in [high], the dotted line a pair of vowels differing in all three features.

Each face of the cube in Figure 1 represents a natural class, e.g. the foreground face is the class of [-back] (rounded) vowels, the right hand face is the [-tense] vowels, the top face is the [+high] vowels. The pair of vowels connected by the heavy line in this figure, /y:ø/, differ in their value for just a single distinctive feature, [high], being [+high] vs [-high], respectively, and are thus enough to identify those natural classes. The pair of vowels connected by the dashed line through the body of cube, /y:ɔ/, differ instead in all three features, being [+high, -back, +tense] vs [-high, +back, -tense], respectively, and cannot identify any of these natural classes. The second condition is the number of pairs of training vowels, which varied from One to Two to Three. The third condition is the feature by which the listeners were trained to categorize the vowels: [back] vs [high] vs [tense]. Table I below lists the training and test vowels used in each of the 18 conditions:

	Feature	1 pair	2 pairs	3 pairs	Test
Body	back	ʊ:ø	ʊ:ø ɔ:y	ʊ:ø ɔ:y u:œ	o:ʏ
	feature?	back high tense	back tense	back	back
	high	ʊ:ø	ʊ:ø u:œ	ʊ:ø u:œ ʏ:o	y:ɔ
	feature?	high back tense	high back	high	high
	tense	u:œ	u:œ y:ɔ	u:œ y:ɔ o:ʏ	ø:ʊ
	feature?	tense high back	tense high	tense	tense
Edge	back	ɔ:œ	ɔ:œ o:ø	ɔ:œ o:ø ʊ:ʏ	u:y
	high	ɔ:ʊ	ɔ:ʊ œ:ʏ	ɔ:ʊ œ:ʏ o:u	ø:y
	tense	ɔ:o	ɔ:o œ:ø	ɔ:o œ:ø ʊ:u	ʏ:y

Table I: Training and test vowels in Body vs Edge conditions for the features [back], [high], and [tense] when 1 vs 2 vs 3 pairs of vowels are used in training.

For each pair of vowels, the left one was assigned to one response category, the right one to the other. For the Body condition, the features that could be relevant for categorizing the vowels for each number of pairs of training vowels are listed below the pairs of vowels themselves, in the rows labeled "feature?". As shown, in this condition, adding more pairs of vowels narrows down choice between possibly relevant features. Furthermore, for all numbers of pairs of training vowels in this condition, adding the test pair identifies the relevant feature. No comparable list is needed for the Edge condition because the relevant feature can be identified with just one pair of training vowels.

If reducing linguistic uncertainty matters, then listeners should categorize the test vowels more accurately in the Body but not the Edge condition when they've been trained with more pairs of vowels: Body: 1 < 2 < 3 vs Edge 1 = 2 = 3. If this outcome is obtained, the categories are abstract prototypes defined by distinctive feature values.

Although linguistic uncertainty isn't reduced by adding pairs of training vowels in the Edge condition, those additional pairs may still make it easier for listeners to categorize the vowels by distinguishing the relevant from the irrelevant physical dimensions along which the contrasting categories differ. Nosofsky [7] showed that observers categorize a stimulus on the basis of its *aggregate* perceptual similarity to all members of its category rather than its similarities to the category's individual members, and that selective attention stretches perceptual distances — the inverse of similarity — along dimensions relevant to that categorization and shrinks them along irrelevant dimensions. Selective attention should be effective for the natural classes to which speech sounds, even foreign ones like these, belong by virtue of their distinctive feature specifications, because distinctive features have reliable phonetic correlates that will wax in salience in proportion to the number of training pairs. Selective attention should be particularly effective for these German vowels, because they all contrast for features that are distinctive in English vowels, too. If so, then for the Edge as well as the Body condition: 1 < 2 < 3. If this outcome is obtained, categories are sets of exemplars, in which selective attention increases perceived similarity between exemplars within a category and likewise perceived dissimilarity between exemplars of contrasting categories.

Finally, it's possible that adding more training pairs may instead make it harder to learn how to categorize the test vowels in both the Body and Edge conditions, because the differences between the phonemes that make up each category may be difficult to ignore. That is, listeners may learn the category as an arbitrary disjunctive set rather than discovering the physical similarity of its members or the distinctive feature value that defines that natural class. If so, then in both the Body and Edge conditions: 1 > 2 > 3. A drop in performance is also predicted if more training pairs increases each vowel's neighborhood density, as that increases the number of competitors [9]. This outcome, too, indicates that categories are sets of exemplars, but the members of the sets are merely gathered together because the listeners cannot select which dimensions to attend to.

For brevity, we will refer to these as the features, selective attention, and the disjunctive sets models.

The remainder of this section describes the stimuli and procedures used to measure listeners' ability to learn to categorize these vowels in the 18 different conditions just described. So far, listeners have been run in 12 of these conditions: Body vs Edge by One vs Three pairs of training vowels by [back] vs [high] vs [tense].

2.2. Stimuli

Stimuli were two syllable German words or pseudo-words of the shape [ˈCVCən], in which the vowel (V) was one of the rounded, nonlow vowels /u,ʊ,y,ʏ,o,ɔ,ø,œ/ of this language. The stimuli were spoken in isolation by four adult native speakers of German, two men (C,D) and two women (A,E). The two women and one of the men (D) spoke a northern German dialect (Kiel, Hamburg) and the other man (C) spoke a strongly northern-influenced southern dialect (Bavaria). Throughout the experiment, stimuli were presented in two kinds of blocks differing in speakers and the consonants flanking the target vowel: (1) Speakers D and E; Contexts [b_p, t_t, g_k, t_k] and (2) Speakers A and C; Contexts [p_p, d_t, k_k, d_p]. A fourth condition in the experiment is thus the Speaker-Context combination, and it measures the extent to which listeners' learning and generalization depends on the particular stimuli they heard. For brevity, the two Speaker-Context combinations will be referred to by the letters identifying the speakers, DE vs AC.

2.3. Procedures and listeners

Listeners first heard four training blocks, two with each combination of speakers and consonantal contexts. 96 trials were presented in each training block, 48 for each category. With One training pair, each vowel is presented 48 times per block (each token 6 times), with Two training pairs, each vowel is presented 24 times (each token 3 times), and with Three training pairs, each vowel is presented 16 times (each token 2 times). After training, listeners heard four training-testing blocks, again two with each combination of speakers and contexts. In the training-testing blocks 48 trials with the test vowels, 24 for each category (each token 3 times), were added to 96 trials with the training stimuli, for a total of 144 trials. Order of stimulus presentation was random within a block and block order was counter-balanced across groups of listeners.

A single stimulus was presented on each trial, and the listeners identified the category to which its vowel belonged by pressing one of two buttons on a box. Following their response, a light came on above the button the listeners should have pressed. This feedback was given on all trials, throughout the initial training and the subsequent training-testing.

Listeners had 2000 ms after the end of the stimulus to respond, the feedback light came on for 500 ms after their response, and a 500 ms pause followed before the next stimulus was presented.

Peak amplitudes, which always occurred within the target vowel, were equalized in the stimuli. They were presented binaurally through TDH-49 headphones at 20 kHz, low-pass filtered with an 11-pole filter at 8133 kHz (energy was down at least 80 dB at the Nyquist frequency).

Listeners were run in groups of 1-4 inside a quiet room with partitions separating one listener from another.

Listeners were native speakers of American English recruited by advertisement from the University of Massachusetts, Amherst undergraduate student body. All listeners were paid for their participation. None reported any current or past speech or hearing pathology. No listener was run who had studied or been exposed for any length of time to any language with front rounded vowels, including German, Dutch, any Scandinavian language, French, any Chinese language, or Korean.

Each listener was assigned to one of the 12 conditions described above, i.e. they had to categorize the vowels for just one distinctive feature, with just one number of pairs of training vowels, in the Body or Edge condition. So far, 4-6 listeners have been run in each condition. We will run more listeners in each condition and also run listeners in the six conditions with Two pairs of training vowels.

3. RESULTS

Responses from the two presentations of a Speaker Context combination were combined within each of the two phases, training and training-testing, of the experiment, to produce three sets or Stages of responses per combination: First Training, Second Training during the training-testing phase, and Testing. Listeners' ability to categorize the stimuli was measured by calculating *d'* values for each of the three sets of responses, treating their task as yes-no classification. The effects of the various experimental conditions on the values of these sensitivity measures were then evaluated in separate repeated measures ANOVAs for each distinctive feature. The within-subjects variables in these analyses were the Stage at which these measures were calculated, First Training vs Second Training vs Testing, and Speaker-Context combination, DE vs AC. To evaluate learning between the two Training stages and generalization from Training to Testing, planned comparisons were carried out between First and Second Training and between Second Training and Testing. The between-subjects variables were Body vs Edge and Number of pairs of training vowels, One vs Three. If listeners abstract feature-based prototypes from training, then Testing performance should be significantly better compared to Second Training performance in the Body but not the Edge condition when Three pairs of training vowels are used rather than just One. On the other hand, this interaction between Body vs Edge and Number of pairs of training should not be significant if they instead learn to attend selectively to the relevant physical dimension. Because separate analyses were carried out for each distinctive feature, the significance criterion was adjusted to .01.

For the feature [back], listeners' performance was only slightly better during Second than First Training and was decided worse during Testing. These differences are reflected in the significant main effect of Stage [F(2,30) = 20.719, p < .001]. Planned comparison shows that this effect is due to the poorer performance during Testing than Second Training [F(1,15) = 31.275, p < .001] and that there's no difference between First and Second Training [F(1,15) = 1.347, p = .264]. Figure 2 shows that differences in performance between Stages depend both on whether the listener was in the Body or Edge condition and on many pairs of vowels they were trained on.

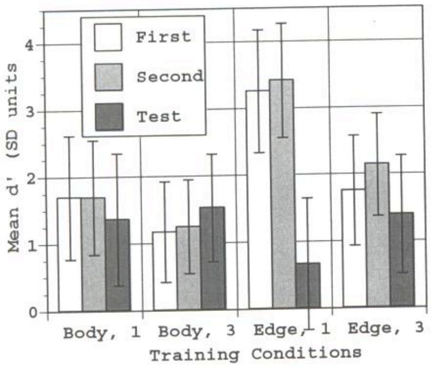

Figure 2. Mean *d'* values (95% confidence intervals) across listeners for categorization by [back] during First and Second Training and Testing Stages.

This figure shows that the novel vowels in the Testing Stage were actually categorized better relative to the old vowels in the Second Training Stage in the Body condition when listeners heard three pairs of training vowels. Additional training vowels also improved categorization of the novel vowels in the Edge condition but that performance remains worse than Second Training. These differences are reflected in significant interactions between Stage and Body vs Edge [F(2,30) = 20.083, p < .001], Number of pairs of training vowels [F(2,30) = 14.824, p < .001], and nearly the combination of these two variables [F(2,30) = 4.221, p = .024]. Planned comparisons show that each of these interactions is significant because Testing performance is different relative to Second Training performance [Stage x Body vs Edge, Testing vs Second Training: F(1,15) = 29.440, p < .001, Second vs First Training: F < 1; Stage x Number, Testing vs Second Training: F(1,15) = 16.957, p = .001, Second vs First Training: F < 1; Stage x Body vs Edge x Number, Testing vs Second Training: F(1,15) = 4.877, p = .043, Second vs First Training: F < 1].

With One training pair, listeners categorized stimuli from the DE Speaker-Context combination better than those in the AC combination in the Body condition (DE 1.817±.903 95% CI vs AC 1.348±.836), but they categorized the stimuli from the AC combination better than those in the DE combination in the Edge condition (AC 2.789±.836 vs DE 2.094±.902). Categorization was no better for one combination than the other with Three training pairs in either the Body or the Edge condition (Body: DE 1.213±.737 vs AC 1.412±.638; Edge: DE 1.762±.808 vs AC 1.756±.748). As a result, Speaker-Context combination interacted marginally with Body vs Edge [F(1,30) = 5.080, p = .040] and significantly with Body x Edge by Number of pairs of training vowels [F(1,30) = 10.339, p = .006]. This clearly isn't evidence of a generally different response to the two Speaker-Context combinations, but instead only an idiosyncratic response from the two groups of listeners who heard One training pair in the Body vs Edge conditions.

Figure 3 shows that Test performance improves even more with more pairs of training vowels in the Body condition for the feature [high] than it did for the feature [back], and that increasing the number of pairs of training vowels does even less good in the Edge condition.

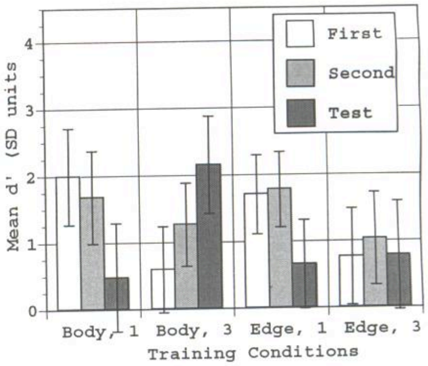

Figure 3. Mean *d'* values (95% CI) for categorization by [high] during First and Second Training and Testing Stages.

This difference between the two conditions is reflected in significant interactions between Stage and Number of training pairs [F(2,30) = 28.162, p < .001] and between Stage and the Body vs Edge by Number of training pairs [F(2,30) = 6.340, p = .005]. (The main effect of Stage is marginally significant [F(2,30) = 4.458, p = .02].) Planned comparisons for the three-way interaction show, however, that neither the difference between Second and First Training Stages nor Testing and Second Training Stages quite reaches significance [First vs Second Training: F(1,15) = 5.512, p = .033, Testing vs Second Training: F(1,15) = 4.506, p = .051].

The four-way interaction of· Speaker-Context combination x Stage x Body vs Edge x Number of training pairs is significant [F(1,30) = 9.236, p = .001], but this can again be no more than an idiosyncratic differences between the responses of groups of listeners.

Figure 4 shows listeners weren't able to categorize the novel vowels for [tense] in the Testing condition any better in the Body than the Edge condition when they heard Three rather than just One pair of training vowels.

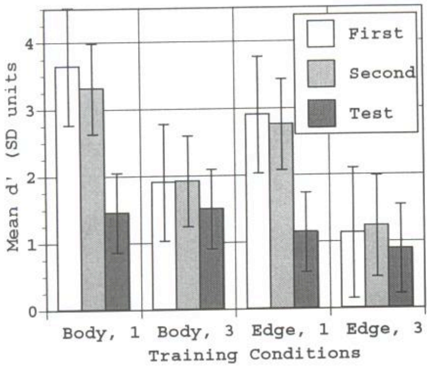

Figure 4. Mean *d'* values (95% CI) across listeners for categorization by [tense] during First and Second Training and Testing Stages.

Instead, they categorize the novel vowels about as well compared to Second Training in both Body and Edge conditions. For this feature, there were significant main effects of Stage [F(2,30) = 24.679, p < .001] and Number of training pairs [F(1,15) = 15.641, p = .001] and a significant interaction between Stage and Number of training pairs [F(2, 30) = 11.747, p < .001]. The interaction stems from the very poor categorization of the Test vowels compared to Second Training with One training pair compared to with Three training pairs. Planned comparisons show that categorization is much worse in Testing than Second Training for One training pair than Three [F(1,15) = 16.822, p = .001] but not in Second Training compared to First Training [F < 1].

The only significant effect of the Speaker-Context combination is an interaction with Stage [F(2, 30) = 6.091, p = .006], which reflects a sharper decline in success in categorizing the vowels in the AC combination than the DE combination from First and Second Training to Testing. Because it doesn't depend on whether the stimuli were drawn from Body or Edge of the cube nor on the Number of training pairs, this is the first genuine exemplar effect in this experiment.

4. DISCUSSION

For two of the three features tested, [back] and especially [high], these results show that listeners have learned feature-based prototypes rather than sets of exemplars. In the Body but not the Edge condition, the novel vowels in the Testing stage were categorized more accurately relative to the Training vowels when Three pairs of training vowels were used rather than just one. The greater improvement in the Body than the Edge conditions disconfirmed the prediction of the attention-weighted exemplar model of equal improvement in both conditions. Contrary to the predictions of the disjunctive exemplar model, phonetic variation is different from speaker and context variation; it helps while they hurt.

These results also show that adding more pairs of training vowels in the Body condition doesn't increase competition between training and testing stimuli. Our listeners undoubtedly to treat these stimuli as non-words. Increasing the number of similar neighbors facilitates non-word recognition, by increasing the likelihood of the of similar phoneme strings [9]. To explain these results, the similarity must be in the natural classes defined by distinctive feature values and not particular sets of exemplars.

No difference in the effect of Number of training pairs was found between the Body and Edge conditions for the feature [tense], however. Listeners' categorization of vowels for that feature also doesn't support either exemplar model, because performance on the novel stimuli is neither improved nor impaired in both the Body and Edge conditions by increasing the Number of training pairs.

These results confirm in detail preliminary results [4], where more training pairs improved categorization of vowels for [back] and [high] but not [tense] in the Body condition, but did not do so in the Edge condition. In that earlier experiment every listener categorized the vowels for all three features, although with different numbers of training vowels for each feature.

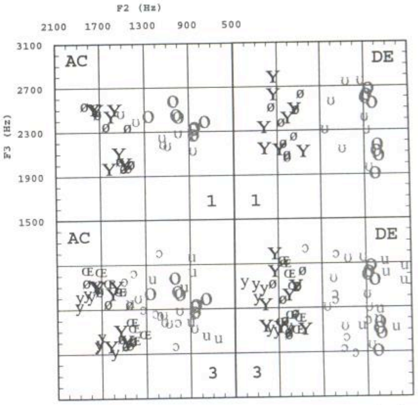

Figure 5. F_2 and F_3 values of stimuli to be categorized as [+back] (gray) vs [-back] (black), AC (left) vs DE (right) Speaker-Context combinations, 1 (top) vs 3 (bottom) training pairs. Tokens of the Test pair, /o:ʏ/, are larger than the Training vowel tokens.

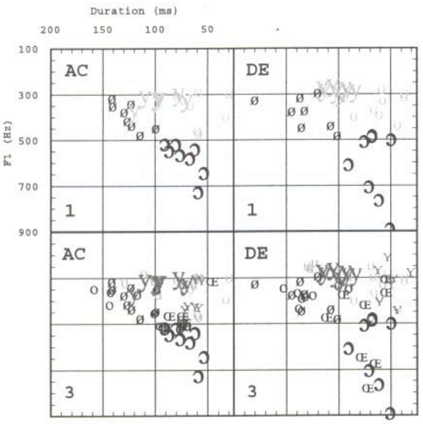

Figure 6. Duration and F1 values of stimuli to be categorized as [+high] (gray) vs [-high] black. The Test pair is /y:œ/. Other conventions as in Figure 5.

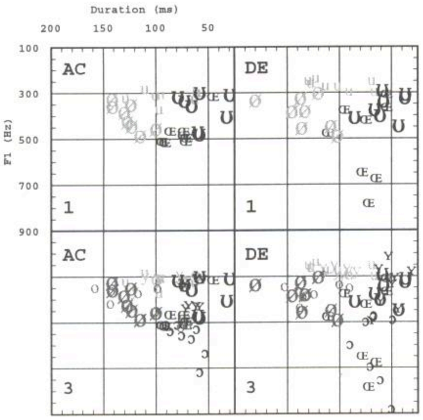

Figure 7. F1 and duration values of stimuli to be categorized as [+tense] (gray) vs [-tense] (black). The Test pair is /ø:ʊ/. Other conventions as in Figure 5.

At the meeting, we will add quantitative modeling to this qualitative comparison of competing models' predictions of category learning and generalization. The distributions of tokens contrasting for [back], [high], and [tense] with respect to pairs of physical dimensions that reliably separate the natural classes in each case are displayed in Figures 5-7.

These displays suggest that feature-based prototypes could be based on deterministic, perhaps even linear rules ones [1]. Vowels with lower F2 values would be categorized as [+back], those with higher F2 values as [-back] (Figure 5). Vowels with shorter durations and lower F1 values would be categorized as [+high], those with longer durations and higher F1 values as [-high] (Figure 6). Vowels with longer durations and perhaps lower F1 values would be categorized as [+tense], those with shorter durations and higher F1 values as [-tense] (Figure 7). Yet these figures also show how categorization could depend on the similarity of each token to all other members of its natural class [7].

5. ACKNOWLEDGEMENTS

This research supported by NIDCD, NIH, grant DC01708, and the U. of Massachusetts. John Kingston, Ling. Dept., U. of Massachusetts, Amherst, MA 01003, jkingston@linguist.umass.edu, 1-413-545-6837.

6. REFERENCES

[1] Ashby, F. G., & Gott, R. E. (**1988**). "Decision rules in the perception and categorization of multidimensional stimuli," *J. Exptl. Psych.: Learn., Mem., & Cog.*, **14**, 33-53.

[2] Goldinger, S.D. (**1996**). "Words and voices: Perception and production in an episodic lexicon," K. Johnson & J. Mullenix (eds.) *Talker variability in speech processing*, (pp. 33-66), San Diego: Academic Press.

[3] Kingston, J., C. Bartels, J. Rice, J.R. Benkí, D. Moore, R. Thorburn, & N.A. Macmillan. (**1996**). "Learning non-native vowel categories," T. Bunnell & W. Idsardi (eds.) *ICSLP 96*, Philadelphia.

[4] Kingston, J., & Moreton, E. (**1998**). "The discovery of natural classes by non-native listeners," *J. Acoust. Soc. Am.*, 103, 2986. (Abstract).

[5] Logan, J.S., Lively, S.E., & Pisoni, D.B. (**1991**). "Training Japanese listeners to identify /r/ and /l/: A first report," *J. Acoust. Soc. Am.*, 89, 874-886.

[6] Marcus, G. F., Vijayan, S., Bandi Rao, S., Vishton, P. M. (**1999**). "Rule learning by seven-month-old infants, "*Science*, 283, 77-80.

[7] Nosofsky, R.M. (**1986**). "Attention, similarity, and the identification-categorization relationship," *J. Exptl. Psych.: Gen.*, 115, 39-57.

[8] Saffran, J. R., Aslin, R. N., & Newport, E. L. (**1996**). "Statistical learning by 8-month-old infants," *Science*, 274, 1926-1928.

[9] Vitevitch, M. S., & Luce, P. A. (**1999**). "Probabilistic phonotactics and neighborhood activation in spoken word recognition," *J. Mem. Lg.*, 40, 374-408.