[i e a u] and Sometimes [o]: Perceptual and Computational Constraints on Vowel Inventories

Marc F. Joanisse and **Mark S. Seidenberg**
University of Southern California
Neuroscience Program
Los Angeles, CA 90089-2520
{MARCJ,MARKS}@GIZMO.USC.EDU

Abstract

Common vowel inventories of languages tend to be better dispersed in the space of possible vowels than less common or unattested inventories. The present research explored the hypothesis that functional factors underlie this preference. Connectionist models were trained on different inventories of spoken vowels, taken from a naturalistic corpus. The first experiment showed that networks trained on well-dispersed five-vowel sets like [i e a o u] learned the inventory more quickly and generalized better to novel stimuli, compared to those trained on less dispersed vowel sets. Experiments 2-3 examined how effects due to ease of perception are modulated by factors related to production. Languages tend to prefer front vowel contrasts over back vowels because the latter tend to be produced with more variability. This caused networks trained on an [i e a u] inventory to perform better than those trained on [i a o u]. Thus both acoustic separation of vowels and variability in how they are realized in speech affect ease of learning and generalization. The results suggest that acoustic and articulatory factors can explain apparent phonological universals.

Introduction

Universal tendencies in languages are often cited as evidence that at least some of language's structure is innate, rather than learned. One such pattern is the overwhelming tendency for vowel inventories to be organized into acoustically well-dispersed and symmetrical forms. For example, there are many more three-vowel languages with a triangular [i a u] inventory than the more lopsided [i a o], or worse [ə e a] (see Figure 1). This tendency holds for inventories with any number of vowels. Figure 1 represents prototypic realizations of vowels; as we will explore later, vowels deviate from these prototypes to varying degrees in actual production.

Formal phonology attributes these phenomena to principles of feature markedness (Chomsky & Halle 1968; Clements & Hume 1995): less marked vowels like /i/ and /u/ are common among the world's languages because their feature specifications are simpler than more highly marked vowels like /ɔ/ or /y/. The optimality of a phoneme inventory thus depends on the markedness of its constituents. One problem with this approach is the lack of independent criteria for deciding what is considered "marked."

An alternative approach seeks to explain phonological patterns in terms of phonetic factors including ease of discrimination and precision of production (Ohala 1990). For example, the more easily a group of vowels are to discriminate from each other, the more likely they are to make up an actual vowel system. This theory has been explored using mathematical models based on acoustic properties of vowels to predict the optimal sets for inventories of different sizes (Liljencrants & Lindblom 1972; Boë, Schwartz, & Valée 1994). Because a vowel's quality is determined by the position of its formants in the acoustic spectrum, such models calculate two vowels' contrastiveness as the distance of their respective formants.

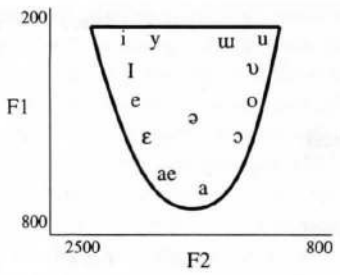

Figure 1: Prototypical locations of vowels in formant space. The dark line indicates the space of possible vowels, given articulatory constraints. Frequencies are plotted in Hz.

The purpose of the present research was to use connectionist networks to explore the hypothesis that languages tend to favor certain vowel sets because they are easier to perceive and produce. Vowels that are more distinct from one another acoustically will be easier to discriminate, easier to learn, and promote better generalization to novel items. Using connectionist networks, rather than the mathematical models used in previous research, allowed us to investigate learning and generalization directly. The first experiment compared the performance of networks trained on acoustically well-dispersed

and poorly-dispersed inventories. Experiments 2 and 3 explored the role of variance in how vowels are produced in speech, and how this biases the world's languages towards front vowels.

Experiment 1

Of the many languages of the world that use five vowels, over 90% of them form the familiar triangle of [i e a o u] (Maddieson 1984). Less common inventories tend to differ from this set by only one or two vowels. In contrast, no vowel system is made up of only back vowels, or only front vowels; such inventories fail to use the entire formant space of vowels and as a result, vowels in these sets are more difficult to distinguish. We hypothesized that poor dispersion would impair the rate with which a simple connectionist network would learn the vowel sets and generalize to novel stimuli.

Method

All inventories used in this experiment have five vowels (Table 1). The first two sets were variations of the typical [i e a o u] type: the *aeiou* set is the most common one, with 58 reported cases (Maddieson 1984), while the *norm5* inventory represents a slightly less common one, represented in 13 of the world's languages. The *schwa* set is very uncommon: only two languages use such an inventory. Finally, the *front5* and *back5* inventories were made up of either five front or back vowels, and are completely unattested in the world's languages.

Attested Sets					
norm5:	i	u	*aeiou:*	i	u
	ɛ	ɔ		e	o
	a			a	
schwa					
(less common):	i	u			
	e ə				
	a				
Unattested Sets					
front5:	i ɪ e ɛ æ		*back5:*	u ʊ o ɔ a	

Table 1: Vowel sets used in Experiment 1.

Between 20 and 64 instances of each vowel type were extracted from the DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus[1], made up of waveforms of elicited speech from different speakers of American English. Only American English speakers from a single dialect region ('North Midland') were used. Each 180 ms. vowel waveform was converted to a matrix of 2032 spectral coefficients, using a Fourier Transform. Each coefficient corresponded to the amplitude of a narrow frequency band for a given point in time. By transforming these coefficients to the 0-1 scale, we obtained the coefficients for each vowel, which provided the input activations of a feedforward network. The network's output consisted of 11 nodes, each corresponding to a different possible vowel. All networks also had a hidden layer of 30 units.

Each of the five sets of vowels were trained on three networks, for a total of 12 networks. On each training trial, the network was presented with the spectral data from a randomly selected instance of a vowel. The probability of each vowel type was held constant, so that the network received an equal number of training trials for each type. The network was trained to map the spectral matrix onto the correct output vector, using the backpropagation algorithm to adjust connection strengths.[2] One vowel was presented at each trial, and all networks were trained on each inventory for 100,000 trials.

Results

To assess the rate at which vowel sets were learned, the proportion of correctly learned training vowels was examined at each 10K training trials. Performance was assessed relative to a criterion of 95% correct on the training set, using a nearest-neighbor scoring method. The *norm5* networks reached this criterion by 30K trials, while the *aeiou* and *schwa* networks reached it by 40K trials. The unattested *front5* set took much longer, 70K trials, while the *back5* network never reached criterion: its asymptotic rate of 92% correct vowels was attained at 90K trials.

We also used a set of testing stimuli to assess each network's generalization to novel tokens. This set consisted of 6 instances of every vowel type in the network's training inventory, also taken from the TIMIT database. The novel vowels were presented to the networks, and the difference between the expected output and the actual output was calculated, using a sum squared error (SSE) formula. Figure 2 shows the mean SSE's and standard errors of each network type on the testing sets. Results show that the attested networks *aeiou* and *norm5* had faster learning rates, indicated by a steeper initial slope, and settled into a lower overall error score when fully trained. The unattested inventories *back5* and *front5*, on the other hand, had much higher error values. The *schwa* set had lower error than these unattested sets, though error was not as low as the better-attested sets.

The higher SSE's in the unattested sets were due to both a greater number of incorrectly identified vowels in the training set, and the fact that these networks failed to learn one of the training vowels altogether. A vowel-by-vowel analysis of each network type's performance is plotted in Figure 3. The height of each bar represents the number correct of 6 testing vowels of each type. In the two unattested inventories, one vowel is completely missing, indicating that a contrast has been neutralized. Overall bar heights also tend to be lower, indicating a higher overall rate of incorrect responses.

Discussion

The main finding from the simulations was that the attested vowel sets were better learned and promoted better general-

[1] The TIMIT database does not contain an equal number of each type of vowel, so all available instances of each vowel were used.

[2] The learning rates and weight ranges for all networks in this paper were set at 0.01.

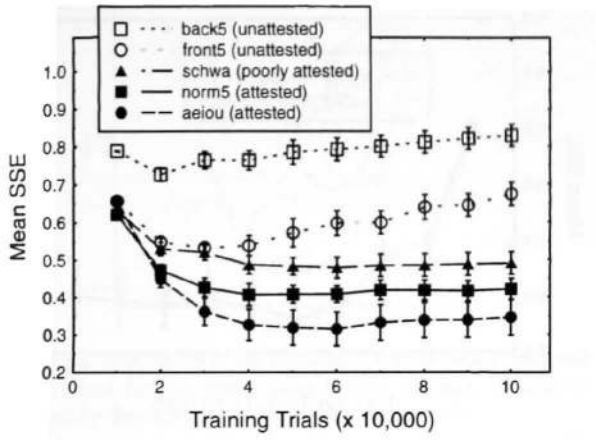

Figure 2: Mean sum squared error on generalization sets for each vowel set, over time. Data are averaged across 3 different simulations for each set.

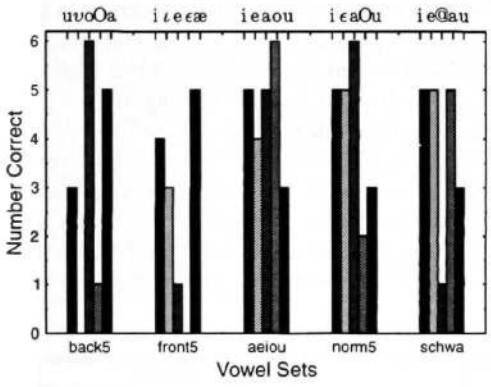

Figure 3: Number of correct responses for each network type at asymptote, broken down by individual vowels in the testing sets.

ization than the unattested sets. Given the architecture that was used, the unattested sets could be not be learned; one vowel in each set was unlearned and mean SSE was increasing rather than decreasing with additional training.

These results indicate that connectionist networks, like humans, show a preference for vowel inventories that are acoustically well-dispersed. Moreover, the model's relative performance on the 3 attested vowel sets corresponds to their relative frequencies in the world's languages. Thus, the simulations accounted for both differences between possible and impossible vowel inventories and differences within the possible sets in terms of relative learnability. The model's behavior was shaped by two main constraints. First, because there is considerable variability in how vowels are realized in speech, the model must develop representations of vowels that allow it to correctly classify many different stimuli as instances of the same type. Second, the model must also be able to accurately differentiate between instances of different types. The vowel inventories that are learnable are thus ones that allow both of these demands to be met simultaneously. The model's relatively accurate performance on novel examples suggests that it developed prototype-like representations that allowed accurate classification of novel stimuli that deviated from previous examples. Humans also exhibit this capacity, as demonstrated by phenomena such as categorical perception (Liberman, Harris, Hoffman, & Griffith 1957).

The learning and generalization performance of these models underscore the relationship between language acquisition and language processing in humans. There is an intimate connection between factors influencing the ease with which children can acquire a hypothetical inventory, and the efficiency with which it can be processed.

In this experiment, none of the attested sets were learned perfectly, but it should be noted that vowel acquisition and recognition normally rely not only on acoustic cues, but also on local phonemic context, lexical knowledge, and discourse information. The conditions under which the networks were trained were relatively impoverished insofar as none of these additional sources of top-down constraint were available.

In summary, the results of the first simulations are compatible with linguistic data showing that vowel inventories that are more highly dispersed are more common than those that are less highly dispersed. Unlike formal theories of markedness, however, our models did not rely on explicit feature hierarchies to explain these facts; rather, they derive from how the models learn given the task they have been asked to perform and the nature of the input.

Experiment 2

Although models like Lindblom's (Liljencrants & Lindblom 1972) stress acoustic distance between vowels, it is also important to consider facts about how they are actually realized in naturalistic speech. In particular, vowels are produced with differing amounts of variability, which also affects learnability. Consider the 4-vowel inventories: the UPSID database of language inventory patterns (Maddieson 1984) lists 12 four-vowel languages with the vowels [i e a u] (the most frequent four-vowel inventory), but only two with [i a o u]; the difference between the two is the choice of a mid vowel. The greater number of languages containing the [i e a u] inventory might imply that this set is better dispersed and therefore easier to learn, but in fact, the two are equally dispersed; looking back at Figure 1 it is clear that the acoustic space between /i/ and /e/ is roughly the same as between /u/ and /o/.

Theories based on acoustic differences between idealized vowels provide no basis for explaining the difference in the distributions of these vowel inventories. However, differences in the variability with which vowels are realized in speech may be what causes a preference towards front vowel contrasts. Beckman et al. (1995) explain that precise articulation of high front vowels is easier to obtain than for the equivalent back vowels, resulting in a smaller amount of F1 variance in vowels like /i/, compared to /u/. This is because high front vowels like /i/ can be produced with great precision by stiffening the genioglossus muscle, and propping the tongue laterally against the dental ridge. This prevents it from falling into

the domain of /e/, facilitating the contrast between the two. The vowel /u/ cannot be produced similarly, since the dental ridge does not extend far back enough; as a result, tongue height cannot be as accurately controlled. This means that back vowels like /u/ or /o/ are more likely to overlap.

We hypothesized that the increased variance in back vowels would affect discriminability, since /u/ and /o/ are more likely to have overlapping distributions. Consequently, the listener's ability to perceive them as different should also be weaker than in front vowels. We tested this by training simple networks on two realistic four vowel inventories: [i e a u] and [i a o u]. The prediction was that the greater variability of /u/ would yield poorer performance on this second set.

Method

Two new vowel sets were created using waveforms extracted from the TIMIT database. The first set, *front4*, consisted of different instances of the vowels [i e a u]. The second set, *back4*, consisted of the vowels [i a o u]; the first inventory is the more frequent one. The training method was similar to the previous experiment, using the backpropagation algorithm to adjust weights after each presentation of a random vowel from the input set. Because the number of vowels to encode was smaller, and the anticipated effect is subtle, 20 hidden units were used for these simulations. Three networks were trained on each inventory.

Results

The proportion of correctly learned training vowels was assessed at 10K-trial intervals. Criterion was set at 95% correct on training vowels. The networks trained on the *front4* inventory took an average 30K training trials to reach criterion. Networks trained on the *back4* inventory needed 40K trials. All networks attained perfect scores by 50K trials.

To test the networks' ability to generalize to novel vowels, a set of 10 vowels of each type (total = 40) was created for each network type. Each testing vowel was presented to the network at increments of 10K trials. The mean SSE across three runs of each network and standard errors are plotted in Figure 4. As with the training set, performance for the *front4* set was slightly better than the *back4* set at some points, although this difference quickly disappeared. Asymptotic error rates for the two networks were identical.

Discussion

The results did not provide strong support for the hypothesis that the less attested *back4* would be harder to learn because of the variability associated with vowel /u/. It is hard to see how small differences observed between the vowel sets would translate into large differences in the frequencies with which they occur in languages. In considering these results, we wondered how the vowel samples we took from the TIMIT database related to the idealized vowels illustrated in Figure 1. Acoustic analyses were performed on the training vowels, by calculating mean formant values for each vowel

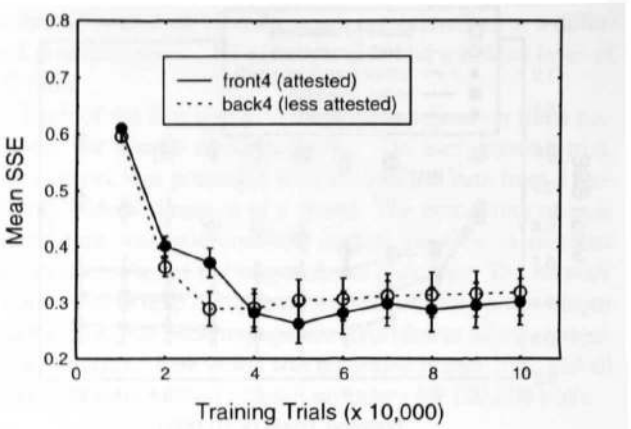

Figure 4: Mean Sum Squared Error scores (and standard errors) for the *front4* and *back4* vowel sets, over the course of training.

type. This revealed an interesting pattern in the back vowels of American English speakers: rather than producing /u/ and /o/ with the very low F2's usually observed in languages, TIMIT speakers produced back vowels that were relatively unrounded and fronted, resulting in F2's that were quite high (Figure 5). By fronting and unrounding /u/ to a greater extent than /o/, the acoustic overlap of these two vowels was reduced (Figure 6), enhancing the mutual distinctiveness of /u/ and /o/. As a result, /u/ and /o/ had less overlap than expected based on idealized representations of the vowels such as in Figure 1. Thus, speakers apparently modified their speech to avoid inventories of vowels that would otherwise be difficult to discriminate.

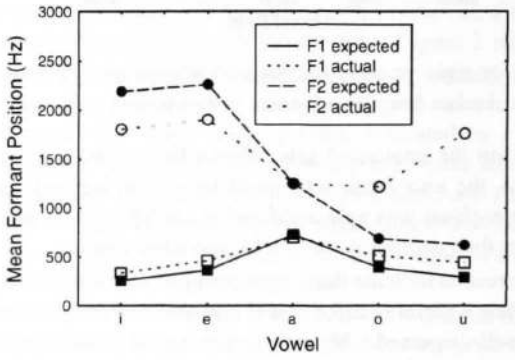

Figure 5: Comparison of typical vowel formant values and those produced by speakers in the TIMIT database. Back vowels /u/ and /o/ differ greatly from their expected positions.

Experiment 3

We have suggested that the *back4* inventory used in the previous experiment contained /u/'s which did not overlap with neighboring /o/'s and this increased the discriminability of the two. This suggests a simple prediction: if the two vowels were moved back to their canonical positions, the variability of the two should create sufficient overlap to significantly impede learning and generalization. To test this, simple feed-

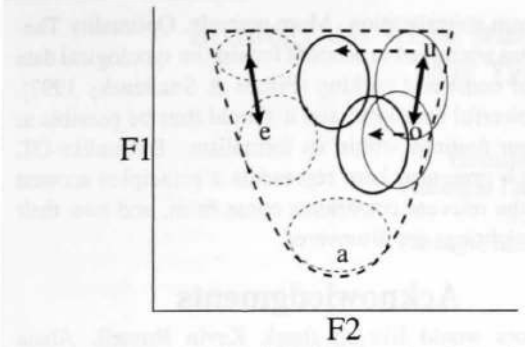

Figure 6: Schematization of the domains of typical back vowels and their American English counterparts (darker ellipses). Overlap is minimized by displacing /u/ to a greater extent than /o/.

forward networks were trained on synthetic vowels that allowed us to control the exact formant and variance parameters of each vowel type.

Method

A set of F1, F2 and F3 means for all five vowels [i e a o u] was devised, based on cardinal vowel positions in Lindblom (1986). These formant values represented vowels' usual positions in languages. Each vowel's standard deviation was obtained from our own analyses of the TIMIT database, along with data from Beckman et al.(1995). Thus, back vowels had a F1 standard deviation that was greater than the corresponding front vowels, due to the effect of poor control over vowel height. The result was a set of vowels with similarly spaced /i/-/e/ and /u/-/o/, but with greater variance in the back vowels.

Using these parameters, 35 instances of each vowel type were generated. Actual synthesis was performed using a parallel-formant speech synthesizer to create 180 ms. waveforms for each vowel. These were then transformed into a set of 2040 spectral coefficients using a Fourier Transform algorithm, and then rescaled to the 0-1 range, and used as the input vectors for our networks.

The synthetic stimuli constituted much cleaner instances of vowels than in previous experiments: length and loudness were all identical, formant frequencies and bandwidths were constant for the entire length of the vowel, and there were no coarticulatory effects at the vowels' onsets and releases. Learning was thus expected to be easier overall; for this reason, a hidden layer of 15 units was used for each network. Each network was trained to associate each set of input vectors to a localist representation of the vowel it corresponded to. Training proceeded as in the previous experiments: a randomly selected input vector was presented to the network at each iteration, and error was computed using the backpropagation algorithm. Three networks of both types were trained, for a total of six networks.

Results

Networks were trained to 100K trials. The criterion of 95% correct on training items was again used to assess overall performance. Each network was tested at intervals of 10K trials. The *front5* network reached criterion by 70K trials, whereas the *back5* network needed 100K trials to reach criterion.

To further assess each network's performance, five novel vowels of each type in the training sets were generated. Each was presented to the network every 10K training trials, and the resulting SSE value was recorded. Figure 7 shows mean error rates and standard errors for each vowel in the training set over the course of training. These results show slower learning of the *back5* set, compared to *front5* set, though in both cases the networks achieved near perfect training by 100K training trials.

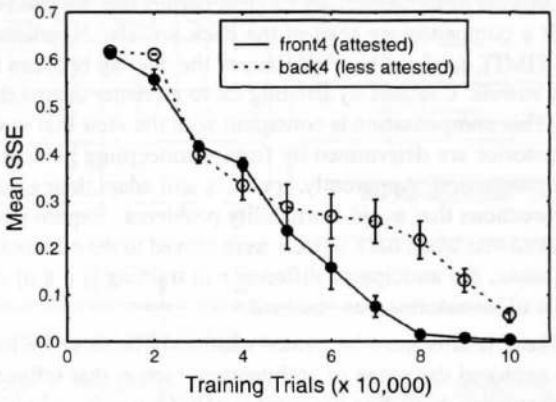

Figure 7: Mean SSE for the synthetic *front4* and *back4* vowel sets, over time. Data are averaged across 3 different network runs for each type.

Discussion

Results of this experiment support the hypothesis that variability in vowels affects learning and generalization in vowel inventories. As predicted, the performance of the *front4* networks was superior to the *back4* networks, in spite of the fact that the relative dispersion of the two inventories was the same. Because of the imprecision with which they tend to be produced, back vowels formed weak and overlapping representations in the *back4* networks. This was less likely to occur in front vowels, since the distinctiveness of /i/ and /e/ can be better maintained by the speaker. This would explain why it is that fewer of the world's 4-vowel languages take the form of [i a o u] than [i e a u].

Conclusion

The simulations we have described examined how factors related to the distinctiveness of vowels could be related to facts about about the distributions of vowels in languages. The first experiment showed that greater dispersion of vowels in a set promoted better learning and generalization, compared to less dispersed inventories of the same size. These results are consistent with the observation that, other factors being equal, contrasts that are more dispersed will be easier for people to discriminate and produce. Given the architecture that we used, the unattested vowel sets were actually unlearnable;

the networks dropped one vowel from each set and also performed worse on the remaining vowels. The second experiment examined another source of constraint, variability in how vowels are realized in speech. Two inventories could contain vowels whose canonical forms are equally distant from each other but overlap in differing degrees because of production variability. We hypothesized that such variability was related to differences in the frequencies of attested vowel sets. This predicted that there would be differences between two attested vowel sets in ease of learning and generalization, and while the effects in Experiment 2 were in the right direction, they were rather small. We then determined that the vowels in the American English corpus that we used exhibit a compensatory shift in the back vowels. Speakers in the TIMIT database have minimized the overlap between the two vowels' domains by fronting /u/ to a greater degree than /o/. This compensation is consistent with the view that vowel inventories are determined by factors concerning perception and production. Apparently, speakers will adapt their speech in directions that avoid learnability problems. Experiment 3 showed that when back vowels were moved to their canonical positions, the anticipated difference in training [i e a u] and [i a o u] inventories was obtained.

These results must be treated cautiously because we have not explored the range of architectural factors that influence performance, including the number of hidden units. Although such factors are likely to affect ease of learning, our assumption is that the observed performance differences between vowel sets will be preserved over a broad range of conditions; this issue needs to be assessed in future work, however.

We also have not examined all of the factors that influence the composition and stability of vowel inventories; there are likely many other sources of discriminability that help determine vowel inventory patterns. For example, some languages use length to contrast acoustically similar vowels. Diphthongization also enhances the contrastiveness of a vowel, by fusing it with another vowel sound. Finally, the proximity of any two formants in a vowel can affect its distinctiveness, which explains why the vowel /y/ is more common that its back counterpart /ɯ/ (Boë, Schwartz, & Valée 1994).

The success of the simulations presented here suggest that this broad range of factors influencing the distributions of vowels in languages may be explainable in terms of constraints on perception, production and learning. The vowel inventories that exist are those that are learnable given these constraints. Further, the relative frequencies of different inventories reflect the graded effects of these variables. Explaining vowel inventories in this way contrasts with the approach taken within generative phonology, which treats these phenomena as the result of innate feature hierarchies that are too complex to be learned. The fact that some combinations of vowels do not occur is then attributed to innate constraints on phonological systems. Our view obviously differs, insofar as it attributes these effects to facts about how vowels are realized in speech, something that such competence theories exclude from consideration. More recently, Optimality Theory (OT) has attempted to account for similar typological data in terms of constraint ranking (Prince & Smolensky 1997). OT is a powerful approach, and it should thus be possible to describe our findings within its formalism. But unlike OT, the research presented here represents a principled account of where the relevant constraints come from, and how their relative weightings are discovered.

Acknowledgments

The authors would like to thank Kevin Russell, Alicja Gorecka, Pat Keating, and two anonymous reviewers for their useful comments on this work. We also wish to acknowledge previous work by Kevin Russell, who originally devised a smaller version of Experiment 1. This research was supported by NIMH grants MH 47566 and MH 01188.

References

Beckman, M. E., T.-P. Jung, S. Lee, K. de Jong, A. K. Krishnamurthy, S. C. Ahalt, & K. B. Cohen (1995). Variability in the production of quantal vowels revisited. *J. of the Acoustical Society of America 97*, 471–90.

Boë, L.-J., J.-L. Schwartz, & N. Valée (1994). The prediction of vowel systems: Perceptual contrast and stability. In E. Keller (Ed.), *Fundamentals of speech synthesis and speech recognition*, pp. 185–213. John Wiley & Sons.

Chomsky, N. & M. Halle (1968). *The Sound Patterns of English*. MIT Press.

Clements, G. & E. Hume (1995). The internal organization of speech sounds. In J. Goldsmith (Ed.), *The Handbook of Phonology*. Blackwell.

Liberman, A., K. Harris, H. Hoffman, & B. Griffith (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology 54*(5), 358–368.

Liljencrants, J. & B. Lindblom (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Journal of Phonetics 48*(4), 839–862.

Lindblom, B. (1986). Phonetic universals in vowel systems. In J. Ohala (Ed.), *Experimental Phonology*. Academic Press.

Maddieson, I. (1984). *Patterns of Sounds*. Cambridge: Cambridge University Press.

Ohala, J. J. (1990). There is no interface between phonology and phonetics: a personal view. *Journal of Phonetics 18*, 153–171.

Prince, A. & P. Smolensky (1997). Optimality: From neural networks to Universal Grammar. *Science 275*, 1604–1610.

Russell, K. How not to learn languages that you can't. Unpublished ms.