Modeling the role of native language knowledge in perceiving nonnative speech contrasts

James L. Keidel*, **Jason D. Zevin**†, **Keith R. Kluender*** and **Mark S. Seidenberg***

* Department of Psychology, University of Wisconsin-Madison

† Graduate Program in Neuroscience, University of Southern California

ABSTRACT

A novel connectionist model of speech perception is presented which accounts for data regarding the perception of foreign speech sounds by native speakers of English. The model was trained on a large corpus of English CV syllables, and then tested on isiZulu stimuli. Results were similar to those obtained in a study with human participants by Best et al. [1]. Importantly, the match between the model's performance and the human data did not depend on the inclusion of articulatory information in the model.

INTRODUCTION

Developmental studies of speech perception indicate that the ability to discriminate some unfamiliar contrasts begins to decline around the first year of life [2]. In the absence of experience with such contrasts, performance remains poor into adulthood. For instance, native Japanese speakers encouter well-documented difficulties learning the English /ɹ/ - /l/ distinction and these individuals do not typically achieve nativelike performance even with extensive training (e.g., [3]). For their part, native English speakers have difficulty with a number of other contrasts, including the Hindi dental-retroflex place contrast [4].

Not all foreign contrasts are equally difficult to perceive, however. The ease (or dificulty) of perceiving a novel contrast depends both on the degree of acoustic similarity between sounds and the relationship of the second-language (L2) sounds to the phonetic inventory of the listener's native language (L1). A number of theoretical models have been proposed to explain the complex interactions that arise between these two factors. Here, we describe a novel connectionist model of speech perception that complements and extends existing work by Flege [5] and Best et al. [1]. The model provides a mechanistic account of the perception of foreign speech sounds, as well as a means of examining long-standing questions about the role of various kinds of representations in speech perception.

In Flege's Speech Learning Model (SLM), L2 speech sounds are perceived relative to L1 prototypes. Thus, the SLM predicts that foreign language (L2) speech sounds should be easiest to acquire if they differ *phonetically* from native language (L1) speech sounds – and from other foreign speech sounds. While the SLM accomodates a wide range of phenomena, it does not make explicit claims about the mechanism by which sounds are assimilated to existing categories. In practice, predictions in the SLM are derived from contrastive analysis of phonological inventories and empirically determined assimilation patterns [6].

Best's Perceptual Assimilation Model (PAM), like the SLM, proposes that assimilation of L2 sounds to the L1 phonetic inventory plays a central role in cross-linguistic speech perception. Unlike SLM, PAM assumes that speech perception is accomplished by direct perception of gestural information. In accordance with this, the model's predictions of patterns of assimilation and discrimination are generated through contrastive analysis of the gestures involved.

In the single category (SC) assimilation pattern, two foreign sounds are mapped onto the same native phone, yielding very poor discrimination performance. For example, the distinction between the isiZulu implosive (/ɓ/) and pulmonic biliabial stop (/b/) was perceived with only about 66% accuracy by native speakers of American English.

Category goodness (CG) distinctions are just that; while both sounds map onto the same category, they do not do so equally well. An example of this is the contrast between the velar ejective stop consonant and the pulmonic voiceless aspirated velar stop (/k'/-/k^h/). Because the former has a much more forceful burst that than either the isiZulu or English /k^h/, it is judged as a significantly worse exemplar of /k^h/. Discrimination for this contrast was therefore predicted to be good, but not perfect. Best et al. found approximately 90% discrimination for these stimuli.

Finally, two category (TC) distinctions involve two sounds that assimilate to different L1 categories. IsiZulu voiced and voiceless lateral fricatives (/ɬ/ - /ɮ/) follow this pattern, with the voiced sound assimilating predominantly to /l/, and the voiceless to /ʃ/. As

might be expected from an /l/-/ʃ/ discrimination task, performance was close to perfect (around 95%).

Thus, success in the discrimination task was related to the manner in which an L2 contrast assimilated to L1 categories, consistent with most, if not all theories of L2 speech perception (including PAM), and with the notion of "acquired similarity" central to the earliest theories of categorical perception in speech (e.g., that of Liberman and colleagues [7]). What remains to be determined, however, is whether these results emerge as a specific consequence of features that differentiate PAM from other theories, such as its commitment to direct realism. Instead, this pattern of data might simply emerge from relations between the statistical structure of the English and isiZulu phonetic inventory, defined on strictly acoustic/auditory dimensions.

As a test of this alternative, we present a connectionist model of speech perception trained on English speech sounds and tested on stimuli similar to those employed by Best et al. This model has two advantages over existing theories. First, because it is computationally explicit, it provides a mechanistic explanation of the phenomena it simulates, and provides a straightforward way of making quantitative predictions. Second, it depends on very general assumptions about statistical learning, rather than particular assumptions about the kinds of internal representations that support speech perception[1].

METHODS

Stimuli

2600 English CV syllables were recorded by eight male native English speakers and one male bilingual Zulu/English speaker. The consonant set included all English stops, fricatives and liquids, as well as the nasal /m/. Each consonant was produced in the vowel contexts /a/, /eɪ/, /i/ /o/, and /u/. Each syllable was digitized at 20 kHz and converted into a cochleagram with a 15ms analysis length and a time step of 10ms using the Praat program (P. Boersma and D. Weenink). Distance between filters was 2 Bark.

Stimuli were then over- and undersampled at rates of 21 and 19 kHz, approximately equivalent to a +/- 5% change in register. This effectively tripled the number of speakers to which the model was exposed, and discouraged overlearning of the training set. Two hundred and nineteen of these syllables were withheld from the training set for testing. In addition, 102 isiZulu stimuli were recorded and cochleagrams made with the same parameters. These stimuli were used exclusively for testing.

[1] Work along the lines presented here may also provide a compelling account of speech perception in general. However, the current claim applies specifically to L2 perception.

Model architecture and training procedure

The network consisted of 5 layers of units. The first hidden layer was connected recurrently to a second layer; this improves the model's ability to track temporal dependiencies in the input. Output units were connected to themselves and each other, allowing for the development of attractor structure. At the beginning of training, each weight in the network was assigned a random value between -0.1 and 0.1. The integration constant was set at 0.3, and a learning rate of 0.001 was used.

The model was trained using the continuous recurrent backpropagation algorithm [8, 9]. The model was run six times with different random weights and order of stimulus presentation on each run. Each stimulus presentation was divided into 50 time "ticks" to capture the time-varying aspects of speech spectra. This input was coupled to two 21-bit acoustic feature vectors, one each for the consonant and vowel. Beginning at time tick 38, activation on the output layer was compared to the target output for the current syllable. Error was then applied to those units whose activation deviated from the target, and changes on the weights to each unit were made on the basis of the magnitude and direction of the error. Training ceased when the sum squared error (SSE) reached asymptote.

Testing procedures

Identification: Identification was simulated by determining the consonant nearest to the literal output of the model. For novel tokens of English phones, percentages reported reflect the match of the model's output to the syllable produced by the speaker. For the isiZulu items, raw percentage of identifications are reported for the modal responses.

Discrimination: Discrimination performance was examined using an analogue of the AXB task[2]. Ten stimuli from each of the categories under study (/ɬ/, /ɮ/, /k'/, /kʰ/, /ɓ/, /b/, all from the same speaker) were used. On each trial, a test stimulus (X) was compared to two other stimuli (A and B), one of which was an exemplar of the same phonetic category, the other of which belonged to a different category. Correct discrimination was scored if the model's response – as determined by computing values on hidden units at a time tick corresponding to the end of the consonant – indicated that the test stimulus was more similar to the stimulus taken from the same category. The proportion of correct responses out of 120 trials run for each contrast on each run of the model is

[2] As tasks such as AXB imply both perceptual discrimination and higher-level psychological processes such as decision-making and working memory, it is not immediately clear how these should be simulated in the model. A Euclidean distance metric was chosen over other potential methods because it makes the fewest assumptions about the nature of these latter processes.

Table 1: Identification and Discrimination Scores for Zulu Contrast

Identification						
	/b/ -	/ɓ/	/k^h/ -	/k'/	/ɮ/ -	/ɬ/
/b/	1	1	-	-	-	-
/k^h/	-	-	.88	.83	-	-
/ʒ/	-	-	-	-	.83	-
/ʃ/	-	-	-	-	-	.78
Discrimination						
	.75		.82		.95	

Note: Non-modal responses included /t^h/ for /k^h/ (.11), /t^h for /k'/ (.17), /z/ for /ɮ/ (.17), and /θ/ for /ɬ/ (.22).

reported as mean accuracy.

RESULTS

English Speech Sounds

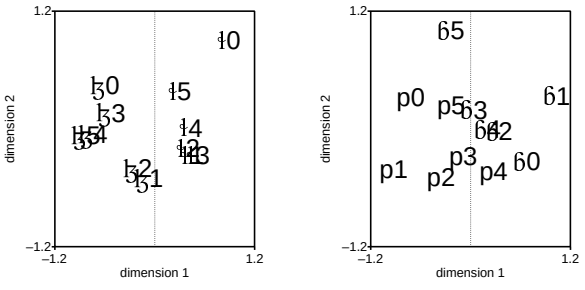
In order to establish that the model had learned to recognize stimuli in its "L1," we tested its performance on both trained and novel English stimuli. Performance on the training set was near perfect (98.4%), and generalization was quite good (94.8% correct). Interestingly, errors consisted mainly of place confusions among stops and anterior fricatives, similar to human performance [10].

Zulu Speech Sounds

Identification: Patterns of assimilation are given in Table 1. Zulu stimuli were always assimilated to English phones, although the consistency of assimilation patterns varied among contrasts (as with human English speakers [1]).

Discrimination: Discrimination performance varied with stimulus type, $F(2, 10) = 17.5, p < .001$. As in the human data, discrimination was best for the TC contrast (/ɮa-/ɬa), somewhat poorer for the CG contrast (/ga/ - /k'a/) and poorest for the SC contrast (/ba/ - /ɓa/). Multidimensional scaling (MDS) plots in Figure 1 provide insight into the internal representations that underlie this discrimination behavior. Note that the lateral fricatives are clearly separable along the dimension that explains the most variance in the dissimilarity matrix, whereas the bilabials form overlapping distributions.

Figure 1: MDS plots of hidden unit activations for two contrasts tested. Vertical lines indicate the origin of the most explanatory dimension



DISCUSSION

A multi-layer connectionist model trained with continuous recurrent backpropagation readily learns to identify a wide range of acoustic stimuli as different consonants and vowels. When the model misidentifies L1 stimuli, it does so in a manner that is largely consistent with the performance of human listeners. Furthermore, the model replicates important findings from Best et al. regarding the relationship between assimilation and discriminability of L2 speech sounds. The pattern of data suggests that the model has learned to distinguish speech sounds in a manner that reflects some important aspects of human speech perception.

Performance on L2 speech sounds is particularly informative. They are perceived as exemplars of familiar (L1) categories. In the course of learning to distinguish English speech sounds, the model learns that certain aspects of the input are highly informative (e.g., VOT within a specific range), but that others are not (e.g., prevoicing, voice quality changes). Thus, when a novel contrast relies on properties of the stimuli which are distinctive in the L1, discrimination is very good. Conversely, when a contrast depends on an aspect of the stimulus the model has learned to ignore, it typically fails to distinguish between the sounds. This behavior is the result of the general ability of associative memory to extract the reliable covariance in a given input domain – in this case, the phonetic inventory of English.

Critically, the model's performance does not depend on knowledge of the articulatory gestures that gave rise to the acoustic waveforms. This suggests that the results observed by Best et al. may lend themselves to an explanation in purely acoustic terms. For example, the acoustic consequences of laryngeal lowering that provide the strongest articulatory cue to the /ɓ/ - /b/ contrast are very subtle, functionally reducing this contrast to a distinction between a pre-voiced and a 0 VOT stop. Because both of these are allophones of /b/ in English, the model learns to ignore this difference and does not reliably distinguish between these two sounds. Interestingly, the larynx *raising* and closing gesture in /k'/ is highly audible because (in tandem with a velar closure and release) it generates a very distinctive "popping" sound followed by silence, on which basis these stimuli can be distinguished from /k^h/. Finally, the acoustic differences between /ɮ/ and /ɬ/ are

both salient and readily assimilable for native English speakers.

Our approach differs from PAM in this respect: On our view, listeners assimilate new sounds as a function of the learned L1 covariance structure defined in the auditory domain; on Best's view, they assimilate in terms of directly perceived speech gestures. The model presented here at least suggests that the direct perception of gestures is not critical to an account of the relationship between assimilation and discrimination observed by Best et al.

CONCLUSIONS

Work with similar models in a very different domain (i.e., reading, [11]) has revealed interesting interactions between generalization and age-limited learning effects. When knowledge of early-learned items generalizes to novel stimuli (as in alphabetic reading), effects of age-limited learning are not observed. When the training corpus of the model was manipulated so that less generalization from early to later items was possible, stronger age-limited learning effects were observed.

Perception of L2 speech sounds offers an opportunity to study a similar interaction between generalization and plasticity more naturalistically. Phonetic inventories overlap among languages. For example, whereas isiZulu contains many consonants which are unfamiliar to native speakers of English, it also contains some contrasts (for example, /b/ - /p^h/) which are essentially identical to the same distinctions in English. Thus, in learning to perceive isiZulu, an English speaker would get at least part of the inventory "for free." A good deal of empirical work – much of it motivated by the SLM (see review in [5]) – suggests that acquisition of L2 speech sounds depends in large part on how these speech sounds are assimilated to existing L1 phones. The current model can be used to generate specific predictions about which speech sounds will be most susceptible to such age-limited learning effects.

We currently are adapting the model to address issues in both developmental speech perception and lexical acquisition. The architecture presented here provides an ideal environment for testing questions about the type of operations that subserve these processes, as well as the time course of their emergence. It is hoped that work in this direction will provide a unified and parsimonious account of early linguistic development.

ACKNOWLEDGEMENTS

Research supported by NIMH grant P50-MH 64445 to MSS and NIDCD grant R01 DC04072 to KRK. Address correspondence to keidel@lcnl.wisc.edu

REFERENCES

[1] C. T. Best, G. W. McRoberts, and E. Goodell, "Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system," *Journal of the Acoustic Society of America*, vol. 109, pp. 775–794, 2001.

[2] J. F. Werker and R. C. Tees, "Cross-language speech perception: Evidence for perceptual reorganization during the first year of life," *Infant Behavior & Development*, vol. 7, no. 1, pp. 49–63, 1984.

[3] S. E. Lively, J.S. Logan, and D.B. Pisoni, "Training Japanese listeners to identify English /r/ and /l/: II. The role of phonetic environment and talker variability in learning new perceptual categories," *Journal of the Acoustic Society of America*, vol. 94, pp. 1242–1255, 1993.

[4] Janet F. Werker and Chris E. Lalonde, "Cross-language speech perception: Initial capabilities and developmental change," *Developmental Psychology*, vol. 24, no. 5, pp. 672–683, 1988.

[5] J. E. Flege, "Second-language speech learning: Theory, findings, and problems," in *Speech perception and linguistic experience: Theoretical and methodological issues*, W. Strange, Ed., Timonium, MD, 1995, pp. 229–273, York Press.

[6] S. Guion, J. Flege, R. Akahane-Yamada, and J. Pruitt, "An investigation of current models of second language speech perception: The case of japanese adults' perception of english consonants.," *Journal of the Acoustic Society of America*, vol. 107, pp. 2711–2724, 2000.

[7] A. M. Liberman, K. S Harris, H. S. Hoffman, and B. C. Griffith, "The discrimination of speech sounds within and across phoneme boundaries," *Journal of Experimental Psychology*, vol. 54, no. 5, pp. 358–368, 1957.

[8] B. A. Pearlmutter, "Gradient calculations for dynamic recurrent neural networks: A survey," *IEEE Transactions on Neural Networks*, vol. 6, no. 5, pp. 1212–1228, 1995.

[9] M. W. Harm, *Division of Labor in a Computational Model of Visual Word Recognition*, Ph.D. thesis, University of Southern California, Los Angeles, CA, 1998.

[10] G. A. Miller and P. Nicely, "An analysis of perceptual confusions among some english consonants.," *Journal of the Acoustic Society of America*, pp. 338–352, 1955.

[11] J. D. Zevin and M. S. Seidenberg, "Age of acquisition effects in reading and other tasks," *Journal of Memory and Language*, vol. 47, pp. 1–29, 2002.