

Learning the sounds of speech: A Hebbian account

Gautam K. Vallabha
James L. McClelland

Center for the Neural Basis of Cognition
Carnegie Mellon University

1 Introduction

The acquisition of speech sound categories shows an interesting paradox. Usually, prolonged exposure to stimuli increases sensitivity to subtle changes^{1,2}. However, in speech prolonged exposure to a category *decreases* the sensitivity, in two ways:

- Loss of sensitivity within a category. This loss results in greater discriminability at category boundaries, and in some cases, a reduced discriminability at the centers of categories³.
- Loss of sensitivity to unused distinctions⁴. This loss makes it much more difficult to acquire the distinctions later on in life, for example when learning a second language.

The loss of sensitivity is similar to that of *acquired equivalence* in general perceptual learning, wherein stimuli associated with a common outcome become less discriminable from each other^{5,6}. It is useful to consider whether speech sound acquisition can be interpreted within this larger context of perceptual learning. This also allows us to consider a unified explanation for the two kinds of sensitivity loss noted above.

In this poster, we describe a general model of perceptual learning, and apply it to the speech domain. In particular, we examine how speech categories can form in first and second language acquisition.

2 Model

The model is constructed around four core principles:

- It should use biologically plausible Hebbian learning mechanisms
- It should be capable of unsupervised learning, but take advantage of feedback when it is available
- It should not rely on "critical period" effects. Difficulty in second language acquisition should preferably be explained by some kind of "representational inertia" rather than by maturational decreases in learning rate.
- It should structurally be a graded, interactive, and stochastic network⁷:
 - Activation is a graded sigmoidal function of net input
 - Activation propagates gradually in time
 - Between-layer connections are excitatory
 - Within layer connections are inhibitory
 - The activation is intrinsically variable

Note: The model is not intended to be a detailed claim about neural function (for example, we don't associate particular layers with brain regions). However, the model principles are consistent with neurophysiology.

Activation update

$$\frac{dnet_k}{dt} = -net_k + \sum_j W_{kj} a_j + \sum_l W_{kl} a_l + \xi_{net}$$

$$\frac{dnet_j}{dt} = -net_j + \sum_i W_{ji} a_i + \sum_k W_{jk} a_k + \sum_l W_{jl} a_l + \xi_{net}$$

$$a_i = I_i + \xi_{input}$$

$$a_j = \begin{cases} \tanh(net_j \cdot \beta_{net}), & net_j > 0 \\ 0, & net_j \leq 0 \end{cases}$$

$$\xi_{input} \sim N(0, \sigma_{input}), \quad \xi_{net} \sim N(0, \sigma_{net})$$

Weight update

$$\Delta W_{jk} = \eta a_j a_k, \quad \sum_k W_{jk}^2 \leq W_{jk}^2$$

$$\Delta W_{ij} = \eta a_i a_j, \quad \sum_j W_{ij}^2 \leq 1$$

$$\Delta W_{ji} = \eta a_j a_i, \quad \sum_j W_{ji}^2 \leq 1$$

$$\eta = \eta_{max} \cdot \exp(-goodness^2 \cdot \beta_{goodness})$$

The model has three layers: an input layer, a representation layer, and an output or category layer. The input layer is topographically mapped to the representation layer, which is fully and reciprocally connected to the category layer.

The units in the representation and category layers excite themselves and inhibit all their neighbors. These intra-layer connections are not modified during learning.

Each exemplar is assumed to be a point in the input space, and it is presented to the network as a "bump" of activity in the input layer. A "category" is a unimodal distribution of exemplars.

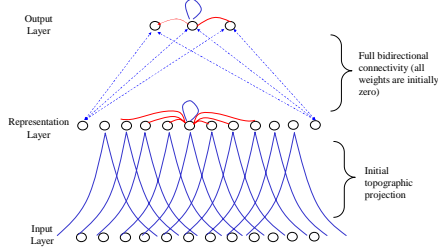
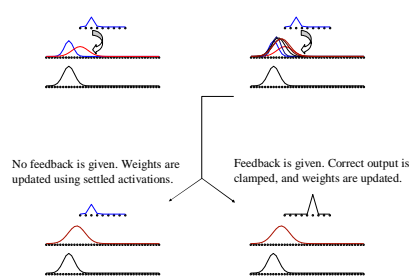


Figure 1. A diagram of the model (the actual model has 2D layers; the figure shows 1D layers for clarity). Blue lines are excitatory connections, red lines are inhibitory.

3 Schematic of model function

- The input to the network is clamped
- The input activity propagates to the representation and category layers
- The category unit recurrently excites the most often activated representation of the category (the "prototypical" representation)
- The representation broadens and shifts towards the prototypical representation



Steps 2-4 occur concurrently, and form a dynamic attractor. The closer two stimuli are to the prototype, the more they are affected by the attractor, and the more difficult it is to distinguish between them.

4 First-language acquisition

The following model illustrates how attractors, once formed, are strengthened by unsupervised exposure.

The three vowel categories to be learnt were [i], [ɪ] and [e]. The network had a 10x10 input layer, a 10x10 representation layer and 3 category units (the input layer covered the F1x2 Bark space). The network was first trained by 6000 trials with feedback being given on 25% of the trials. This was followed by 6000 trials during which no feedback was given. This training had two effects:

- Representational inertia: The weak attractors induced during the initial training became stronger even when there was no feedback.
- Phoneme boundary effect: As the attractors became stronger, discriminability decreased within-category and increased across-category.

This model has a limitation in that initial feedback is needed to position the attractors. This is rather unrealistic, and we are currently exploring how attractors can form entirely without supervision.

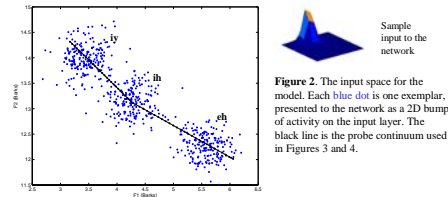


Figure 2. The input space for the model. Each blue dot is one exemplar, presented to the network, as a 2D bump of activity on the input layer. The black line is the probe continuum used in Figures 3 and 4.

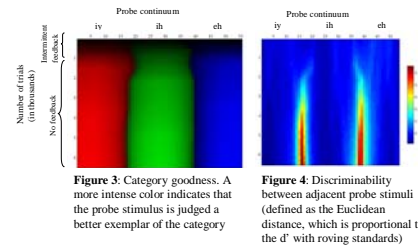


Figure 3: Category goodness. A more intense color indicates that the probe stimulus is judged a better exemplar of the category

Figure 4: Discriminability between adjacent probe stimuli (defined as the Euclidean distance, which is proportional to the d' with roving standards)

5 Second-language acquisition

Once the native-language attractors are established in a perceptual space, they pull in new sounds and prevent the establishment of separate attractors. New categories can still be formed from naturalistic exposure, but only if the new sounds are sufficiently far from existing attractors (cf. Flege's Speech Learning Model⁸).

In a laboratory setting, category acquisition can be facilitated by giving intensive feedback or by presenting the sounds in a controlled manner. We modeled this acquisition using data from an experiment⁹ in which native speakers of Japanese were trained to distinguish the English r-l contrast. We focus here on the *adaptive* training (aka. perceptual fading) conditions:

- The words *lock* and *rock* were recorded from a native English speaker.
- A *lock-rock* continuum was generated from these utterances by spectral interpolation
- Interpolation was also used to create extreme or exaggerated tokens of *rock* and *lock*

Training

- Subjects were initially asked to label extreme tokens of *lock* and *rock*. If they were successful, the distance between the tokens was reduced. If they made mistakes, the distance was decreased.
- Subjects in the "Feedback" condition tried by-trial feedback about their labeling; subjects in the "No Feedback" condition did not get any information.

Modeling r-l acquisition: Assumptions

- The perception of English r and l categories is hindered by the attractor formed by the Japanese apical tap [ɾ]¹⁰.
- Exemplars are characterized by their F2 and F3 onsets.
- The exemplars of each class are distributed as shown in Figure 5.

Procedure

- The model was trained with Japanese tap exemplars until learning stabilized
- It was briefly exposed to the extreme stimuli (so that it could label extreme r and l tokens).
- It was put through the same training conditions as the subjects.

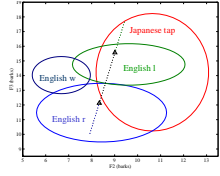


Figure 5. The input space for the model. Each ellipse is the 86% likelihood contour of a bivariate Gaussian distribution. The dots are the *lock-rock* continuum (colored dots are the extreme tokens). The triangles are the original l and r tokens, denoted 0 and 1.0, respectively.

Results are shown in Figure 6. The model qualitatively fits the experimental data, particularly the upward tilt of the untrained categorization curve and the development of discrimination peaks at the category boundary (the location of the peak and the upward tilt of the untrained discrimination curve are not matched, though).

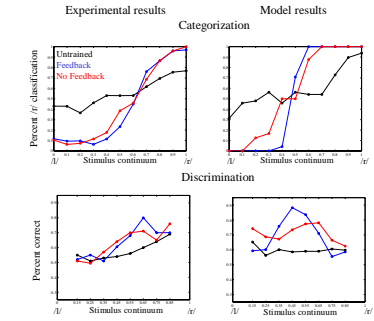


Figure 6. A comparison of the experimental and model results

6 Conclusions

The Hebbian interactive model offers a useful way to think about the creation and maintenance of perceptual categories. The key idea is that recurrent excitation from "higher level" layers can modulate the representations of earlier layers, and provide a basis for outcome-modulated representations (similar ideas have been proposed for visual category formation¹¹). This mechanism is particularly appealing when applied to speech since it can potentially incorporate both Kuhl's Native Language Magnet theory⁹ and Flege's Speech Learning Model⁸.

Ongoing work

The model currently assumes phoneme-based localist output units. This usage is rather dubious given the variety of allophonic and contextual variation in speech. In addition, the model does not address the relation between the old and new categories. In the data modeled in Section 5, the Japanese subjects' classification of [ɾ] and [l] improved after three 20-minute sessions. Do the new categories really have the same status as native language categories that have been entrenched since infancy? We are exploring these issues by seeing whether the model can discover its category representations and whether this can lead to different kinds of categories.

References

- Resnanais, G.H., Schreiner, C.H., & Merzenich, M.M. (1993). *J. Neurosci* 13(1):87-103.
- McLennan, D., & Callaway, N.L. (1999). *J. Exp. Psychol. Human* 25(2):543-560.
- Kuhl, P.K. (2000). *Proc. Natl. Acad. Sci. USA* 97(22):11850-11857.
- Kuhl, P.K. et al. (1992). *Science* 255:606-608.
- Livingston, K.R., Andrews, J.K., & Harad, S. (1998). *J. Exp. Psychol. Learn* 24(5):732-753.
- Gaumnitz, F.H. et al. (1999). *J. Acoust. Soc. Am.* 106(5):2900-2912.
- McClelland, J.L. (1992). In (Meyer D.E., & Konham, S., eds) *Attention and Performance XIV* (p. 655-688). MIT Press.
- Flege, J.E. (1995). In W. Strange (ed), *Speech Perception and Linguistic Experience* (p. 233-272). York Press.
- McClelland, J.L. et al. (2002). *Cogn. Affect. Behav. Neurosci.* 2(2):89-108.
- Gaun, S.G. et al. (2000). *J. Acoust. Soc. Am.* 107(5):2711-2724.
- Golobene, R.L. et al. (1996). In *Proc. 18th Ann. Conf. Cognitive Science Society* (pp. 243-248). Lawrence Erlbaum.