SI Text

Estimation of the speakers' covariance matrices for OME and TOME

In order to minimize bias due to the small number of data points, the covariances for each category for each speaker were estimated using small-sample estimation. For a given vowel category, let C_s be the covariance matrix for speaker s and let $C_{avg} = \sum_s C_s / N$, where N is the number of speakers. Then, for each speaker s, calculate the mixture of C_s and C_{avg} that maximizes the average leave-one-out likelihood (49). This mixed (or regularized) matrix C_{s+avg} is the covariance matrix for speaker s for the given vowel category. The regularized covariance matrices were used to generate the 8000 training points and 2000 test points on each run (for OME), and the 32000 training points and 2000 test point on each run (for TOME). One English speaker had too few productions for this estimation to work and was excluded from the analyses.

Analysis of unsuccessful runs with OME

When the unsupervised learning was unsuccessful, it was typically because two categories were incorrectly merged (or more rarely, because a category was split into two). When there were several unsuccessful runs for a speaker, the failure was always due to the same kind of merger. For example, for the English speaker with no successful runs, /ɪ/ always merged with /ɛ/. Table 2 shows the number of speakers with a particular merger, for both languages. Generally, the categories tend to merge more across vowel color rather than length, consistent with the results in Table 1 showing a greater *d′* for length. It may be noted that the English runs have /ɪ–e/ rather than /ɪ–i/ mergers; the reason is that English /ɪ/ and /e/ have similar F1 values whereas English /ɪ/ and /i/ differ sharply on all three dimensions. Another notable point is the rarity of splits. The parametric nature of OME encourages the discovery of unimodal categories. Consequently, a split (where two estimated categories carve up a unimodal distribution) is very unstable because as soon as one category gains a slight edge, it will eventually completely suppress the competing category. In non-parametric learning, this unimodality constraint is absent and thus splits are more likely.

Within- and Cross-language generalization with OME

Within-language generalization. The training consisted of a single run for each speaker *s*. If the run was successful (see *Evaluation of OME* in *Methods*), the estimated categories were used to classify 2000 exemplars from another speaker *k* of the same language, resulting in a confusion matrix $CM_{s,k}$. The generalization $G_{s,k}$ from speaker *s* to speaker *k* was defined as $100 \cdot Trace(CM_{s,k}) / \sum_i \sum_j CM_{s,k}(i,j))$, and the average generalization *G* as $\sum_s \sum_{k \neq s} G_{s,k} / (K \cdot N–1)$, where *K* is the number of successfully-trained speakers and *N* the total number of speakers.

Cross-language generalization. This measure evaluated the consistency with which exemplars from distinct categories in the test language were assigned to distinct categories in the trained language. Consider first the English training. There was a single run for each English speaker *s*. If the run was successful, the estimated categories were used to classify 2000 exemplars from Japanese speaker k, resulting in a confusion matrix $CM_{s,k}$, where the rows were the (test language) categories /i, iː, e, eː/ and the columns were the (trained language) categories /ɪ, i, ɛ, e/. Let $CM'_{s,k}$ be $CM_{s,k}$ reordered to maximize Trace($CM_{s,k}$). Then, the generalization $G_{s,k}$ from English speaker s to Japanese speaker k was defined as $100 \cdot Trace(CM'_{s,k}) / \sum_i \sum_j CM'_{s,k}(i,j))$, and the average English-to-Japanese generalization G_{EJ} as $\sum_s \sum_k G_{s,k} / (K \cdot N)$, where K is the number of successfully-trained English speakers and N the total number of Japanese speakers. A similar procedure was used to calculate the Japanese-to-English G_{JE}.

Note that $CM_{s,k}$ was reordered separately for each combination of training and test speaker. While this may overestimate the amount of “cross-language generalization”, it avoids assumptions about which category in one language is closest to a given category in the other language.

Inputs and initialization for TOME

Scaling of the inputs. Each input stimulus to the TOME algorithm had to be a point in the 25 x 25 x 25 space (representing F1 x F2 x Duration). Thus, values in each input dimension had to be scaled to be in the 1 … 25 range. For each speaker and each run, the training distributions of the speaker were used to generate the 32000 training points. Then, the *F1* values was rescaled as

$$F1_{scaled} = [(F1 - F1_{min}) / (F1_{max} - F1_{min})] \times (N - 1) + 1$$

where $N = 25$ is the number of units along the F1 dimension, and the extrema were calculated over the training points. The same equation was used, mutatis mutandis, for the *F2* and *Duration* values. Lastly, the extrema values for the training points were also used to scale the 2000 test points.

Initialization. There were $R = 512$ initial categories. For each category unit r, the conditional probabilities over the 25x25x25 input space were initialized to be a spherical Gaussian with mean μ_r and covariance $\beta \cdot I$. The means were systematically placed over the middle third of the 25x25x25 input space; specifically, each μ_r was $[a\ b\ c]^T$ where a, b and c varied over {5 7 9 11 13 15 17 19}. The variance β was set to 1.5, and the mixing probabilities were initialized to $1/R$.

Simulation details for Figure 3

The OME simulations in Figure 3 used the same parameters as the OME vowel learning. In order to do this, the stimuli drawn from the input distribution (Figure 3a) were rescaled to *z*-scores prior to the OME training. After training, the discovered categories were scaled back to the original space (Figure 3b).

The TOME simulations in Figure 3 used the same parameters as the TOME vowel learning (except with a uni-dimensional input space with 50 input units instead of a 25x25x25 input space, and with $\alpha = 0.9$).

Table 2. Learning performance for unsuccessful runs

Language	Median percent correct (supervised)	Number of speakers with mergers[†]						Number of speakers with splits			
		i–iː	e–eː	i–e	iː–eː	i–eː	e–iː	i	iː	e	eː
Parametric model, OME											
English	68.9 (90.5)			4	4	2			1		
Japanese	68.9 (90.5)	2		2	1						
Nonparametric model, TOME											
English	72.1 (92.1)	2		2	6	4		4	3	3	3
Japanese	75.2 (90.9)	1	1	3	1			2	4	4	4

[†] There were 19 English speakers and 10 Japanese speakers.

A run was counted a failure if 95% of the test points were accounted by fewer than 4 categories (“mergers”) or greater than 4 categories (“splits”). The percent-correct value was averaged across all the unsuccessful runs of a speaker; the reported value is the median across speakers. The parenthetical value is the performance of the supervised training. Blank cells indicate zero values. For brevity, the Japanese category labels /i, iː, e, eː/ are also used to designate the English categories /ɪ, i, ɛ, e/.