Psych 229: Language Acquisition

Lecture 11

Speech Perception

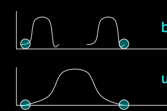
Vallabha et al. (2007): Statistical Learning of Phonemic Contrasts

Testbed: Category emergence for English & Japanese vowel contrasts

Trajectory: 6 month olds have language-specific vowel distinctions

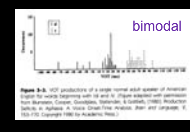
How a statistical method might work

There are two clues to how such a mechanism might work. The first is that infants are sensitive to the statistical distributions of speech tokens. For example, infants exposed to a stimulus continuum with a bimodal distribution were better able to distinguish the end points of the continuum, as compared with infants who were exposed to a unimodal distribution (12).



Motherese makes acoustic properties more salient

The second clue is that infant-directed speech is acoustically different from adult-directed speech, tending to have a slower tempo, increased segment durations, enhanced pitch contours, and exaggerated vowel formants (14–16). Thus, it is possible that the acoustic distributions of infant-directed speech facilitate rapid and robust vowel learning.



Vallabha et al. (2007): Statistical Learning of Phonemic Contrasts

The input

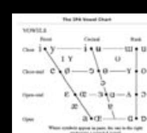
Werker *et al.* (17) recorded the infant-directed speech of English and Japanese mothers. The English mothers produced two vowel pairs, /ɪ/–/i/ and /ɛ/–/e/, in 16 monosyllabic nonce words in both spontaneous and read contexts, while the Japanese mothers produced /i/–/iː/ and /e/–/eː/. These categories occur in the same general region of a multidimensional vowel space defined by formant frequency and duration, but have different phonetic realizations in the two languages. For example, the English /ɪ/ and /i/ differ in both formant frequency and duration, whereas the Japanese /i/–/iː/ differ almost solely in duration (for simplicity, we refer to the vowel pairs in both languages as "length" contrasts, although the English contrasts are sometimes referred to as "lax" vs. "tense").

Important: It's not labeled

Speech to children does not contain category labels, lacks information about the number of categories to be learned, and contains exemplars of different categories in intermixed order.

Vallabha et al. (2007): Statistical Learning of Phonemic Contrasts

A quick look at formants (F1, F2)



F1: depends on whether the sound is more open or closed. (Varies along y axis.) F1 increases as the vowel becomes more open and decreases as vowel closes.

F2: depends on whether the sound is made in the front or the back of the vocal cavity. (Varies along X axis). F2 increases the more forward the sound is.

Idea: As long as speakers use the same values for these formants, they will produce the same vowel.

Vallabha et al. (2007): Statistical Learning of Phonemic Contrasts

The nature of the learning algorithm: incremental

Furthermore, language learners are likely to rely on an online learning procedure: one that adjusts category representations as each exemplar comes in, rather than storing a large ensemble of exemplars and then calculating statistics over the entire ensemble.

Use incremental Expectation Maximization (EM)

These challenges are addressed by the work reported here. We present an algorithm that can be seen both as a variant of EM and as an extension of competitive learning models. The model simultaneously estimates the number of categories in an input ensemble and learns the parameters of those categories, adjusting its representations online as each new exemplar is experienced (24). The algorithm is applied to the problem of discovering the category structure in the infant-directed speech recorded by Werker *et al.* It is "parametric" in that it treats the distribution of speech sounds in a category as an *n*-dimensional Gaussian, and estimates the sufficient statistics of each distribution. We later present a nonparametric variant to investigate the robustness of the learning principles and how they relate to neurologically motivated models (9, 27, 28).

Vallabha et al. (2007): Statistical Learning of Phonemic Contrasts

A brief look at Expectation Maximization

Used for finding the maximum likelihood estimates of parameters in probabilistic models

There are unknown (latent) variables in the model.

Algorithm alternates between doing an expectation step, which computes the expectation of the likelihood by using the latent variables, and a maximization step which computes the maximum likelihood estimates using the expected likelihood found in the previous step. It can then go back to the expectation step, using the results of the maximization step.

Vallabha et al. (2007): Statistical Learning of Phonemic Contrasts

Learner receives sequence of vowel tokens and tried to form categories based off these from all possible vowel categories (1000)

The algorithm treats the vowel stimuli as coming from a set of Gaussian distributions corresponding to a set of vowel categories. Each vowel category is a multivariate Gaussian distribution that has its own overall tendency ("mixing probability") of contributing a token to the data ensemble. The tokens are sampled independently and at random from the ensemble of Gaussians, so that the probability of encountering a particular vowel token is unaffected by the previously encountered tokens. The goal is to recover, given just the sequence of vowel tokens, the number of Gaussians, the parameters of each Gaussian and the respective mixing probabilities. Although this formulation

The vowel portion of each production was characterized by three parameters: the location of the first and second formants (F1 and F2, respectively, measured from the first quarter of the vowel) and the duration of the steady-state. The Gaussians were derived separately for each vowel category of each speaker (see *Methods*; one English speaker was excluded because of an insufficient number of productions). The four Gaussians for each speaker (henceforth, the "training distribution") were used to generate 2,000 data points for each vowel category, for a total of 8,000 training tokens for that speaker.

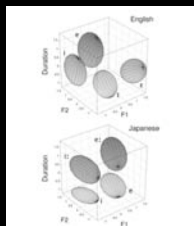


Fig. 1. The Gaussian distributions for the English /i,ɪ,e,ɛ/ and Japanese /i,iː,e,eː/, computed over the read vowels from all speakers (the tokens were z-scored for each speaker before the analysis). The ellipsoids are equal-probability surfaces, 1 SD along each principal axis, enclosing ~19% of the total probability mass. Note that these are aggregate tendencies; the vowel categories of individual speakers varied greatly, covered a wider range, and overlapped with each other considerably.

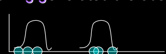
Vallabha et al. (2007): Statistical Learning of Phonemic Contrasts

Each run of the algorithm is initialized with 1,000 equally probable Gaussian categories with randomly initialized means (Fig. 2). On each trial, one token is randomly drawn, with replacement, from the set of 8,000 for that speaker (see *Methods*). The algorithm first calculates the "responsibility" of each category for the token (the responsibility is proportional to the probability of the token given the category's current mean and covariance matrix times its mixing probability). Next, it updates the means and covariance matrices of all categories based on the current token, with more responsible categories receiving larger updates. Finally, it increments the mixing probability of the winning category (i.e., the category with the greatest responsibility) by a small amount η, and reduces the mixing probabilities of all others so that the total probability sums to 1; this update enforces the constraint that each data point should belong to only one category (24). As the training progresses, categories that are very far from input data clusters end up with very low mixing probabilities and "drop out" of the competition. At the end of training, the categories "left standing" are the final estimated categories of the algorithm.

For each token
-"responsibility" of each potential category is calculated
- more responsible categories get larger updates to their means & covariances
- mixing probability (measure of success) of most "responsible" category [estimated] is updated a small amount

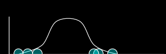
Vallabha et al. (2007): Statistical Learning of Phonemic Contrasts

Basic Idea: Hypotheses are assigned probabilities based on their likelihoods of having generated the observed data



Hypothesis probability raised

Hypothesis: 2 categories
high probability of generating data seen



Hypothesis probability lowered

Hypothesis: 1 category
low probability of generating data seen

Vallabha et al. (2007): Statistical Learning of Phonemic Contrasts

Fig. 2. An illustration of unsupervised OME learning for 2D inputs. Before training, there are a large number of categories (grid of black circles), all equally probable and each sensitive to a different part of the input space (shaded ellipses). During training, the inputs (filled points) are presented one at a time. After training, categories that do not match the inputs have mixing probabilities close to zero (open circles in the grid), whereas the dominant categories have adapted to the input clusters.

- 50,000 data points to train on
- 2,000 data points tested on

Measure of Success

Each test point was classified with the category that had the greatest likelihood for that point. The run was considered "successful" if 95% of the test points were classified into four categories. For evaluation purposes, the categories were also assigned labels (e.g., the category to which most of the /i/ tokens were classified was labeled /i/).

Vallabha et al. (2007): Statistical Learning of Phonemic Contrasts

Evaluation Metrics

This allowed the test performance to be characterized by a confusion matrix, from which three measures were derived: the percent-correct (the proportion of the 2,000 test points that were correctly classified), the length d' (sensitivity in distinguishing /i,e/ from /iː,eː/ in Japanese speech, /ɪ,ɛ/ from /i,e/ in English speech), and the spectrum d' (sensitivity in distinguishing /i,iː/ from /e,eː/ in Japanese speech, /ɪ,i/ from /ɛ,e/ in English speech). Ten independent runs were carried out per speaker, with training and test points being drawn anew on each run.

Language	No. of speakers w/successful runs*	Average no. of successful runs*	Median percent correct†	Median d' for length discrim.†	Median d' for spectrum discrim.†
		Parametric model, OME			
English	18 of 19	7.7 ± 2.8	92.7 (93.4)	3.91 (3.90)	3.19 (3.22)
Japanese	10 of 10	7.9 ± 3.0	91.1 (91.9)	4.09 (4.09)	3.32 (3.30)

*Speakers with successful runs, with 10 runs per speaker.
†Percent-correct and d' values are medians across speakers of the average over successful runs within a speaker. Parenthetical values show supervised training results.

Vallabha et al. (2007): Statistical Learning of Phonemic Contrasts

What about inter-speaker variation within the same language?
Does that affect the categorization ability?

One issue here is that, although English and Japanese vowel spaces are clearly different (Fig. 1), there is also considerable variability between speakers of the same language. This raises the following question: Can the productions of an individual speaker support the discovery of speaker-general but still language-specific structure? To assess this, training with each speaker was tested with all other speakers of either the same language [within-language generalization (WLG)] or the other language [cross-language generalization (CLG)]. In the latter case, test performance was measured by the consistency with which exemplars from distinct categories in the test language were assigned to distinct categories in the trained language [see supporting information (SI)]. The WLG proved to be substantially greater than the CLG: the average WLG was 69% (English training) and 77% (Japanese training), whereas the average CLG was 51% (English training) and 53% (Japanese training). Almost identical results are obtained with supervised training. It therefore is clear that the productions of individual speakers contain substantial language-specific information. Even so, the superiority of the same-speaker test performance (92.7% and 91.1%; Table 1) over the WLG suggests that robust acquisition of vowel categories depends on exposure to multiple speakers (32).

Vallabha et al. (2007): Statistical Learning of Phonemic Contrasts

Assumption of the OME mode: categories have Gaussian distribution

Part of the success of the OME algorithm stems from the assumption that the categories are Gaussian. This places strong constraints on the category representations and limits the number of parameters to estimate for each category (namely, its mean and variance).

A model that doesn't do this: TOME

). We now consider an alternative that allows us to examine the extent to which learning can occur without the Gaussian constraint while also moving closer to a possible neurobiological implementation. In this variant, the distribution of each category is represented nonparametrically, by dividing the input space into many small regions and tabulating the proportion of inputs in each region (33). This scheme has a natural "neural network" interpretation: the proportions can be encoded as connection weights between neuron-like units standing for the input regions and units standing for the category representations. The resulting learning algorithm has similarities to connectionist models of categorization (31), topographic map-based perception (8, 18, 19), and competitive Hebbian learning (34, 35), and we refer to it as "Topographic OME" (TOME).

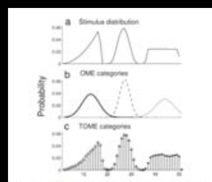


Fig. 3. A comparison of representations learned by OME and TOME for a 1D input space (see SI). (a) The distribution from which the input stimuli were drawn. (b) The three categories discovered by OME. Each line shows the conditional probability of one category, multiplied by its mixing probability. (c) The three categories discovered by TOME, with 50 input units. The distributions are shown as discrete to emphasize the histogram-based category representation in TOME.

Hebbian learning: neural network, "cells that fire together wire together" - building associations

Vallabha et al. (2007): Statistical Learning of Phonemic Contrasts

TOME process

For learning the infant-directed vowels, we used a 3D (25 × 25 × 25) grid of input units (the three dimensions being F1, F2, and duration), along with R = 512 category units. The mixing probabilities for the categories were initialized to 1/R, and the conditional probabilities were initialized to spherical Gaussian distributions, with the centers placed systematically over the input space (see SI). As with the OME learning, training was done separately for each speaker, with 10 runs per speaker. On each run, 32,000 input stimuli were drawn from the same training distributions used for the OME learning, and rescaled to match the input grid. On each trial, one stimulus was drawn from this set of 32,000 (with replacement) and presented to the network (see *Methods*).

Operation of OME. On each trial, the algorithm goes through six steps, summarized as (*i*) get the input stimulus D; (*ii*) calculate the likelihood of D for each category r; (*iii*) calculate the responsibility for each category r; (*iv*) update the parameters for each category r; (*v*) update the mixing probability for winning category $\hat{r}$; and (*vi*) Ensure mixing probabilities sum to 1. p_r is the likelihood of the data point given category r, and $Resp_r$ is the corresponding responsibility.

Operation of TOME. On each trial, the algorithm goes through six steps paralleling those in OME. p_r is the likelihood of the data point given category r, $Resp_r$ is the corresponding responsibility, $w^r_{i,j,k}$ is the connection between the input unit at (i, j, k) and category unit r, and the indices i, j, and k range from 1 . . . 25.

TOME results

Language	No. of speakers w/successful runs*	Average no. of successful runs*	Median percent correct†	Median d′ for length discrim.†	Median d′ for spectrum discrim.†
		Nonparametric model, TOME			
English	18 of 19	5.4 ± 2.9	83.0 (91.3)	3.78 (3.83)	2.70 (3.06)
Japanese	10 of 10	5.5 ± 1.6	85.2 (91.2)	4.05 (3.98)	3.11 (3.25)

*Speakers with successful runs, with 10 runs per speaker.
†Percent-correct and d′ values are medians across speakers of the average over successful runs within a speaker. Parenthetical values show supervised training results.

Vallabha et al. (2007): Statistical Learning of Phonemic Contrasts

Success?

The results of the OME and TOME training suggest that speech categories contain enough acoustic structure for explicit unsupervised acquisition of both the number of categories and their respective distributions.

Table 1. Learning performance for successful runs

Language	No. of speakers w/successful runs*	Average no. of successful runs*	Median percent correct†	Median d′ for length discrim.†	Median d′ for spectrum discrim.†
		Parametric model, OME			
English	18 of 19	7.7 ± 2.8	92.7 (93.4)	3.91 (3.90)	3.19 (3.22)
Japanese	10 of 10	7.9 ± 3.0	91.1 (91.9)	4.09 (4.09)	3.32 (3.30)
		Nonparametric model, TOME			
English	18 of 19	5.4 ± 2.9	83.0 (91.3)	3.78 (3.83)	2.70 (3.06)
Japanese	10 of 10	5.5 ± 1.6	85.2 (91.2)	4.05 (3.98)	3.11 (3.25)

*Speakers with successful runs, with 10 runs per speaker.
†Percent-correct and d′ values are medians across speakers of the average over successful runs within a speaker. Parenthetical values show supervised training results.

Discussion: Do we want perfect performance in these models, or do we want flawed performance since infants must go through stages of learning?

Relation to vowel category acquisition

The success of the OME algorithm has several implications for theories of vowel acquisition. The current results show that infant-directed speech in English and Japanese contains enough acoustic structure to bootstrap the acquisition of (at least some) vowel categories. In tandem with other work (8, 9, 18, 22–25), this provides a mechanistic underpinning and feasibility assessment of the proposal that, for at least some speech sounds, infants initially have a homogeneous auditory space that develops category structure although experience.

A note on the implementational level

Both OME and TOME represent categories by dedicating a single category unit to each one. This fact should not be viewed as a claim about neural implementation, because it is unlikely that there are neurons in the brain dedicated to individual categories. It is more likely that category representations should be sought in the collective activity of neural populations, and that this distributed activity exhibits behavior akin to localist representations (28, 46).

Vallabha et al. (2007): Statistical Learning of Phonemic Contrasts

Now, back to speech acquisition - domain-specific vs. domain-general?

Finally, the OME algorithm has implications for the larger debate about the nature of speech acquisition, namely, whether it is guided primarily by innate, domain-specific constraints that unfold over time or by the statistics of the speech stimuli. The present work is based on a position between these two extremes. Although it incorporates an innate bias for Gaussian-distributed categories, such a bias appears to be justified for stop consonants (47) as well as vowel spectra (29). Moreover, such a bias is very generic and unlikely to be relevant only to speech (31, 48); the gradient descent algorithm that underlies OME is also very general. The use of relatively domain-general principles together with domain-specific input statistics has been shown to account for phenomena such as stage-like development (49) and quasi-regularity (50), and the success of the OME algorithm suggests that such an approach may prove fruitful in the domain of speech category acquisition. Within the present approach, an issue for further research is whether something approximating the bias favoring Gaussian or unimodal category structure used in the OME version of the model can be incorporated in a future version of the biologically more realistic TOME model, while still preserving TOME's ability to model non-Gaussian distributions should the input deviate from the Gaussian constraint.

Gaussian distribution assumption = domain-general bias?

How important is biological plausibility in the learning algorithm?