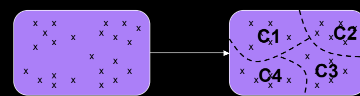
Psych 215L:
Language Acquisition

Lecture 5

Speech Perception II

Speech Perception: Computational Problem

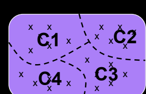
Divide sounds into contrastive categories



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



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

Testbed: Category emergence for English & Japanese vowel contrasts

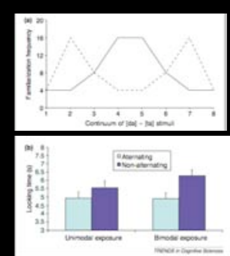
Trajectory: 6-month-olds have language-specific vowel distinctions

Statistical learning

"…infants exposed to a stimulus continuum with a bimodal distribution were better able to distinguish the endpoints of the continuum, as compared with infants who were exposed to a unimodal distribution…"

Maye et al. 2002

on 6 and 8-month-old infants



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

Motherese

"…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…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

Sounds: Vowel contrasts in English and Japanese

English contrasts: contrast in quality (tense vs. lax) and a bit in duration

/ɪ/ vs. /i/ /ɛ/ vs. /e/

"ih" "ee" "eh" "ey"

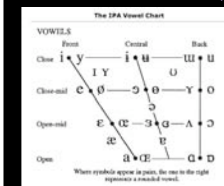
Japanese contrasts: contrast almost solely in duration (short vs. long)

/i/ vs. /i:/ /e/ vs. /e:/

"ee" "eeee" "ey" "eeey"

"These categories occur in the same general region of a multidimensional vowel space defined by formant frequency and duration, but have different phonetic realizations n the two languages. For example, the English /ɪ/ and /i/ differ in both formant frequency and duration, whereas the Japanese /i/-/i:/ differ almost solely in duration."

Formants



High F1

Low F1

High F2

Low 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): Learning Algorithm

"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."

"The model simultaneously estimated 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…It is 'parametric' in that it treats the distribution of speech sounds in a category as an *n*-dimensional Gaussian, and estimated the sufficient statistics of each distribution. We later present a nonparametric variant…"

Incremental Expectation Maximization

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

There are unknown (latent) variables in the model that generate the observable data in the input (e.g. where the vowel category centers are in acoustic space).

Algorithm cycles between doing an expectation step and a maximization step

Expectation: computes the expectation of the likelihood of the actual data encountered by using the current values of the latent variables

Maximization: computes the maximum likelihood estimates using the expected likelihood found in the expectation step

Example EM problem

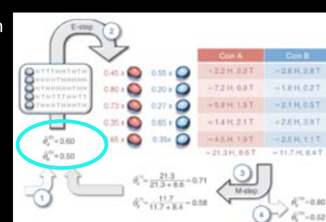
Problem: determine bias in two coins, A and B

Bias: (θ_A, θ_B)

Start with an initial bias guess:

θ_A = 0.6 (60% heads)

θ_B = 0.5 (50% heads)

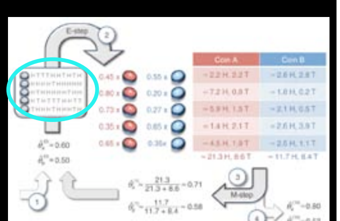


Example EM problem

Problem: determine bias in two coins, A and B

Bias: (θ_A, θ_B)

Have data set:
5 sets of 10 coin tosses, but don't know which coin was tossed for each set



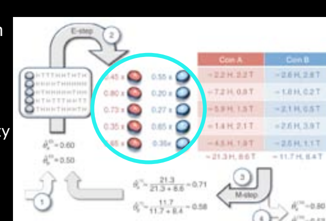
Example EM problem

Problem: determine bias in two coins, A and B

Bias: (θ_A, θ_B)

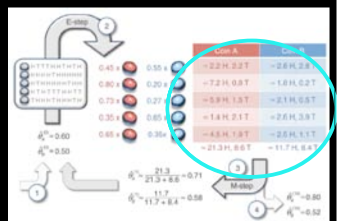
In the E-step, a probability distribution over possible completions is computed using the current parameters.

Ex: HTTTHHTHTH
Normalized prob that A generated this = .45
Normalized prob that B generated this = .55



Example EM problem

Problem: determine bias in two coins, A and B
Bias: (θ_A, θ_B)

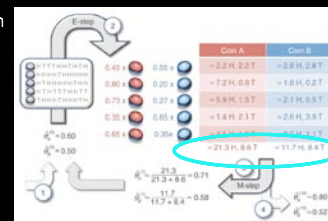


The counts shown in the table are the expected numbers of heads and tails according to this distribution.

Ex: HTTTHHTHTH
A = 0.45 of heads
5 heads = 5 *.45 = 2.25H (and 2.25T)
B = 0.55 of heads
5 heads = 5 * .55 = 2.75H (and 2.75T)

Example EM problem

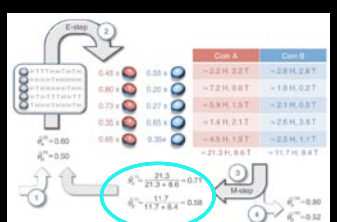
Problem: determine bias in two coins, A and B
Bias: (θ_A, θ_B)



Sum to get expected distribution of heads and tails for each coin.

Example EM problem

Problem: determine bias in two coins, A and B
Bias: (θ_A, θ_B)

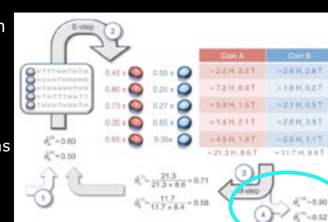


In the M-step, new parameters are determined using the current completions.

Ex: New θ_A estimate
= 21.3H/ (21.3H + 8.6T)
= 0.71 (was 0.6 before)

Example EM problem

Problem: determine bias in two coins, A and B
Bias: (θ_A, θ_B)

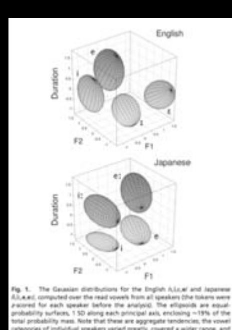


4. After several repetitions of the E-step and M-step, the algorithm converges.

θ_A = 0.8
θ_B = 0.52

Vallabha et al. (2007): Algorithm & Data

"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 goal is to recover, given just the sequence of vowels tokens, the number of Gaussians, the parameters of each Gaussian and the respective mixing probabilities."

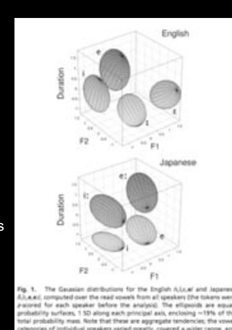


Vallabha et al. (2007): Algorithm & Data

Parameters (3): locations of the first and second formants (F1 and F2) and the duration of the vowel categores.

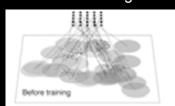
Gaussians derived separately for each vowel category.

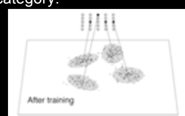
Four Gaussians for each speaker became training distribution and were used to generate 2,000 data points for each vowel category, for a total of 8,000 training tokens per speaker.



Vallabha et al. (2007): Algorithm & Data

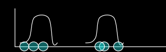
"The algorithm first calculates the 'responsibility' of each category for the token…Each run of the algorithm is initialized with 1,000 equally probable Gaussian categories with randomly initialized means…On each trial, one token is randomly drawn, with replacement, from the set of 8,000 for that speaker….Next, it updates the [current category parameters], 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…and reduces the mixing probabilities of all others…enforces the constraint that each data point should belong to only one category."





Vallabha et al. (2007): Algorithm

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): Algorithm & Data

"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."

Vallabha et al. (2007): Testing the Model

Training: 50,000 data points to train on
Testing: 2,000 data points tested on

"Each test point was classified with the category that had the greatest likelihood for that point. The [test run] was considered 'successful' if 95% of the test points were classified into four categories. For evaluation purposes, the categories were also assigned labels…[measures] the percent-correct…the length *d'* (sensitivity in distinguishing /ɪ,ɛ/ from /i,e/ in English speech), and the spectrum *d'* (sensitivity distinguishing /i, i:/ and /e, e:/ in Japanese speech, /ɪ,i/ from /ɛ,e/ in English speech."

Vallabha et al. (2007): Evaluating the Model

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): Inter-Speaker Variation & Categorization

"…there is also considerable variability between speakers of the same language…Can the productions of an individual speaker support the discovery of speaker-general but still language-specific structure?"

"…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 [CLG 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"

Vallabha et al. (2007): Inter-Speaker Variation & Categorization

"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)…therefore clear that the productions of individual speakers contain substantial language-specific information. Even so, the superiority of the same-speaker test performance…over the WLG suggests that robust acquisition of vowel categories depends on exposure to multiple speakers"

Vallabha et al. (2007): A Different Model

"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."

"…moving closer to a possible neurobiological implementation…distribution of each category is represented nonparametrically…scheme has a natural 'neural network' interpretation…resulting algorithm has similarities to connectionist models of categorization….refer to it as 'Topographic OME'"

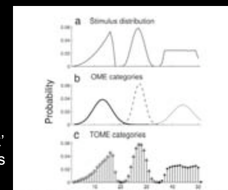


Fig. 3. A comparison of representations learned by OME and TOME for a 1D input space (see SI). (a) The distribution from which the input stimuli were drawn. (b) The three categories discovered by OME. Each line shows the conditional probability of one category, multiplied by its mixing probability. (c) 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.

Vallabha et al. (2007): OME vs. TOME model

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.

TOME isn't as good as OME…but which one matches children's behavior more?

Vallabha et al. (2007): Implications

"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…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 through experience."

A note on the implementational level: "Both [models] represent categories by dedicating a single category unit to each one…more likely that category representations should be sought in the collective activity of neural populations..."

Vallabha et al. (2007): Domain-general vs. Domain-specific

"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 justified for stop consonants as well as vowel spectra. Moreover this bias is very generic and unlikely to be relevant only to speech…use of relatively domain-general principles together with domain-specific input statistics has been show to account for [many] phenomena…the success of the OME algorithm suggests that such an approach may prove fruitful in the domain of speech category acquisition."

Future work: "…whether something approximating the bias…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."