A Role for the Developing Lexicon in Phonetic Category Acquisition

Feldman, Griffiths, Goldwater, and Morgan

Psychological Review 2013

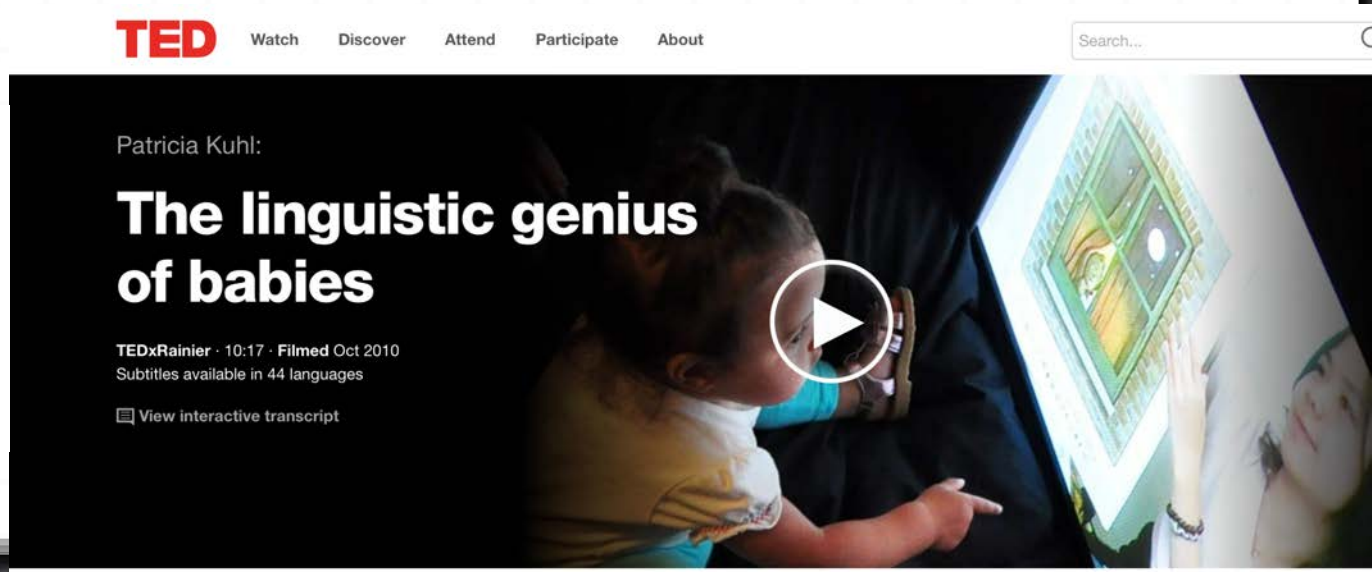
Presentation by Georgina Lean

Summary

- Use a bayesian model to illustrate how feedback from segmented words might constrain phonetic category learning by providing information about which sounds occur together in words.
- Simulations demonstrate that word-level information can successfully disambiguate overlapping English vowel categories.
- Provide a framework for incorporating top-down constraints into models of category learning.

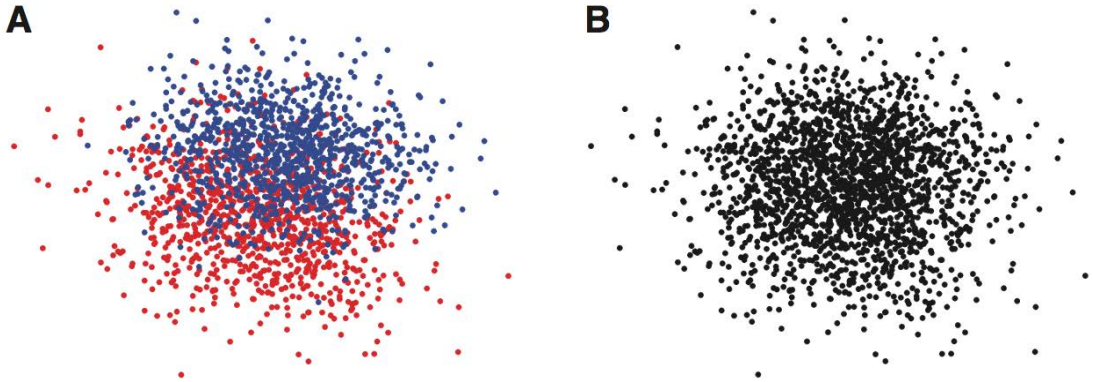
Statistical learning theories

- Infants acquire each layer of structure by observing statistical dependencies in their input.
- Infants show robust sensitivity to statistical patterns.



Statistical learning

- Statistical learning is a domain general strategy for discovering structure in the world.
- Distributional learning has been proposed as a statistical learning mechanism for phonetic category acquisition.
  - Most effective when the categories have very little overlap.



Distributional Learning

- Overcome the overlapping categories problem by using feedback from higher levels of structure to constrain category acquisition.
  - Specifically use a developing lexicon for feedback
- Focus on developing linguistic categories (specifically modeling linguistic categories)

- First we introduce the idea of modeling category learning as density estimation and show how distributional learning can be viewed in this framework.
- Then we show that distributional learning can be challenging when categories have a high degree of overlap
- Conclude by showing that qualitative behavior of our lexical-distributional model mirrors patterns from experiments on sound category learning, suggesting that people behave as interactive learners

Distributional Learning

- What is it?
  - Density estimation: learning a category requires estimating a probability distribution over the items that belong to the category.
  - Categorization becomes probablisitic inference.
- Linguistic Example:
  - Phonetic category acquisition

Computational models that have been used to investigate the utility of distributional learning for phonetic category acquisition

- Gaussian mixture model:
  - Mixture models assume that there are several categories and that each of the observed data points was generated from one of these categories.

Mixture Model

- Inferring a probability distribution p(x|c) when z_i is known
  - The learner knows which stimuli belong to the category

$$\mu_c = \frac{1}{n_{z_i=c}} \sum x_i$$

$$\sum_c = \frac{1}{n_{z_i=c}} \sum (x_i - \mu)(x_i - \mu)^T, \quad (1)$$

Mixture Model

- Inferring z_i when the probability distribution $p(x|c)$ and frequency $p(c)$ is known

$$p(c|x) = \frac{p(x|c)p(c)}{\sum_{c'=1}^{C} p(x|c')p(c')}, \quad (2)$$

Problem

- However, language learners acquiring phonetic categories do not have either of these values.

- Solution?
  - Expectation maximization: provides a principled solution to these types of problems by searching for the parameters and category labels that maximize the probability of the data

Inferring the Number of Categories

- 2 options:
  - Gradient descent model (McMurray et al (2009) and Vallabha et al (2007)): prune excess phonetic categories that are not needed
    - Focus on a voicing contrast in consonants (McMurray) and vowels (Vallabha)
    - Drawback: the model cannot find a set of globally optimal category parameters, only converges to a locally optimal solution
  - Dirichlet process (Ferguson 1973): infinite mixture models
    - Uses Gibbs sampling algorithm
    - Allows for a direct comparison between the distributional and lexical-distributional learning strategies.

Simulation 1: The Problem of Overlapping Categories

- Two sounds might be assigned to the same category for a distributional learner if they are too similar
- To explore this challenge, we test the ability of distributional learning models to recover the vowel categories

Table 1
Normalized Empirical Probabilities of Each Vowel Computed From the Phonematized CHILDES Parental Frequency Count

Vowel	Empirical probability	
	In word tokens	In word types
/æ/	.080	.068
/ɑ/	.125	.105
/ɔ/	.038	.035
/ɛ/	.067	.075
/e/	.039	.048
/ɝ/	.035	.083
/ɪ/	.177	.169
/i/	.077	.099
/o/	.061	.041
/ʊ/	.041	.019
/ʌ/	.176	.229
/u/	.083	.030

Note. CHILDES = Child Language Data Exchange System.

Simulation 1: The Problem of Overlapping Categories

- A successful model is based on
  - its ability to recover the correct number of categories
  - Its ability to identify which sounds from the corpus are in each category

Table 2
Phonetic Categorization Scores From the Infinite Mixture Model (IMM) and Gradient Descent Algorithm (GD) in Simulation 1

	All speakers		Men only	
Variable	IMM	GD	IMM	GD
Number of categories	10	6	11	8
F-score	0.453	0.480	0.699	0.727
Variation of information	3.195	2.677	1.678	1.440

Note. The true number of phonetic categories is 12.

Simulation 1: The Problem of Overlapping Categories

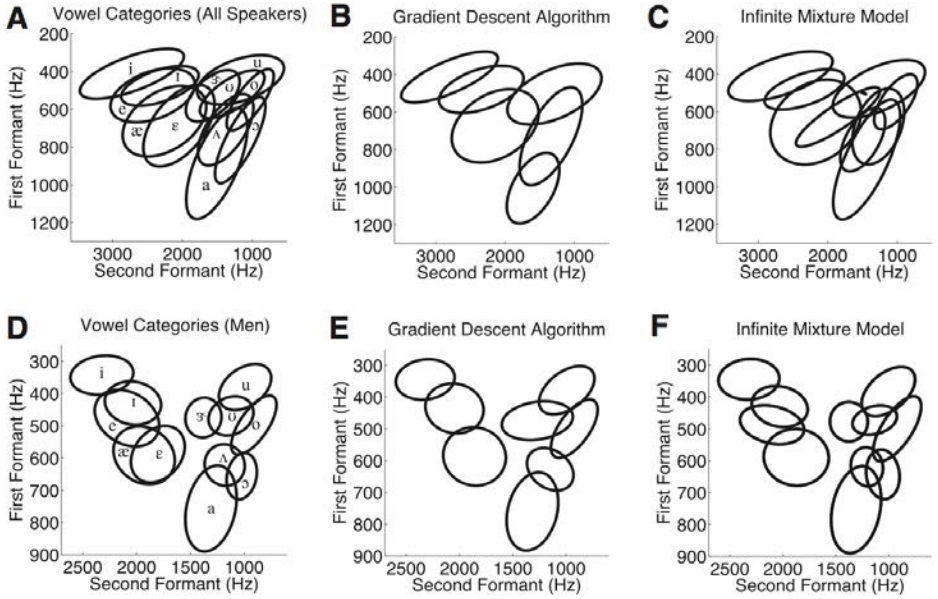


Figure 2. Results from Simulation 1. Ellipses delimit the area corresponding to 90% of vowel tokens corresponding to vowel categories for all speakers from Hillenbrand et al. (1995) that were used to generate the first corpus (A) and the resulting categories found by the gradient descent algorithm (B) and the infinite mixture model (C); and vowel categories for men only from Hillenbrand et al. that were used to generate the second corpus (D) and the resulting categories found by the gradient descent algorithm (E) and the infinite mixture model (F). Figure 2A adapted with permission from "Acoustic Characteristics of American English Vowels," by J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, 1995, *Journal of the Acoustical Society of America, 97,* p. 3103. Copyright 1995 by Acoustical Society of America.

Incorporating Lexical Constraints

- Infants use their sensitivity to transitional probabilities to begin learning potential word forms from their developing lexicon
- Interaction between sound and word learning is not present in distributional learning theories.
- Words= acoustic tokens in the corpus
- Lexical items= categories that represent groupings of acoustic tokens
  - Words are categorized into lexical items

The new Lexical-Distributional Model

- Distributional model's hypotheses consist of sets of phonetic categories
- Lexical distributional model's hypotheses are combinations of sets of phonetic categories and sets of lexical items
  - Learners optimize lexicon to best explain the word tokens in the corpus

Simulation 2-4

- Simulation 2: illustrates the model's basic behavior
  - Lexical items only consists of vowels
- Simulation 3: tests performance on a lexicon of English words from child-directed speech
- Simulation 4: Speaker variability is reduced

Simulation 2: Lexical-Distributional Learning of English Vowels

Table 3

Phonetic Categorization Scores for the Lexical-Distributional Model (L-D), Infinite Mixture Model (IMM), and Gradient Descent Algorithm (GD) in Simulation 2, Averaged Across All 10 Corpora

Variable	L-D	IMM	GD
Number of categories	11.9	8	5.5
F-score	0.919	0.519	0.545
Variation of information	0.671	2.762	2.426

Note. The true number of phonetic categories is 12.

Table 4

Lexical Categorization Scores for the Lexical-Distributional Model (L-D) and Baseline Model in Simulation 2, Averaged Across All 10 Corpora

Variable	L-D	Baseline
F-score	0.799/0.854	0.523
Variation of information	1.263/0.921	1.853

Note. The first number evaluates performance by treating each cluster as separate, regardless of phonological form, and the second number treats all clusters with identical phonological forms as constituting a single lexical item. The mean number of lexical items recovered is not shown, as the target number of lexical items differed across the 10 corpora.

Simulation 3: Information Contained in the English Lexicon

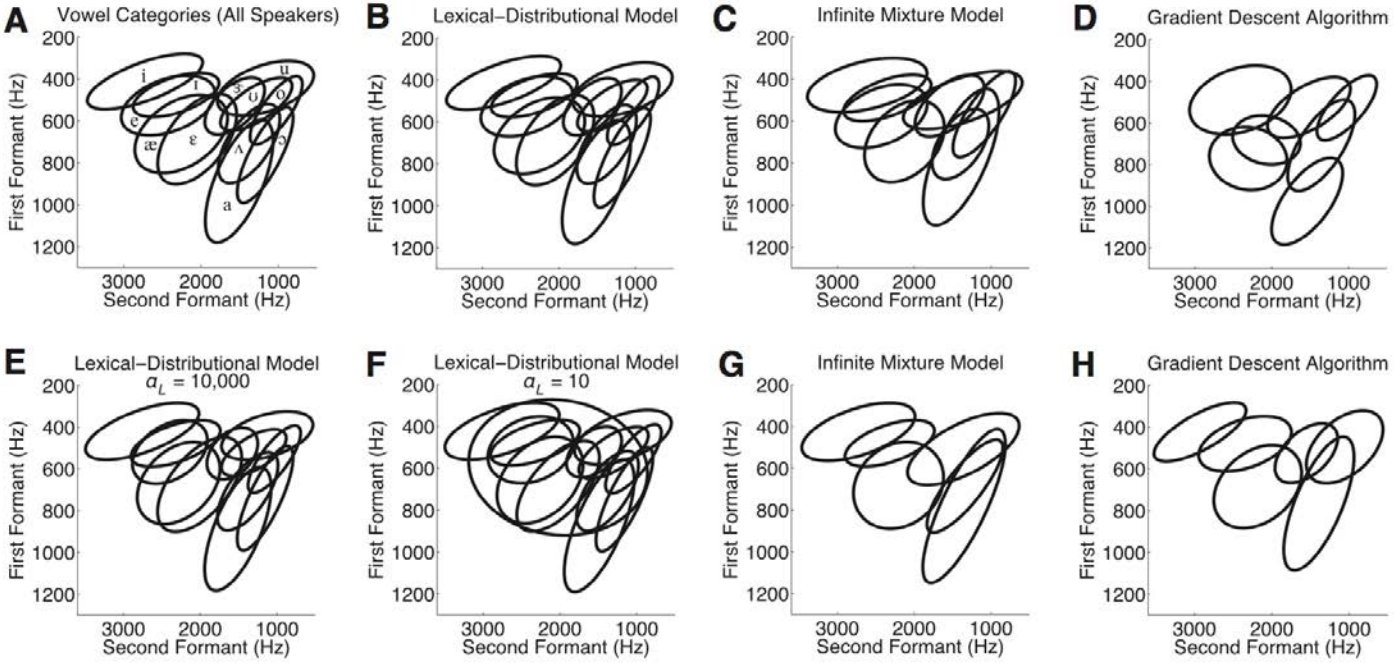


Figure 3. Results of Simulations 2 and 3. Ellipses delimit the area corresponding to 90% of vowel tokens for Gaussian categories computed from men's, women's, and children's production data in Hillenbrand et al. (1995; A), recovered in Simulation 2 by the lexical-distributional model (B), the infinite mixture model (C), and the gradient descent algorithm (D), and recovered in Simulation 3 by the lexical-distributional model with α_L = 10,000 (E), the lexical-distributional model with α_L = 10 (F), the infinite mixture model (G), and the gradient descent algorithm (H). Figure 3A adapted with permission from "Acoustic Characteristics of American English Vowels," by J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, 1995, *Journal of the Acoustical Society of America, 97,* p. 3103. Copyright 1995 by Acoustical Society of America.

Simulation 3: Information Contained in the English Lexicon

Table 5

Phonetic Categorization Scores for the Lexical-Distributional Model (L-D), Infinite Mixture Model (IMM), and Gradient Descent Algorithm (GD) in Simulation 3

Variable	L-D					IMM	GD
	$\alpha_L = 1$	$\alpha_L = 10$	$\alpha_L = 100$	$\alpha_L = 1{,}000$	$\alpha_L = 10{,}000$		
Number of categories	14	13	13	12	12	6	6
F-score	0.719	0.756	0.755	0.745	0.709	0.448	0.483
Variation of information	2.085	1.803	1.790	1.765	1.959	2.949	2.699

Note. The true number of phonetic categories is 12 for each corpus.

Table 6

Lexical Categorization Scores for the Lexical-Distributional Model (L-D) and Baseline Model in Simulation 3

Variable	L-D					Baseline
	$\alpha_L = 1$	$\alpha_L = 10$	$\alpha_L = 100$	$\alpha_L = 1{,}000$	$\alpha_L = 10{,}000$	
Number of categories	900/899	926/920	958/934	1,164/989	1,602/1,086	852
F-score	0.908/0.924	0.919/0.933	0.901/0.918	0.830/0.919	0.610/0.854	0.840
Variation of information	0.368/0.340	0.321/0.290	0.412/0.324	0.705/0.338	1.389/0.538	0.459

Note. The first number treats each cluster as separate, regardless of phonological form, and the second number treats all clusters with identical phonological forms as belonging to a single lexical item. The true number of lexical items is 1,019.

Simulation 4: Reduced Speaker Variability

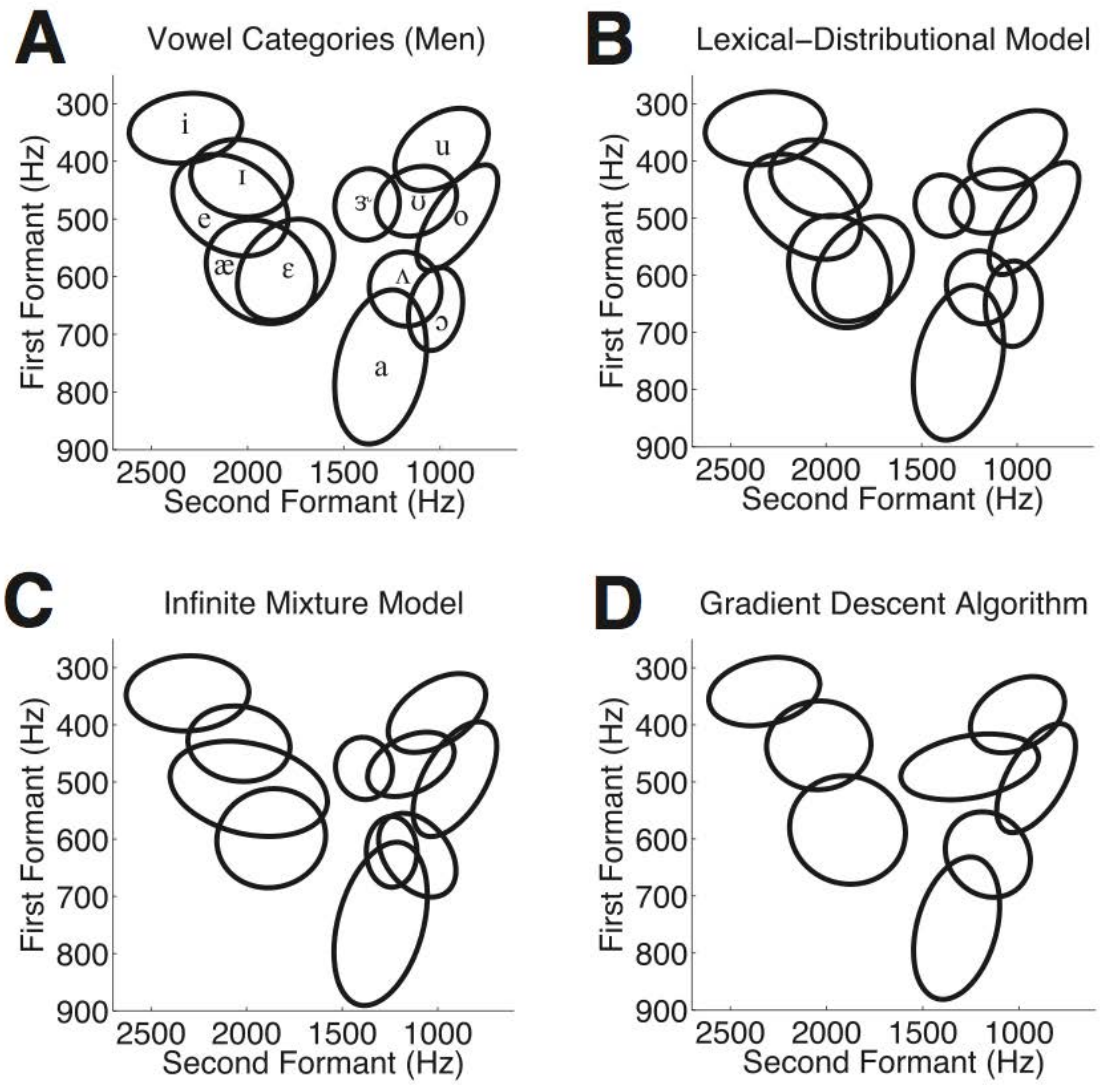


Figure 5. Results of Simulation 4. Ellipses delimit the area corresponding to 90% of vowel tokens for Gaussian categories computed from men's production data in Hillenbrand et al. (1995; A) and recovered in Simulation 4 by the lexical-distributional model with $\alpha_L = 10{,}000$ (B), the infinite mixture model (C), and the gradient descent algorithm (D).

General Discussion