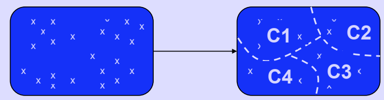
Psych 215L: Language Acquisition

Lecture 4
Speech Perception

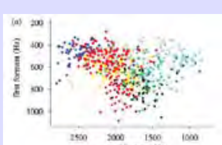
Speech Perception: Computational Problem

Divide sounds into contrastive categories



Speech Perception: Computational Problem

Remember that real world data are actually much harder than this… (from Swingley 2009)



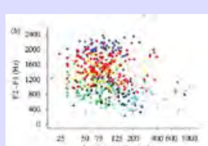


Figure 2. (a) First and second formants in about 700 vowels of one mother's speech to her infant. Each colour/shape combination indicates a different vowel. (b) Second formant minus first formant plotted against raw duration, for vowels of one mother's speech to her infant. Each colour/shape combination indicates a different vowel.

Order of acquisition?

"It is often implicitly assumed…infants first learning about the phonetic categories in their language and subsequently using those categories to help them map word tokens onto lexical items. However, infants begin to segment words from fluent speech as early as 6 months (Bortfeld, Morgan, Golinkoff, & Rathbun, 2005) and this skill continues to develop over the next several months (Jusczyk & Aslin, 1995; Jusczyk, Houston, & Newsome, 1999). Discrimination of non-native speech sound contrasts declines during the same time period, between 6 and 12 months (Werker & Tees, 1984). This suggests an alternative learning trajectory in which infants simultaneously learn to categorize both speech sounds and words, potentially allowing the two learning processes to interact."

What we know about infants

Maye, Werker, & Gerken 2002: infants show sensitivity to statistical distribution of acoustic data points

Mixture of Gaussians (MoGs) modeling approaches building on this ability:

- Boer and Kuhl 2003: Expectation Maximization (EM) algorithm (Dempster, Laird, & Rubin 1977) to learn the locations of three vowel categories from formant data.

- Toscano & McMurray 2008, Vallabha et al. 2007: EM to learn multiple dimensions for both consonant and vowel data

- McMurray, Aslin, and Toscano 2009: gradient descent algorithm similar to EM to learn a stop consonant voicing contrast.

Feldman, Griffiths, & Morgan 2009

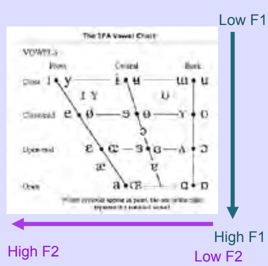
Use MoG approach within a non-parametric Bayesian framework.

Why? Allows extension of the model to the word level (instead of only including the phonemic category level).

Phonetic dimensions used to describe input data:
- formant values (F1, F2)
- voice onset time

Words: Sequences of phonetic values, where each phoneme corresponds to a discrete set of phonetic values

Formants



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.

Sample Input

Input Stream: ADAABDABDC

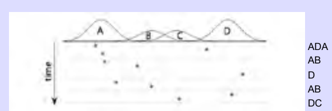


Figure 1: A fragment of a corpus presented to the model. Asterisks represent speech sounds, and lines represent word boundaries. The model does not know which categories generated the speech sounds, and needs to recover categories A, B, C, and D from the data.

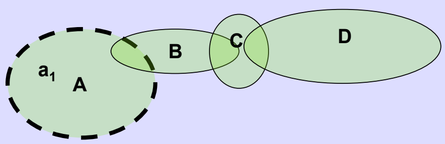
Learner's job is to recover
(1) A, B, C, D distributions
(2) words ADA, AB, D, AB, and DC

Distributional Model

Model goal: learn the phoneme inventory (ignore information about words and word boundaries)

Phoneme inventory = {A, B, C, D, …}

Sounds are assumed to be produced by the speaker selecting a category from the phoneme inventory and then sampling a phonetic value from the Gaussian associated with that category.



Distributional Model

Learner inference process: Dirichlet process (Ferguson 1973)

Properties of the Dirichlet process:

(1) Allow learner to consider potentially infinite number of categories

(2) Bias (α) determines how strong preference for fewer categories is

Learner begins with a prior that is very weak (so real data will overshadow it and learner will adjust beliefs accordingly).

Learner goal: Recover the sequence of categories that produced the observed sounds (acoustic values).

Distributional Model

Speech sounds are initially given random category assignments:

Initial Assignment: D D A B E F C A B A

Distributional Model

Assignments updated after each sweep through the corpus, based on the other assignments currently made.

Initial Assignment: D D A B E F C A B A
Assignment 1: A D A B E F B B C F

Probability of assignment of sound in position j of word i (w_{ij}) to category c:

$$p(c|w_{ij}) \propto p(w_{ij}|c)p(c)$$

Distributional Model

Assignments updated after each sweep through the corpus, based on the other assignments currently made.

Initial Assignment: D D A B E F C A B A
Assignment 1: A D A B E F B B C F

Prior p(c) is given by the Dirichlet process below, where categories that already have many sounds (# of sounds = n_c) are more likely to get a new sound assigned to them, though there is some probability α that a new category is formed:

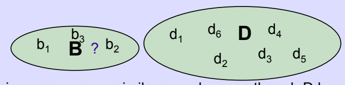
$$p(c) = \begin{cases} \frac{n_c}{\sum_c n_c + \alpha} & \text{for existing categories} \\ \frac{\alpha}{\sum_c n_c + \alpha} & \text{for a new category} \end{cases}$$

Distributional Model

Assignments updated after each sweep through the corpus, based on the other assignments currently made.

Initial Assignment: D D A B E F C A B A
Assignment 1: A D A B E F B B C F

The likelihood $p(w_{ij} \mid c)$ takes into account the other sounds already assigned to that category. Categories where sounds are very different from the current sound are less likely.



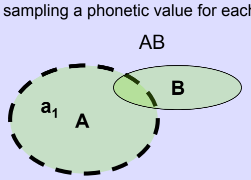
B more likely since many more similar sounds, even though D has more sounds total.

Lexical-Distributional Model

Model goal: learn the phoneme inventory and the lexicon, where lexical items are sequences of phonemes

Phoneme inventory = {A, B, C, D, …}
Lexicon = {ADA, AB, D, DC, …}

The corpus is generated by a speaker selecting a word from the lexicon, and then sampling a phonetic value for each phoneme in that word.

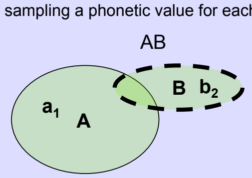


Lexical-Distributional Model

Model goal: learn the phoneme inventory and the lexicon, where lexical items are sequences of phonemes

Phoneme inventory = {A, B, C, D, …}
Lexicon = {ADA, AB, D, DC, …}

The corpus is generated by a speaker selecting a word from the lexicon, and then sampling a phonetic value for each phoneme in that word.



Lexical-Distributional Model

Learner inference process: Dirichlet process (Ferguson 1973) over phonemes and lexicon items

Properties of the Dirichlet process:

(1) Allow learner to consider potentially infinite number of categories (phonemes or lexicon items)

(2) Bias (α) determines how strong preference for fewer categories is (phonemes: fewer categories)

(lexicon: fewer items, shorter items)

Learner goal: Recover the sequence of categories that produced the observed sounds (acoustic values) and the sequence of words produced (by identifying the lexicon items that produced them).

Lexical-Distributional Model

Words initially hypothesized and assigned to random lexical items, and speech sounds in those words are initially given random category assignments:

Initial Assignment: DD ABE FC A BA

Lexical Items: {'DD', 'ABE', 'FC', 'A', 'BA'}

Lexical-Distributional Model

Assignments updated after each sweep through the corpus, based on the other assignments currently made.

Initial Assignment: DD ABE F C A BA
Assignment 1: AD ABE F B B CF

Lexical items = {'AD', 'ABE', 'F' 'B', 'CF'}

Probability of assignment of word$_i$ to lexical item k:

$$p(k|w_i) \propto p(w_i|k)p(k)$$

Lexical-Distributional Model

Assignments updated after each sweep through the corpus, based on the other assignments currently made.

Initial Assignment: DD ABE F C A BA
Assignment 1: AD ABE F B B CF

Lexical items = {'AD', 'ABE', 'F' 'B', 'CF'}

Prior p(k) is given by the Dirichlet process below, where lexical items that already have many tokens (# of tokens = n_k) are more likely to get a new word assigned to them, though there is some probability β that a new lexical item is formed:

$$p(k) = \begin{cases} \frac{n_k}{\sum_i n_i + \beta} & \text{for existing categories} \\ \frac{\beta}{\sum_i n_i + \beta} & \text{for a new category} \end{cases}$$

B freq = 2
AD freq = 1

Lexical-Distributional Model

Assignments updated after each sweep through the corpus, based on the other assignments currently made.

Initial Assignment:	DD ABE F C A BA	Lexical items = {'AD', 'ABE', 'F' 'B', 'CF'}
Assignment 1:	AD ABE F B B CF	

The likelihood $p(w_i \mid k)$ takes into account the categories required to produce the lexical item, with w_{ij} being the category in position j of word i and c_{kj} being the category in position j of lexical item k.

$$p(w_i|k) = \prod_j p(w_{ij}|c_{kj})$$

P(B | 'B') = prob(B | position 1 of 'B')

Lexical-Distributional Model

Assignments updated after each sweep through the corpus, based on the other assignments currently made.

Initial Assignment:	DD ABE F C A BA	Lexical items = {'AD', 'ABE', 'F' 'B', 'CF'}
Assignment 1:	AD ABE F B B CF	

Part 2: Probability of category c to position j in lexical item k:

$$p(c|w_{\{k\}j}) \propto p(w_{\{k\}j}|c)p(c)$$

P(B | position 1 of 'B') = ?

Lexical-Distributional Model

Assignments updated after each sweep through the corpus, based on the other assignments currently made.

Initial Assignment:	DD ABE F C A BA	Lexical items = {'AD', 'ABE', 'F' 'B', 'CF'}
Assignment 1:	AD ABE F B B CF	

Part 2: Prior p(c) is same as before (based on number of sounds currently in that category)

$$p(c) = \begin{cases} \frac{n_c}{\sum_c n_c + \alpha} & \text{for existing categories} \\ \frac{\alpha}{\sum_c n_c + \alpha} & \text{for a new category} \end{cases}$$

Lexical-Distributional Model

Assignments updated after each sweep through the corpus, based on the other assignments currently made.

Initial Assignment:	DD ABE F C A BA	Lexical items = {'AD', 'ABE', 'F' 'B', 'CF'}
Assignment 1:	AD ABE F B B CF	

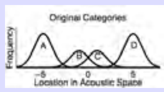
The likelihood $p(w_{\{k\}j} \mid c)$ takes into account all phonetic values associated with all words assigned to lexical item k. Categories where sounds are very different from the current sounds associated with words assigned to the lexical item are less likely.

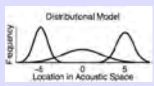
'B' = {B} = {b1, a1, b2, b3}

p(position 1 of 'B' | all known B values) = ?

Testing the Models

For distributional model:
1200 acoustic values sampled from these distributions:
400 A, 200 B, 200 C, 400 D

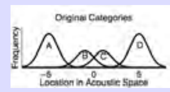




B and C interpreted as a single category

Testing the Models

For lexical distributional model:
1200 acoustic values sampled from these distributions:
400 A, 200 B, 200 C, 400 D



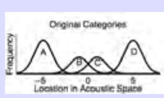
+ a corpus of fluent speech made up of lexical items

Uninformative (B/C) corpus: AB, AC, DB, DC, ADA, D
Why uninformative? Easier to encode this lexicon as AX, DX, ADA, D

Input stream: each of these 6 tokens repeated 100 times

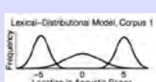
Testing the Models

For lexical distributional model:
1200 acoustic values sampled from these distributions:
400 A, 200 B, 200 C, 400 D



+ a corpus of fluent speech made up of lexical items

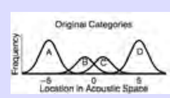
Uninformative (B/C) corpus:
B and C (unsurprisingly) are merged



(Upshot: Minimal pairs are harmful to phonemic category learning)

Testing the Models

For lexical distributional model:
1200 acoustic values sampled from these distributions:
400 A, 200 B, 200 C, 400 D



+ a corpus of fluent speech made up of lexical items

Informative (B/C) corpus: AB, DC, ADA, D
Why informative? Can't encode this lexicon any more compactly

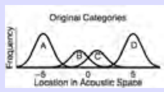
Input stream: 200 AB, 200 DC, 100 ADA, 100 D

Testing the Models

For lexical distributional model:

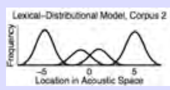
1200 acoustic values sampled from these distributions:

400 A, 200 B, 200 C, 400 D



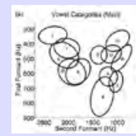
+ a corpus of fluent speech made up of lexical items

Informative (B/C) corpus: Now B and C are found as separate (small acoustic differences viewed as relevant)

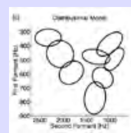


Testing the Models

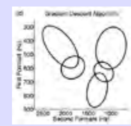
Distributional models on men's vowel data



Distributional model merges many categories together.

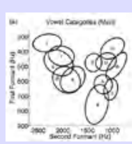


The gradient descent algorithm used by Vallabha et al. 2007 has the same problem.

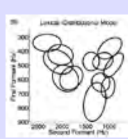


Testing the Models

Lexical-distributional model on men's vowel data: includes made-up corpus of 5000 word tokens (presumably with no minimal pairs)

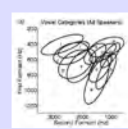


Lexical-distributional model makes fine distinctions.

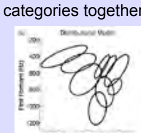


Testing the Models

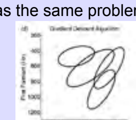
Distributional models on men, women, & children's vowel data: much more overlap in categories



Distributional model merges many categories together.



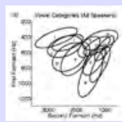
The gradient descent algorithm used by Vallabha et al. 2007 has the same problem.



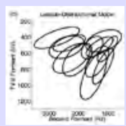
Testing the Models

Lexical-distributional model on men, women, & children's vowel data: much more overlap in categories

+ 5000 hypothetical lexical items (presumably with no minimal pairs)



Lexical-distributional model again makes many fine distinctions.



Accuracy & Completeness Scores

Hit = two sounds correctly placed in same category

False alarm = two sounds incorrectly placed in same category

Miss = two sounds incorrectly placed in different categories

Accuracy = hits/(hits + false alarms)

Completeness = hits/(hits + misses)

		Lexical-Distrib.	Distrib.	Gradient Descent
(a)	Accuracy	0.97	0.63	0.56
	Completeness	0.98	0.93	0.94
(b)	Accuracy	0.99	0.54	0.40
	Completeness	0.99	0.85	0.95

Table 1: Accuracy and completeness scores for learning vowel categories based on productions by (a) men and (b) all speakers. For the Bayesian learners, these were computed at the annealed solutions; for the gradient descent learner, they were based on maximum likelihood category assignments.

Note: Annealing = method of allowing more variability during learning early on (allows a learner to escape local maxima more easily)

Take-away points

"…not wish to suggest that a purely distributional learner cannot acquire phonetic categories. The simulations presented here are instead meant to demonstrate that in a language where phonetic categories have substantial overlap, an interactive system, where learners can use information from words that contain particular speech sounds, can increase the robustness of phonetic category learning."

Take-away points

"The first key assumption is that speech sounds in phonetic categories follow the same Gaussian distribution regardless of phonetic or lexical context. In actual speech data, acoustic characteristics of sounds change in a context-dependent manner due to coarticulation with neighboring sounds (e.g. Hillenbrand, Clark, & Nearey, 2001). A lexical-distributional learner hearing reliable differences between sounds in different words might erroneously assign coarticulatory variants of the same phoneme to different categories, having no other mechanism to deal with context-dependent variability. Such variability may need to be represented explicitly if an interactive learner is to categorize coarticulatory variants together."

Take-away points

"A second assumption concerns the lexicon used in the vowel simulations, which was generated from our model. Generating a lexicon from the model ensured that the learner's expectations about the lexicon matched the structure of the lexicon being learned, and allowed us to examine the influence of lexical information in the best case scenario. However, several aspects of the lexicon, such as the assumption that phonemes in lexical items are selected independently of their neighbors, are unrealistic for natural language. In future work we hope to extend the present results using a lexicon based on child-directed speech."

Elsner, Goldwater, & Eisenstein 2012

Learning from realistic child-directed speech, and learning both consonants and vowels.

(a)	intended:	/ju want wʌn/	/want e kuki/
(b)	surface:	[jə wãʔ wʌn]	[wan ə kuki]
(c)	unsegmented:	[jəwãʔwʌn]	[wanəkuki]
(d)	idealized:	/juwantwʌn/	/wantekuki/

Figure 1: The utterances *you want one? want a cookie?* represented (a) using a canonical phonemic encoding for each word and (b) as they might be pronounced phonetically. Lines (c) and (d) remove the word boundaries (but not utterance boundaries) from (b) and (a), respectively.

Model: "Feldman et al. 2009 use a real-valued representation for vowels (formant values), but assume no variability in consonants, and treat each word token independently. In contrast, our model uses a symbolic representation of sounds, but models variability in all segment types and incorporates a bigram word-level language model."

Elsner, Goldwater, & Eisenstein 2012

"…model that simultaneously learns a lexicon, a bigram language model, and a model of phonetic variation, while using only the noisy surface forms as training data. It is the first model of lexical-phonetic acquisition to include word-level context and to be tested on an infant-directed corpus with realistic phonetic variability…the model recovers lexical items more effectively than a system that assumes no phonetic variability; moreover, the use of word-level context is key to the model's success. Ultimately, we hope to extend the model to jointly infer word boundaries along with lexical-phonetic knowledge, and to work directly from acoustic input."

Experimental support for the lexical-distributional model

Feldman, Griffiths, & Morgan (2011)

- Investigated whether human learners are sensitive to the word context in which a sound is found when identifying phonetic categories

- Adult learners heard nonsense words involving the *ah*-*aw* continuum (F2 formant variation)

Lexicon 1 example: *litah, gutaw*

(Informative for *aw* vs *ah* as separate categories)

Lexicon 2 example: *gutah, gutaw, litah, litaw*

(Uninformative for *aw* vs. *ah* as separate categories)

Stimulus Number	Second Formant (Hz)
1	1517
2	1474
3	1432
4	1391
5	1351
6	1312
7	1274
8	1237

Experimental support for the lexical-distributional model

Feldman, Griffiths, & Morgan (2011)

- Adult participants tested on far contrast (*ta*$_1$ vs. *ta*$_8$), near contrast (*ta*$_3$ vs. *ta*$_6$), and control contrast (*mi* vs. *mu*)

- Learners with lexicons informative for two categories distinguished all the contrasts tested by the second half of testing while learners with uninformative lexicons distinguished only the control contrast. This suggests they can use word context when identifying phonetic categories.

- Caveat: Adults may use information differently than infants who haven't completed word segmentation yet.