Noisy-channel models: II
Lecture 4 (part 1)

Klinton Bicknell & Roger Levy

LSA Institute 2015
U Chicago
A new view of reading

• rich linguistic knowledge and perceptual input combine to identify text

• trade-off between prior probability and visual input: when linguistic context provides more information, need less perceptual input to be confident
A new view of reading

A new bottleneck in reading: gathering visual input

- eye movement decisions (how long to fixate, where to move next) made to get visual input (cf. Legge et al., 1997)

- get more perceptual information about a word by fixating it longer/more

- language system determines the most useful visual input to get
  -> when/where to move the eyes

  - reading as *active*

  - principled solution: optimal control, decision theory
How would this link yield interesting linguistic effects?

trade-off between prior probability and visual input

- efficient language users will make fewer and shorter fixations on words with high probability given context
- efficient language users will be more likely to skip over words with high probability given context

look at quantitative predictions of this account for shape of linguistic effects via a computational model
A computational model

probabilistic inference

efficient eye movement control

realistic environment

Bicknell & Levy (2010, 2012)
A computational model

**probabilistic inference**

- combines visual information with probabilistic language knowledge using methods from computational linguistics

**visual information as weighted finite-state automata (wFSAs)**

```
0: a:a/0.7, b:b/0.3
1: a:a/0.2, b:b/0.8
2: a:a/0.1, b:b/0.9
3: <eps>
```

**probabilistic language knowledge**

- **PCFG:**
  - $S \rightarrow NP \ VP \ / \ 0.9$
  - $S \rightarrow S \ Conj \ S \ / \ 0.05$

- **n-grams:**
  - $p(York|New) = 0.4$
  - $p(Brunswick|New) = 0.01$

**efficient composition**

**efficient inference:** closed under intersection with wFSAs

Bicknell & Levy (2010, 2012)
A computational model

**efficient eye movement control**

- determines efficient eye movement behavior in response to incremental comprehension using machine learning methods
- to maximize reward $R$, a linear function of speed and accuracy
  - first-cut accuracy: log probability of sentence string under model beliefs after reading
- *parameterized* behavior policies control the eyes
- determine parameters that maximize $R$ using reinforcement learning (PEGASUS algorithm, Ng & Jordan, 2000)

*fit model behavior to maximize reward, not fit*

Bicknell & Levy (2010, 2012)
A computational model

realistic environment

• embed model in realistic environment using findings from psychophysics and oculomotor control
• exponentially decreasing visual acuity
• saccade initiation delays
• normally-distributed motor error

Bicknell & Levy (2010, 2012)
A computational model

probabilistic inference

efficient eye movement control

realistic environment

Bicknell & Levy (2010, 2012)
Model simulation results

**methods**

- evaluate model on human effects of word frequency and predictability

- **frequency**: word's overall rate of occurrence in language
  - more frequent: dog
  - less frequent: parsnip

- **predictability**: probability of word in context
  - more predictable: The children went outside to *play* …
  - less predictable: My friend really likes to *play* …

Bicknell & Levy (2012)
Model simulation results

methods

• simulate model reading typical psycholinguistic sentences
• analyze three word-based eye movement measures
• only fit one free parameter of model (visual input rate) to match average human word reading time. (others fit to optimize reading efficiency)

• any match to effect shape falls out of principles of efficient identification and linguistic structure

Bicknell & Levy (2012)
Model simulation results

**frequency**

<table>
<thead>
<tr>
<th></th>
<th>human</th>
<th>model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaze duration (ms)</td>
<td>325</td>
<td>300</td>
</tr>
<tr>
<td>Log frequency</td>
<td>-6-5-4-3-2</td>
<td>-6-5-4-3-2</td>
</tr>
</tbody>
</table>

**predictability**

<table>
<thead>
<tr>
<th></th>
<th>human</th>
<th>model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaze duration (ms)</td>
<td>325</td>
<td>300</td>
</tr>
<tr>
<td>Log predictability</td>
<td>-6-4-2</td>
<td>0-6-4-2</td>
</tr>
</tbody>
</table>

**gaze duration**: total duration of all fixations on word

Bicknell & Levy (2012)
Model simulation results

**Frequency**

- **Refixation probability**: probability of making more than one fixation on the word

**Predictability**

-Bicknell & Levy (2012)-
Model simulation results

**frequency**

<table>
<thead>
<tr>
<th>Skipping probability</th>
<th>Log frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>model</td>
</tr>
</tbody>
</table>

**predictability**

<table>
<thead>
<tr>
<th>Skipping probability</th>
<th>Log predictability</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>model</td>
</tr>
</tbody>
</table>

Skipping probability: probability of not directly fixating word

Bicknell & Levy (2012)
Model simulation results

frequency

predicts shape of all effects

predictability

Bicknell & Levy (2012)
Part One: Summary

- implemented model efficiently controlling eyes to gather perceptual info for text identification
- model well predicts shape of linguistic effects on eye movements, fitting only one parameter to average reading time
- provides motivated reason why linguistic effects appear on eye movements in reading
  - reason is not because language processing operations take substantial time (in fact, they are instant in the model)
Reading

[perceptual input] → probabilistic inference → beliefs about identity of the text → rich probabilistic language knowledge

likelihood → prior

rational action

[comprehension behavior]

principle:
Get more input about words with low prior probability given context [cf. surprisal]
Part Two

Viewing reading as visual information gathering solves the puzzle of regressions.
The puzzle

~10% of saccades move the eyes back (‘regress’) to previous words

Why would it be useful to regress to previous words so often?
A solution

confidence falling

‘From the closet, she pulled out a #acket for the upcoming match’

\[ p(\text{jacket}) = 0.9 \quad \rightarrow \quad p(\text{jacket}) = 0.4 \]

\[ p(\text{racket}) = 0.1 \quad \rightarrow \quad p(\text{racket}) = 0.6 \]

\[ p(\text{packet}) = 0.0 \quad \rightarrow \quad p(\text{packet}) = 0.0 \]

confidence high \quad \rightarrow \quad confidence low

Maybe regressions an effective way to deal with this?

Bicknell & Levy (2010)
A solution

simulations

• compare behavior policies that make regressions to those that don't
• regressive policies are strictly more efficient (both faster and more accurate)

but do humans make regressions when confidence falls?

Bicknell & Levy (2010)
Human regressions

methods

• Dundee corpus: large corpus of eye movements reading newspaper editorials [Kennedy & Pynte, 2005]

• derived broad coverage measure of confidence falling: $\Delta_c$

  • $\Delta_c = \log\text{frequency} - \log\text{predictability}$
    (ask me about the derivation!)

• logistic mixed-effects regression model predicting whether or not the eyes will make a regression to a previous word [statistics]
  [eyes] [statistics]

Bicknell & Levy (2011)
Human regressions

methods

- predictors
  - $\Delta_c$ for current word (word$_n$) & previous word (word$_{n-1}$) for late-triggered regressions
  - variables suggested by other accounts of regressions (frequency, length, landing position, saccade length, fixation duration)
  - predict more regressions for high $\Delta_c$ (on either word)

Bicknell & Levy (2011)
Human regressions

- more regressions when current word or previous word makes confidence fall
- confidence falling one of most reliable predictors

Relevant for today

Bicknell & Levy (2011)
Part Two: Summary

explaining regressions

• efficient reader will update uncertain beliefs about prior material on the basis of new input: confidence falling

• confidence falling & regressions
  • regressions help make reading more efficient
  • confidence falling explains short-range regressions in naturalistic text

• can differentiate types of ‘processing difficulty’
  • words of low prior probability -> more/longer fixations
  • losing confidence in prior material -> regressions
1. Get more input about words with low prior probability given context [cf. surprisal]
2. Regress to previous words if confidence about them falls [cf. EIS]
Summary

This part of today's lecture argued that comprehension is well understood as probabilistic inference on noisy perceptual input.
Conclusion

what this buys us

• part 1: a solution to the problem of local coherences
  • they change beliefs about what readers thought they saw (EIS)
• part 2: a principled reason why reading times on a word should be well described by surprisal
  • need less visual information to be confident in identity of words that have high contextual probability
• part 3: a principled reason why readers should make regressions
  • language often makes readers doubt what they thought they saw
lots more recent work in this area (papers on class website)

- Gibson et al. (2013): evidence for noisy-channel inferences about what sentences mean
  - "The ball kicked the girl" can mean "The girl kicked the ball"
- Lewis et al. (2013): evidence that readers rationally change their eye movement control given their goal functions (as predicted by a computational model)
Probabilistic models of language acquisition
Lecture 4 (part 2)

Klinton Bicknell & Roger Levy

LSA Institute 2015
University of Chicago
Situating language acquisition

day 1: sound categorization

Figure 5.2: Likelihood functions for /b/–/p/ phoneme categorizations, with \( \mu_b = 0 \), \( \mu_p = 50 \), \( \sigma_b = \sigma_p = 12 \). For the input \( x = 27 \), the likelihoods favor /p/.

Figure 5.3: Posterior probability curve for Bayesian phoneme discrimination as a function of VOT. The conditional distributions over acoustic representations, \( P_b(x) \) and \( P_p(x) \) for /b/ and /p/ respectively (the likelihood functions), and the prior distribution over /b/ versus /p/.

We further simplify the problem by characterizing any acoustic representation \( x \) as a single real-valued number representing the VOT, and the likelihood functions for /b/ and /p/ as normal density functions (Section 2.10) with means \( \mu_b \), \( \mu_p \) and standard deviations \( \sigma_b \), \( \sigma_p \) respectively. Figure 5.2 illustrates the likelihood functions for the choices \( \mu_b = 0 \), \( \mu_p = 50 \), \( \sigma_b = \sigma_p = 12 \). Intuitively, the phoneme that is more likely to be realized with VOT in the vicinity of a given input is a better choice for the input, and the greater the discrepancy in the likelihoods the stronger the categorization preference. An input with no negligible likelihood for each phoneme is close to the “categorization boundary,” but may still have a preference. These intuitions are formally realized in Bayes’ Rule:

\[
P(b | x) = \frac{P(x | b)P(b)}{P(x)}
\]  

and since we are considering only two alternatives, the marginal likelihood is simply the weighted sum of the likelihoods under the two phonemes:

\[
P(x) = P(x | b)P(b) + P(x | p)P(p).
\]  

If we plug in the normal probability density function we get

\[
P(b | x) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left[-\frac{(x - \mu_b)^2}{2\sigma_b^2}\right]P(b)
\]

\[
P(p | x) = \frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left[-\frac{(x - \mu_p)^2}{2\sigma_p^2}\right]P(p)
\]  

In the special case where \( \sigma_b = \sigma_p = \sigma \) we can simplify this considerably by cancelling the...
Situating language acquisition

**day 1: inferring sound category** $c$ **from sound token** $S$

- $c \sim$ discrete choice, e.g., $p(p) = p(b) = 0.5$
- $S|c \sim \text{Gaussian}(\mu_c, \sigma^2)$
Statistical Model

day 2a: sound similarity

Actual Stimulus

Perceived Stimulus
Statistical Model

day 2a: inferring target production $T$ from sound token $S$

Choose a category $c$ with probability $p(c)$

Articulate a target production $T$ with probability $p(T|c)$

$$p(T|c) = N(\mu_c, \sigma_c^2)$$

Listener hears speech sound $S$ with probability $p(S|T)$

$$p(S|T) = N(T, \sigma_S^2)$$
day 2b: incremental parsing

The most famous garden-paths: reduced relative clauses (RRCs) versus main clauses (MCs)

From the valence + simple-constituency perspective, MC and RRC analyses differ in two places:

The horse raced past the barn fell.

...(that was)
Statistical Model

day 2b: inferring syntactic structure $T$ from words $w$

$T \sim \text{PCFG}$

words are leaves of the trees
Statistical Model

day 3a: noisy channel sentence processing

The coach smiled at the player tossed the frisbee
The coach smiled at the player thrown the frisbee
The coach smiled toward the player tossed the frisbee
The coach smiled toward the player thrown the frisbee
Statistical Model

day 3a: inferring words $w$ and trees $T$ from perceptual input $I$

- A tree $T \sim \text{PCFG}$
- Words are leaves of the tree
- Visual input $I \sim \text{noise}(w)$
Situating language acquisition

**day 3b: sentence meaning judgments**

The zarg prolted the cherid to a really gromious flig.

Which is more likely?

- The cherid is in a new place.
- The cherid has a new owner.

LOCATIVE inference
POSSESSIVE inference
day 3b: inferring meaning $M$ from linguistic form $F$ and processing pressures $P$

$F \sim \text{function}(M, P)$
Situating language acquisition

in all these cases

• step 1: identify the relevant sources of information

• step 2: make a generative model in which
  
  • the thing to be inferred was a latent variable
  
  • the relevant information was used to specify a prior for the latent variable or the likelihood of the data given that latent variable

• step 3: apply Bayesian inference (relatively easy here: given prior and likelihood)
Problems in acquisition

step 1

let's apply step 1 (identify relevant information sources) to problems in acquisition
Problems in acquisition

learning to segment words

[yuwanttusidəbuk]? -> you want to see the book?

[e.g., Goldwater et al., 2009]
Problems in acquisition

learning word meanings

[e.g., Frank et al., 2009]
Problems in acquisition

learning phonotactics

[bli:k]

[mdwi:] Polish mdły: 'tasteless'

[e.g., Hayes & Wilson, 2008]
Problems in acquisition

learning syntax

The boy is hungry. -> Is the boy hungry?

The boy who is smiling is happy. -> ???

Is the boy who is smiling happy?

*Is the boy who smiling is happy?

[e.g., Perfors et al., 2011]
Problems in acquisition

learning the sound categories in the language from the sounds

• are long VOTs [t] and short VOTs [d] functionally different sounds or just natural variation?

[e.g., Pajak et al., 2013]
Problems in acquisition

steps 2 and 3: construct a generative model, perform inference

- a bit different than before

- we'll go in depth through an example of sound category learning

- note: much of the work we'll be discussing is by Dr. Bożena Pająk (now a learning scientist at Duolingo), who also made many of the visualizations you'll see
Distributional information

**learning sound categories**

- how do we learn that [t] and [d] are different categories?
- information source 1: distributional information

![Distributional information with VOT](image)
experimental evidence

- We know that babies and adults use distributional information to help infer category structure [Maye & Gerken, 2000; Maye et al., 2002]

- Pajak & Levy (2011) performed an experiment replicating this with adults, which I'll describe

Distributional information

sound class: FRICATIVES

- Length /s/ /ss/
- Singleton
- Geminate
Experimental data (Pajak & Levy 2011)

- Distributional training:
  - adult English native speakers exposed to words in a new language, where the middle consonant varied along the length dimension

  [aja]_{145\text{ms}}
  [ina]_{205\text{ms}}
  [ila]_{115\text{ms}}
  [ama]_{160\text{ms}}
  ...

Experimental data (Pajak & Levy 2011)

- Expt 1: Sonorants
- Expt 2: Fricatives

Graph showing the familiarization frequency across different stimuli length continua (in msec) for bimodal and unimodal conditions.
Experimental data (Pajak & Levy 2011)

Testing:
- participants made judgments about pairs of words

  Example: [ama]-[amma]
  "Are these two different words in this language or two repetitions of the same word?"

- dependent measure: proportion of ‘different’ responses (as opposed to ‘same’) on ‘different’ trials
- if learning is successful, we expect:

  ‘DIFFERENT’ RESPONSES

  Bimodal training > Unimodal training
Experimental data (Pajak & Levy 2011)

Stimuli length continuum (in msec)

Familiarization frequency

- bimodal
- unimodal

test stimuli
Experimental data (Pajak & Levy 2011)

Expt1 - trained
Expt1 - untrained

Proportion of 'different' responses

- Bimodal
- Unimodal
Distributional information

distributional information is used

• adults and babies use distributional information to infer categories

• next steps: put this information into a generative model and perform inference
A generative model

our model from before of where sounds come from

- $c \sim$ discrete choice, e.g., $p(p) = p(b) = 0.5$
- $S|c \sim Gaussian(\mu_c, \sigma^2_c)$

still pretty appropriate!
A generative model

a model of where categories and sounds come from

\[ y \sim N(\mu_i, \Sigma_i) \]  
\[ i \sim \text{discrete}(\phi) \]
\[ n = \# \text{observations} \]
\[ m = \# \text{categories} \]
\[ \phi \sim \text{distribution}() \]
\[ \mu_i \sim \text{distribution}() \]
\[ \Sigma_i \sim \text{distribution}() \]

' mixture of Gaussians'

\[ \sigma^2 \text{ before} \]

\( \text{set in advance} \)
Inference

**final step: inference**

- now we have generative model with prior on variables of interest and well-defined likelihood
- but inference is usually quite hard
- two common classes of methods for inference, which I'll describe now
Inference

**inference method 1: Expectation Maximization (EM)**

- if we knew the category parameters (means, variance, probability), it would be easy to categorize datapoints like before

- if we knew how the datapoints were categorized, it would be easy to find category parameters

  - Q: how?

- idea behind EM:

  - randomly initialize category assignments of datapoints

  - estimate category parameters from current assignments

  - now assign datapoints to categories based on category parameters

  - estimate category parameters from current assignments

  - …
Inference

EM illustrations (on board)

• bimodal data

• two categories
Inference

EM issues

- only finds most likely value for variables, not posterior distribution on them
- can get stuck in 'local optima'

![Diagram of EM issues with points A, B, and C, showing the process of moving between local optima.](Image)
Inference

inference method 2: Markov chain Monte Carlo

- like EM, the model is in some state at each point in time (with current guesses for latent variables)

- like EM, the model transitions between states

- unlike EM, the model sometimes moves to lower probability states

- if you calculate the transition probabilities correctly, the amount of time the model spends in each state is proportional to its posterior probability

- thus: don't just get most likely variables out, but full posterior distribution

- also: guaranteed to converge (not getting stuck in local optima)

  - however, this guarantee is only with infinite time…
mixture of Gaussian models to learn phonetic categories

- many groups have had success, at least on simple problems (Vallabha et al., 2007; McMurray et al., 2009; Feldman et al., 2009)

- but it seems clear that other information sources are needed too, because categories overlap too much

- next up: work by Pajak and colleagues on another information source
BULGARIAN consonants

<table>
<thead>
<tr>
<th></th>
<th>Bilabial</th>
<th>Labiodental</th>
<th>Dental/Alveolar</th>
<th>Post-alveolar</th>
<th>Palatal</th>
<th>Velar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nasal</td>
<td>hard</td>
<td>m</td>
<td>(m̥)</td>
<td>n</td>
<td>(n̥)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>soft</td>
<td>m̥j</td>
<td></td>
<td>j</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plosive</td>
<td>hard</td>
<td>p</td>
<td>t</td>
<td>d</td>
<td>k</td>
<td>g</td>
</tr>
<tr>
<td></td>
<td>soft</td>
<td>p̥j</td>
<td>t̥j</td>
<td>d̥j</td>
<td>c</td>
<td></td>
</tr>
<tr>
<td>Affricate</td>
<td>hard</td>
<td>ts</td>
<td>tʃ</td>
<td>dz</td>
<td>tʃ̥</td>
<td>dʒ</td>
</tr>
<tr>
<td></td>
<td>soft</td>
<td>ts̥j</td>
<td>dz̥j</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fricative</td>
<td>hard</td>
<td>f</td>
<td>s</td>
<td>z</td>
<td>x</td>
<td>(y)</td>
</tr>
<tr>
<td></td>
<td>soft</td>
<td>f̥j</td>
<td>s̥j</td>
<td>z̥j</td>
<td>(x̥j)</td>
<td></td>
</tr>
<tr>
<td>Trill</td>
<td>hard</td>
<td>r</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>soft</td>
<td>r̥j</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximant</td>
<td>hard</td>
<td>(w)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>soft</td>
<td></td>
<td></td>
<td>j</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lateral</td>
<td>hard</td>
<td>t</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>soft</td>
<td></td>
<td></td>
<td>λ</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What may help solve this problem

- Learning may be facilitated by languages’ extensive re-use of a set of phonetic dimensions (Clements 2003)

<table>
<thead>
<tr>
<th>THAI vowels</th>
<th>Front</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>unrounded</td>
<td>unrounded</td>
</tr>
<tr>
<td></td>
<td>short</td>
<td>long</td>
</tr>
<tr>
<td>Close</td>
<td>/i/</td>
<td>/iː/</td>
</tr>
<tr>
<td>Close-mid</td>
<td>/e/</td>
<td>/eː/</td>
</tr>
<tr>
<td>Open-mid</td>
<td>/ɛ/</td>
<td>/ɛː/</td>
</tr>
<tr>
<td>Open</td>
<td>/a/</td>
<td>/aː/</td>
</tr>
</tbody>
</table>

(Wikipedia: Thai language)
Learning may be facilitated by languages’ extensive re-use of a set of phonetic dimensions (Clements 2003)

Existing experimental evidence supports this view

- both infants and adults generalize newly learned phonetic category distinctions to untrained sounds along the same dimension (McClaskey et al. 1983, Maye et al. 2008, Perfors & Dunbar 2010, Pajak & Levy 2011)
How do people learn phonetic categories?

- What are the mechanisms underlying generalization?
- How do learners make use of information about some phonetic categories when learning other categories?
How do people learn phonetic categories?

- Pajak et al.’s proposal:
  - in addition to learning specific categories, people also learn category types

\[ \text{singleton} \leftrightarrow \text{geminate} \]
Experimental data illustrating generalization across analogous distinctions (from Pajak & Levy 2011)

Our computational proposal for how this kind of generalization might be accomplished

Simulations of Pajak & Levy’s data
Experimental data (Pajak & Levy 2011)

sound class: FRICATIVES

length /s/ /ss/

singleton geminate

sound class: SONORANTS

length /n/ /nn/
Experimental data (Pajak & Levy 2011)

- Distributional training:
  - adult English native speakers exposed to words in a new language, where the middle consonant varied along the length dimension

\[ \text{[aja]}_{145\text{ms}} \]
\[ \text{[ina]}_{205\text{ms}} \]
\[ \text{[ila]}_{115\text{ms}} \]
\[ \text{[ama]}_{160\text{ms}} \]
...
Experimental data (Pajak & Levy 2011)

Stimuli length continuum (in msec)

Familiarization frequency

Expt1:sonorants

[n]-----[nn]
[m]-----[mm]
[l]-----[ll]
[j]-----[jj]

Expt2:fricatives

[s]-----[ss]
[f]-----[ff]
[j]-----[jj]
[θ]-----[θθ]

This difference reflects natural distributions of length in different sound classes.

Training

bimodal
unimodal
Experimental data (Pajak & Levy 2011)

- Testing:
  - participants made judgments about pairs of words
    - **Example:** [ama]-[amma]
      - “Are these two different words in this language or two repetitions of the same word?”
  - dependent measure: proportion of ‘different’ responses (as opposed to ‘same’) on ‘different’ trials
  - if learning is successful, we expect:
    - ‘DIFFERENT’ RESPONSES
      - Bimodal training > Unimodal training
  - to assess generalization, testing included both trained and untrained sound classes (i.e., both sonorants and fricatives)
EXPT 1: ALIGNED CATEGORIES

Experimental data (Pajak & Levy 2011)

Stimuli length continuum (in msec)
Experimental data (Pajak & Levy 2011)

EXPT 1: ALIGNED CATEGORIES

EXPT 2: MISALIGNED CATEGORIES

Expt1–trained

Expt1–untrained

Expt2–trained

Expt2–untrained
Computational modeling

① How can we account for distributional learning?

② How can we account for generalization across sound classes?
Modeling phonetic category learning

- **Mixture of Gaussians approach**

\[
d_i \sim \mathcal{N}(\mu_{z_i}, \sigma_{z_i}^2)
\]

**Phonetic category**

**Datapoint (perceptual token)**

Inference

<table>
<thead>
<tr>
<th>length</th>
<th>/s/</th>
<th>/ss/</th>
</tr>
</thead>
<tbody>
<tr>
<td>120ms</td>
<td>201ms</td>
<td>165ms</td>
</tr>
</tbody>
</table>
Modeling phonetic category learning

- **Our general approach** (following Feldman et al. 2009):
  - learning via nonparametric Bayesian inference
  - using Dirichlet processes, which allow the model to learn the number of categories from the data

\[
\begin{align*}
G_0 &
\sim DP(g, H) \\
z_i &
\sim G_0 \\
d_i &
\sim \mathcal{N}(\mu_{z_i}, \sigma_{z_i}^2) \\
\mu_s, \sigma_s^2 &
\sim \mathcal{N}(\mu_0, \sigma_0^2) \\
\mu_{ss}, \sigma_{ss}^2 &
\sim \text{InvChiSq}(n_0, \sigma_0^2)
\end{align*}
\]

\(i \in \{1..n\}\)
Computational modeling

① How can we account for distributional learning?

② How can we account for generalization across sound classes?
Proposal

- In addition to acquiring specific categories, learners infer **category types**, which can be shared across sound classes.

- This means that already learned categories can be directly re-used to categorize other sounds.

- To implement this proposal, we use a **hierarchical Dirichlet process**, which allows for sharing categories across data groups (here, sound classes).
Modeling generalization: HDP

prior

phonetic category

datapoint (perceptual token)

\[ G_c \]

\[ z_{ic} \sim G_c \]

\[ d_{ic} \sim \mathcal{N}(\mu_{zic}, \sigma_{zic}^2) \]

\( c = \text{sonorants} \)

\( c \in C \)

fricatives

<table>
<thead>
<tr>
<th>length</th>
<th>/s/</th>
<th>/ss/</th>
</tr>
</thead>
</table>

sonorants

<table>
<thead>
<tr>
<th>length</th>
<th>/n/</th>
<th>/nn/</th>
</tr>
</thead>
</table>

fric-sg fric-gem

<table>
<thead>
<tr>
<th>length</th>
<th>/s/</th>
<th>/ss/</th>
</tr>
</thead>
</table>

\( \mu_{\text{fric-sg}} \)

\( \mu_{\text{fric-gem}} \)

\( \sigma_{\text{fric-sg}} \)

\( \sigma_{\text{fric-gem}} \)

son-sg son-gem
But people are able to generalize even when analogous category types are implemented phonetically in different ways.

We want the model to account for potential differences between sound classes.
Accounting for differences between sound classes:

learnable class-specific ‘offsets’ by which data in a class are shifted along a phonetic dimension (cf. Dillon et al. 2013)

\[ H : \mu \sim \mathcal{N}(\mu_0, \sigma_k^2) \]
\[ \sigma^2 \sim \text{InvChiSq}(v_0, \sigma_0^2) \]
\[ G_0 \sim \text{DP}(\gamma, H) \]
\[ G_c \sim \text{DP}(\alpha_0, G_0) \]
\[ z_{ic} \sim G_c \]
\[ f_c \sim \mathcal{N}(0, \sigma_f^2) \]
\[ d_{ic} \sim \mathcal{N}(\mu_{z_{ic}}, \sigma_{z_{ic}}^2) + f_c \]
Simulation results

EXPT 1: ALIGNED CATEGORIES

Proportion of 2−category inferences

HUMAN DATA

EXPT 2: MISALIGNED CATEGORIES

Proportion of 'different' responses
Modeling category learning

other work on category learning

- Feldman et al. (2009) add a (latent) lexicon to category learning
probabilistic models in acquisition

- in general: still working on trying to incorporate many sources of information

- it seems that infants, children, and adults use many sources of information to learn language

- one very exciting information source: other languages
Conclusions

computational psycholinguistics

• this course gave a very broad overview of many areas: perception, comprehension, production, acquisition

• gave a sense for how probabilistic models can be used

• also covered much of the essential technical knowledge to use probabilistic models

• this is still very much the beginning of probabilistic modeling in the study of language: very many open questions and models yet to build!

• hope you enjoyed!