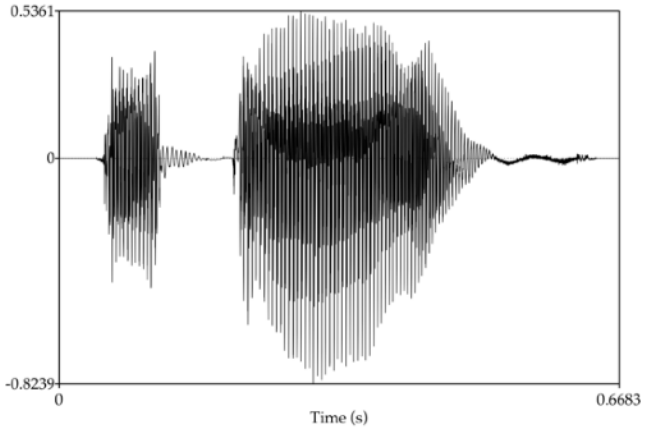
Speech perception

- Recognizing meaningful word sequences from sound (or sign) is a remarkable feat of human intelligence



"above"

- Other species don't do this, as far as we know

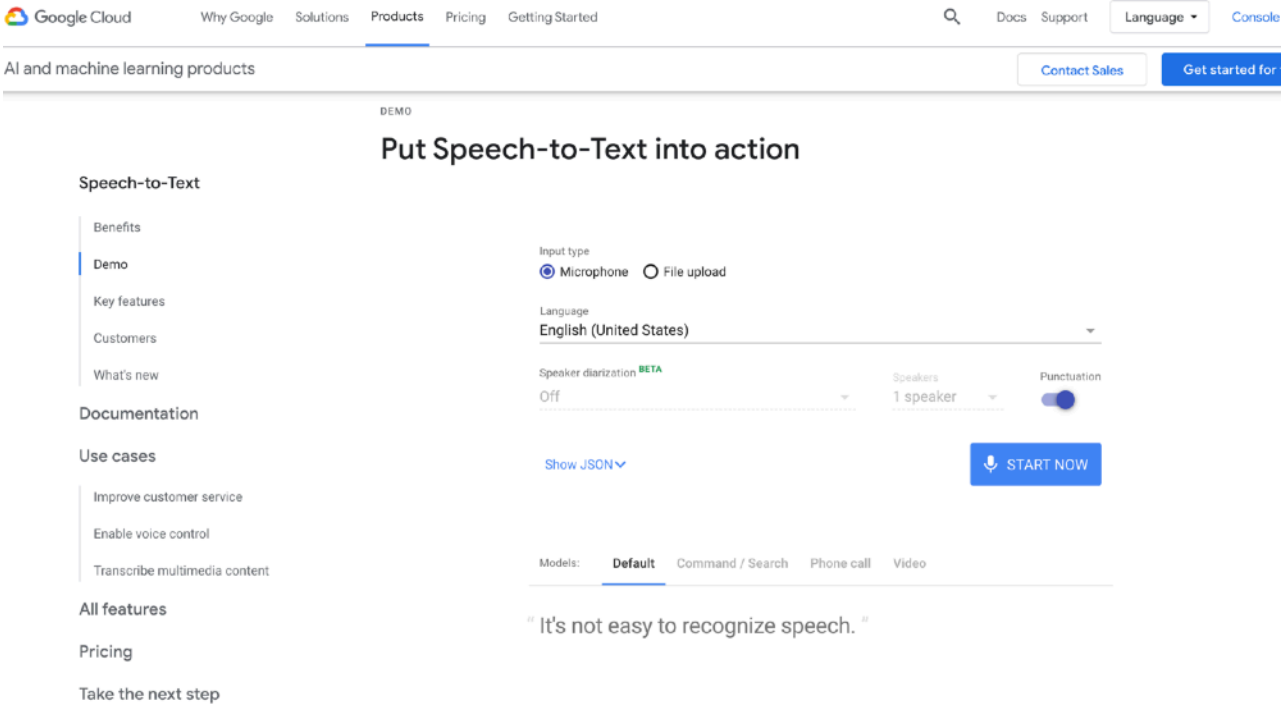
Speech perception

- The problem's harder than everyday experience suggests
  - What are all the sentences that each recording could be?

It's not easy to recognize speech *Phil and Mary are young cousins* *Yanny*

It's not easy to wreck a nice beach *Phil and Mary are our young cousins* *Laurel*

- Computer speech recognition is impressive but still fragile

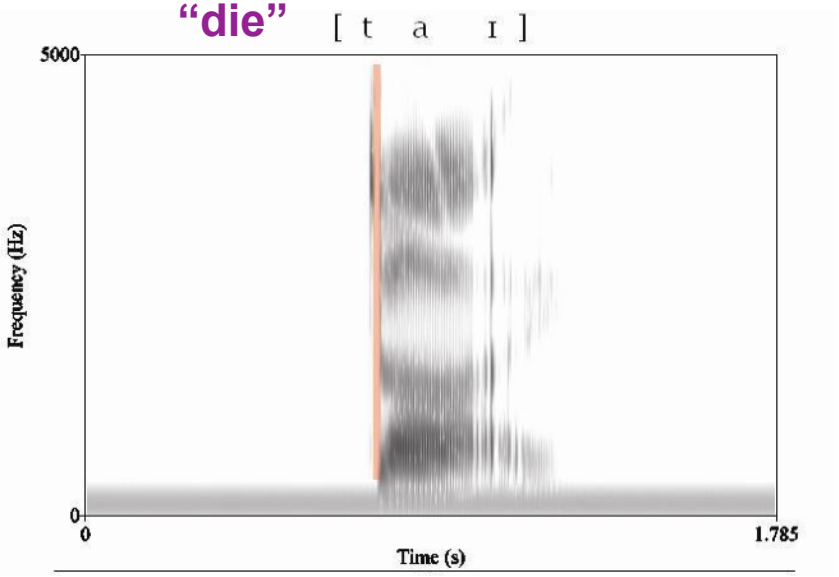


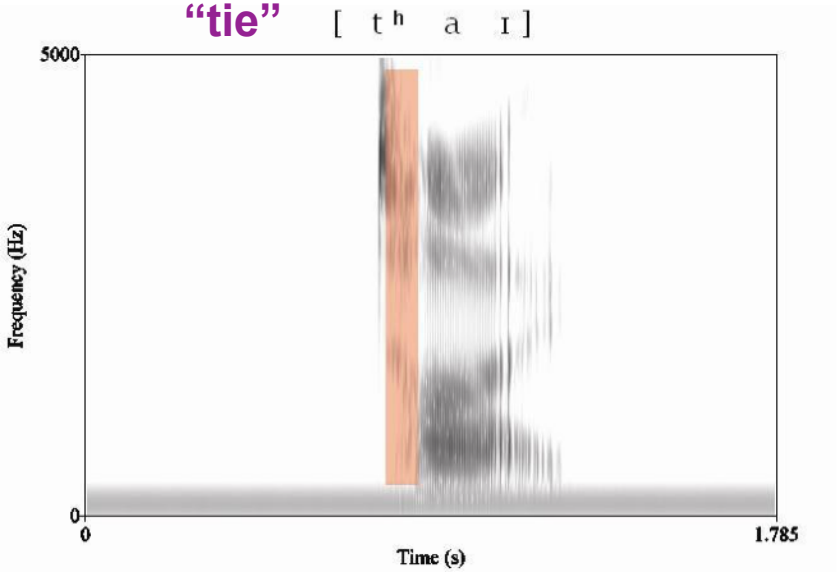
A minimal problem in speech perception

- We'll break down this hard & complex problem into something much simpler, and then scale up
- How to discriminate two minimally different sounds?

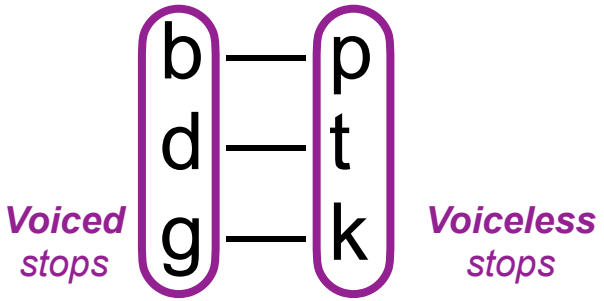
Label each one *When do you hear the transition?*

A minimal problem in speech perception

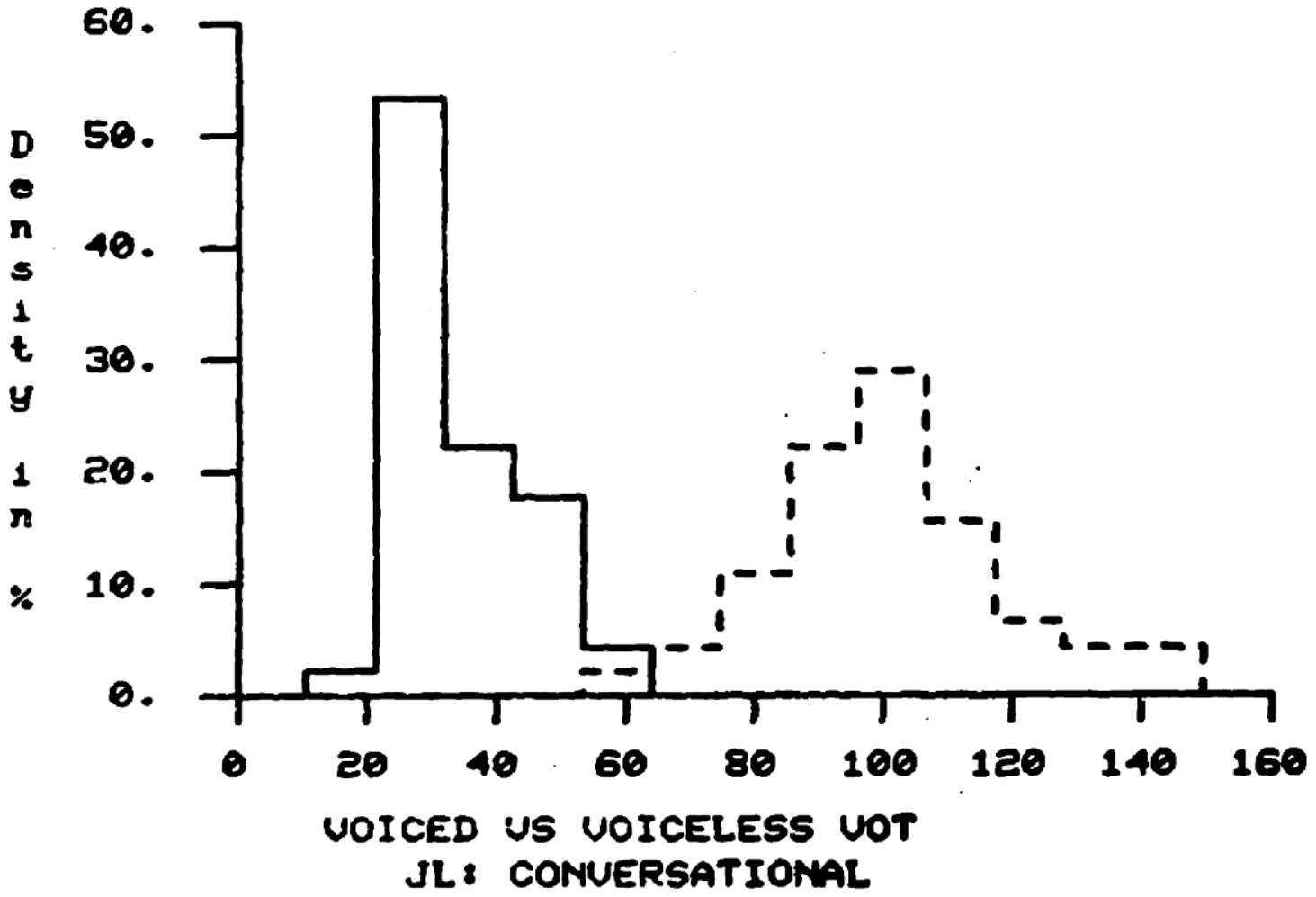




- This distinction involves a single "phonetic dimension", **Voice Onset Time (VOT)**

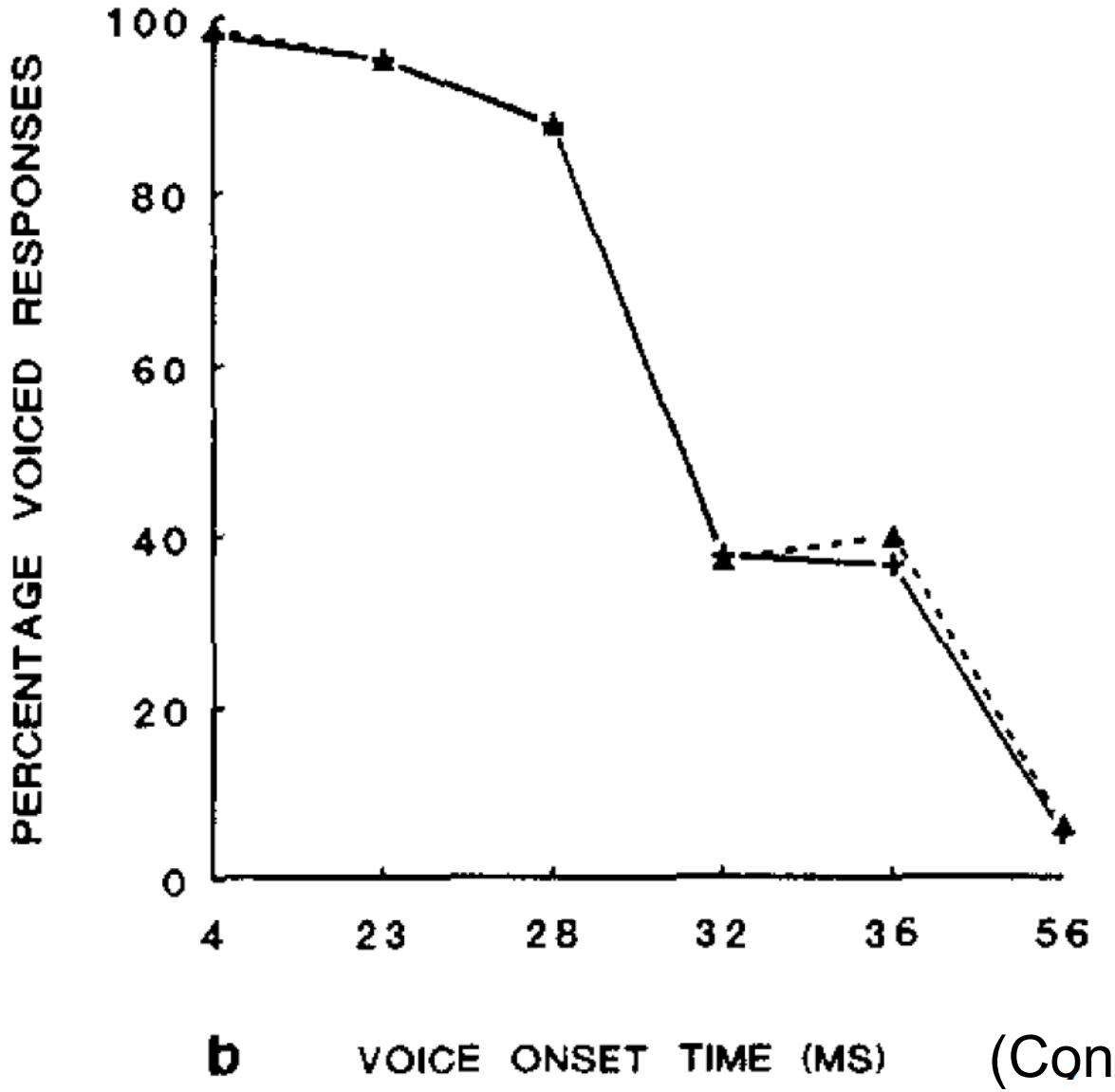


Distribution of VOTs



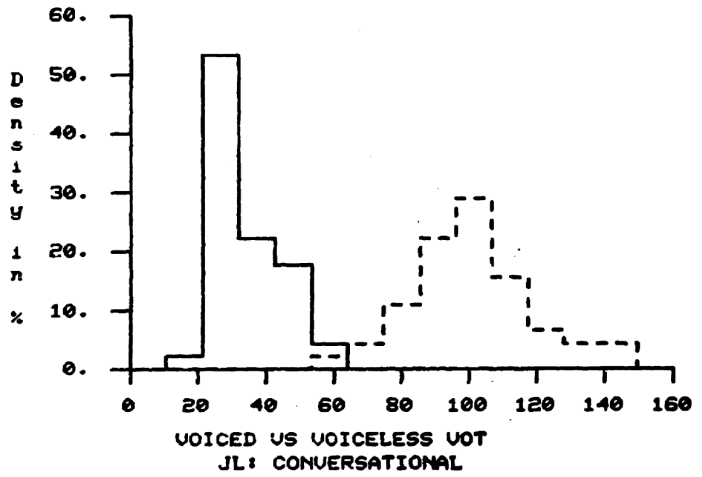
(Chen, 1980)

Human Categorization Curve



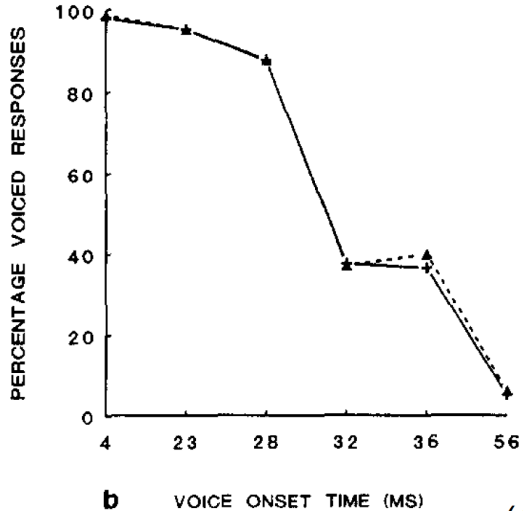
(Connine et al., 1991)

***Generative**, or **forward,** distributional knowledge*



$P(\text{VOT} \mid \text{class})$

***Classification** ("reverse-engineering") behavior*



$P(\text{class} \mid \text{VOT})$

How can we reconcile these two distributions?

Rational analysis

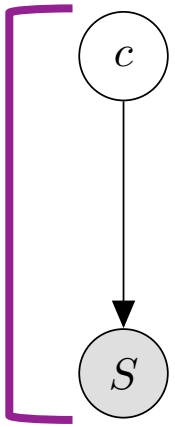
- Background assumption: cognitive agent is optimized via evolution and learning to solve everyday tasks effectively

1. Specify precisely the goals of the cognitive system
2. Formalize model of the environment to which the cognitive system is adapted
3. Make minimal assumptions re: computational limitations
4. Derive predicted optimal behavior given 1–3
5. Compare predictions with empirical data
6. If necessary, iterate 1–5

(Anderson, 1990, 1991)

Modeling VOT-based recognition

- Assume sound category *c* manifests as speech signal *S*



Generative process

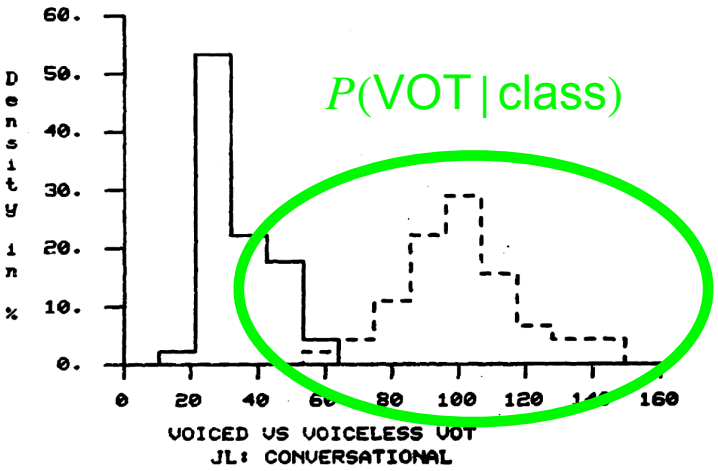
$$P(c|S) = \frac{P(S|c)P(c)}{\sum_{c'} P(S|c')P(c')}$$

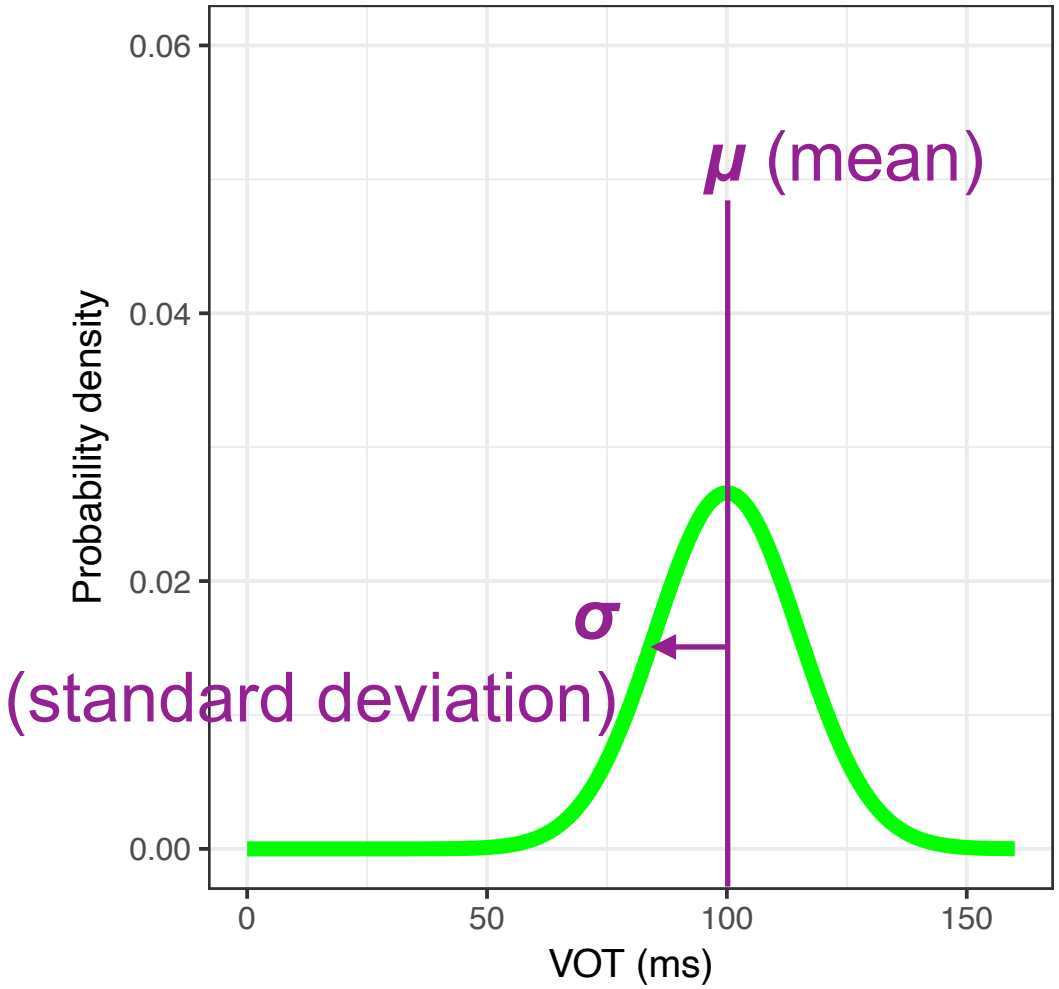
- We can use Bayesian inference to infer *c* from *S*
- To make this work, we need to choose:
  - a ***prior*** P(*c*); and
  - a ***likelihood*** P(*S*|*c*)
- Out of context, the prior might be ***uniform***

$$P(c = /\text{b}/) = P(c = /\text{p}/) = \frac{1}{2}$$

Likelihood for a phonetic dimension

- The ***normal*** (a.k.a. ***Gaussian***) ***distribution*** is a reasonable proxy





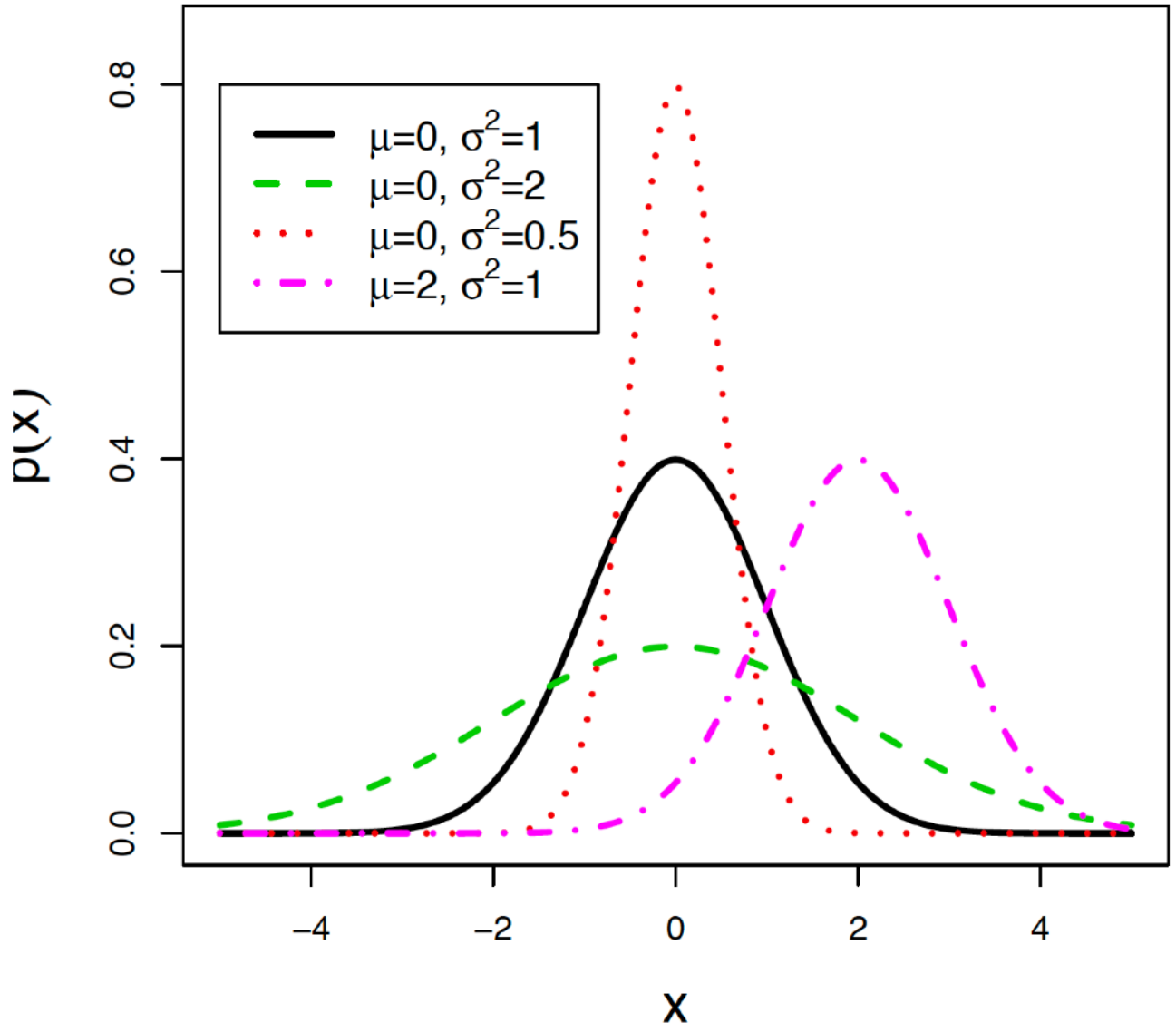
$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]$$

Squared deviation from mean

(normalizing constant)

Scaled by variance

Gaussian parameters



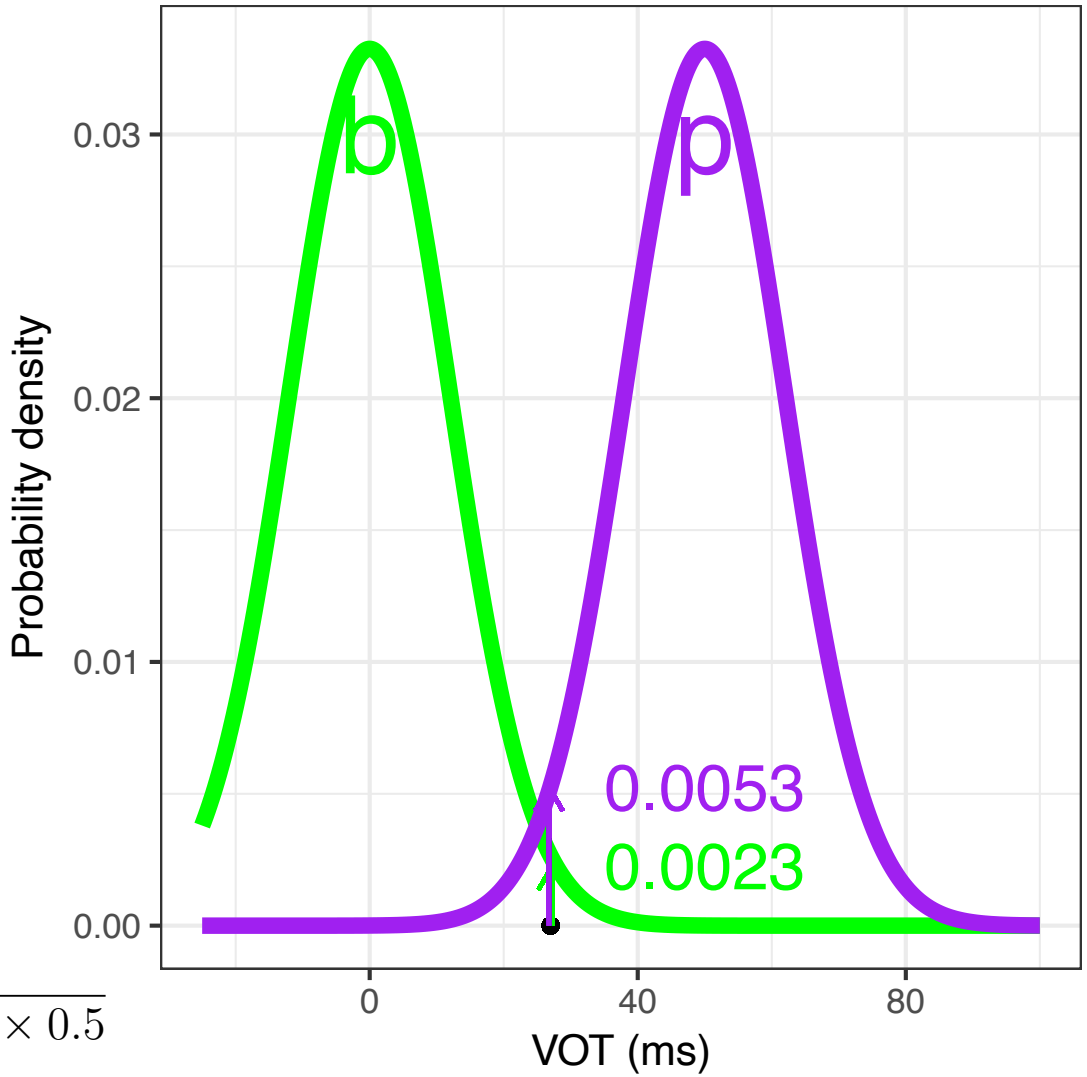
Inference of phonetic category

P(/b/)=P(/p/)=0.5

$$P(c|S) \propto P(S|c)P(c)$$

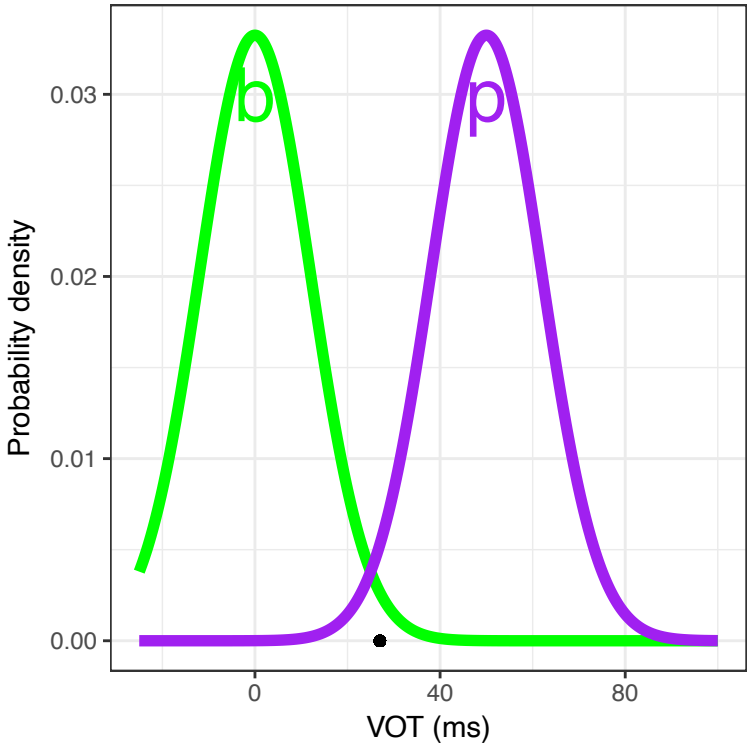
$$S|c \sim N(\mu_c, \sigma^2)$$

"S is normally distributed with mean μ_c and variance σ^2

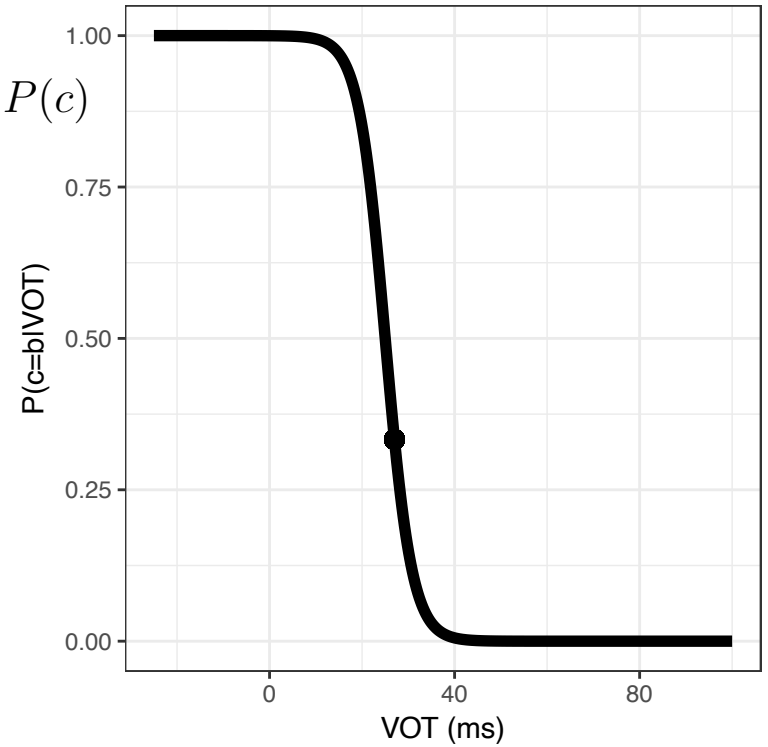


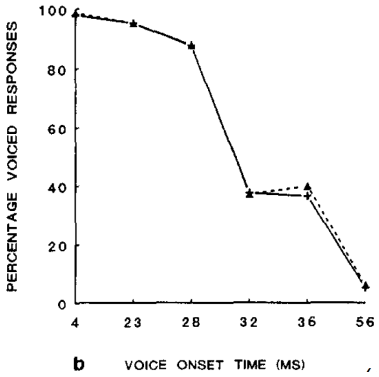
$$P(c = \mathrm{p}|S) = \frac{0.0053 \times 0.5}{0.0053 \times 0.5 + 0.0023 \times 0.5}$$

Bayesian categorization curve

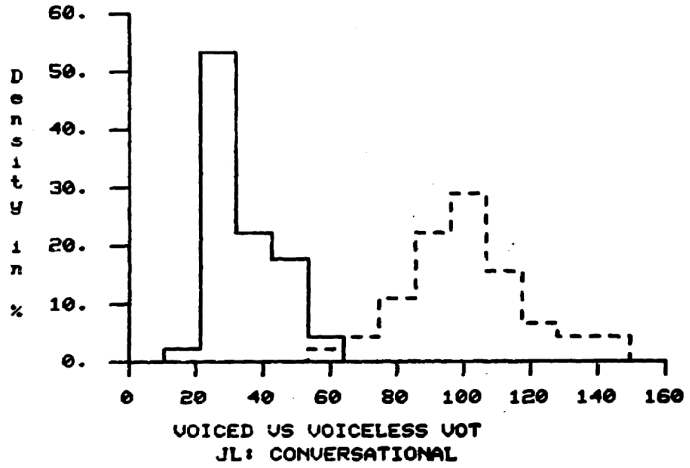


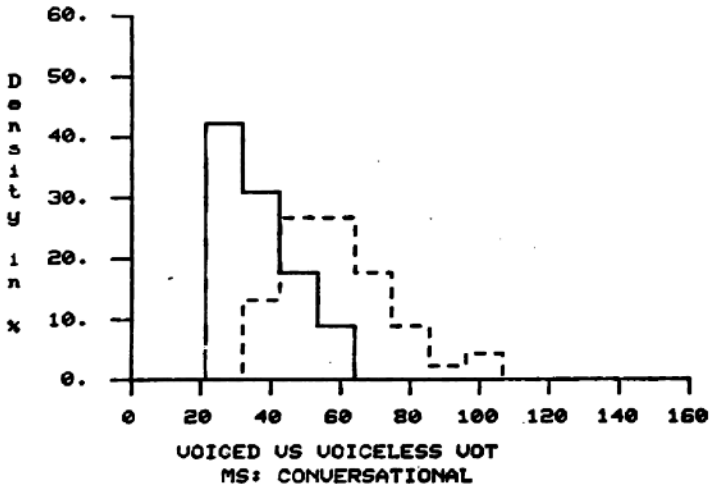
$$P(c|S) \propto P(S|c)P(c)$$

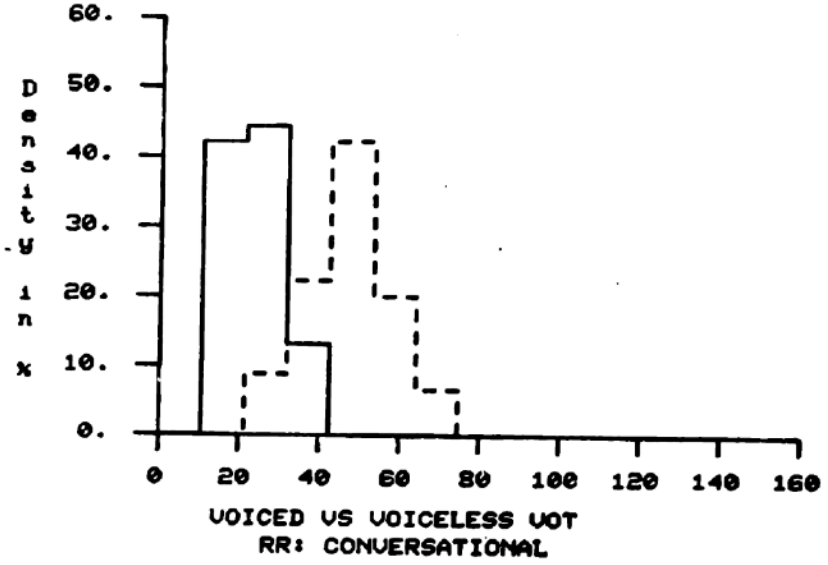


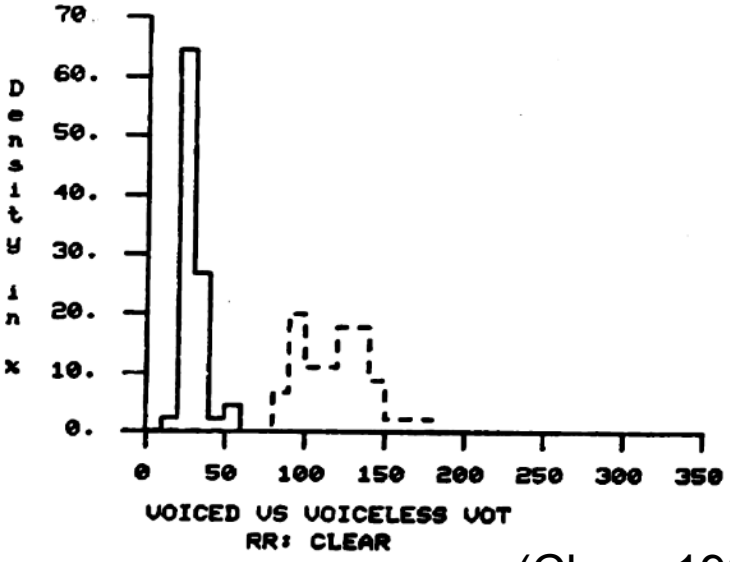


Inter-speaker variability





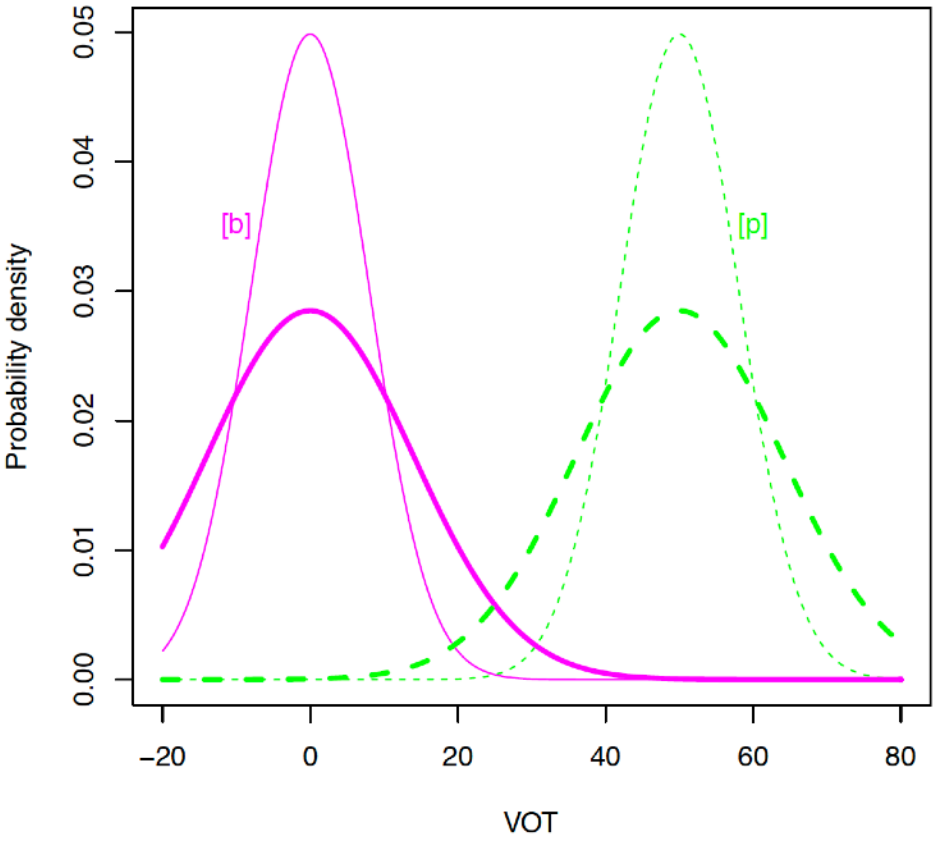




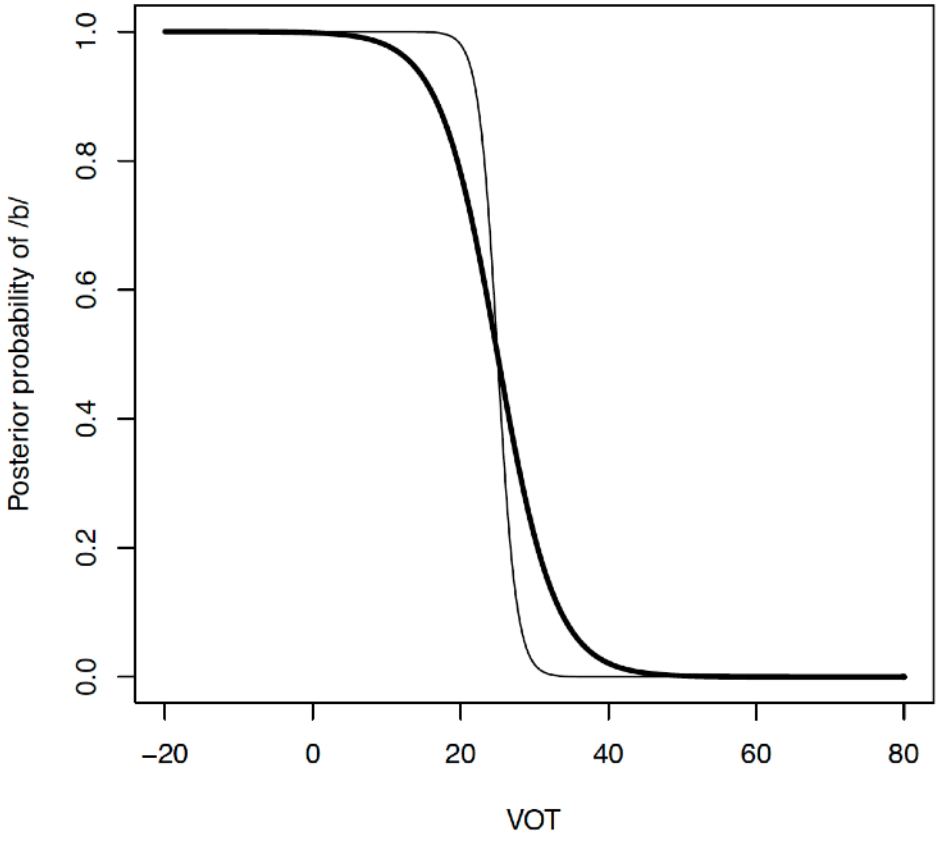
(Chen, 1980)

Ideal observer predictions

$P(\text{VOT} \mid \text{class}, \text{Speaker})$



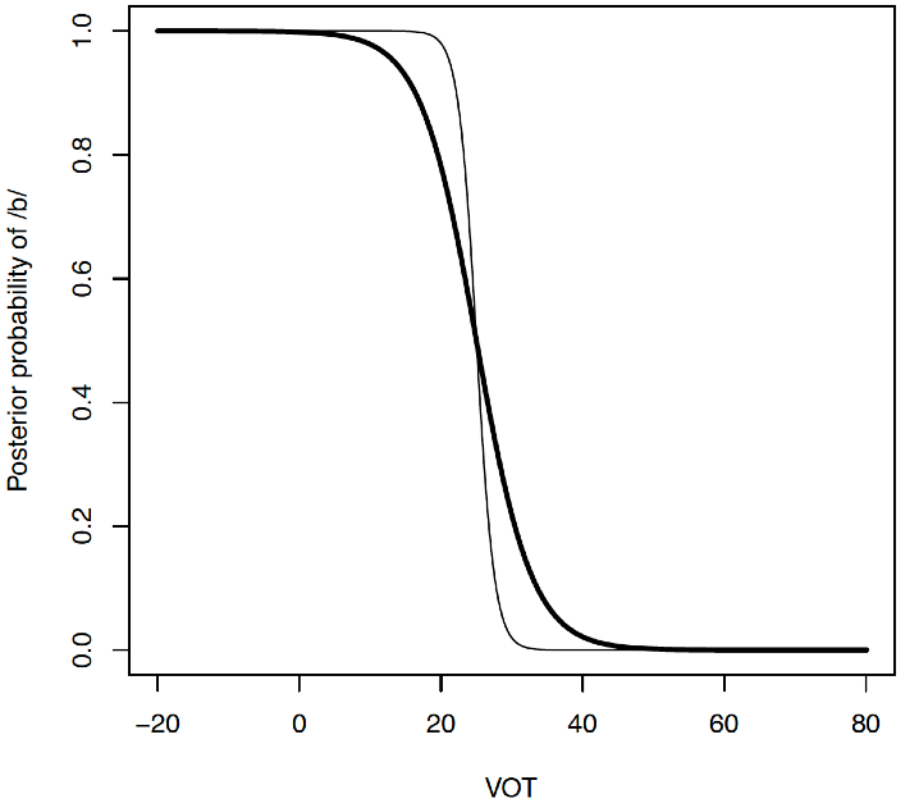
$P(\text{class} \mid \text{VOT}, \text{Speaker})$



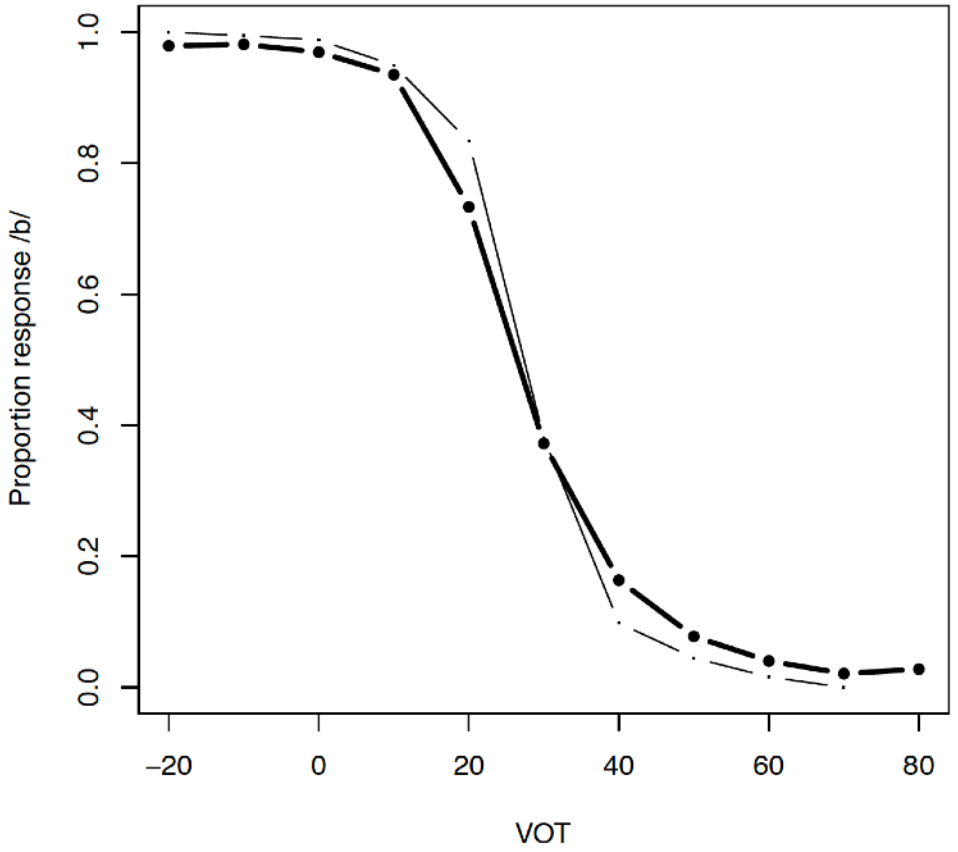
Testing effects of environment variability

- Clayards et al. (2008) tested this prediction
  - trained participants with different-variance Gaussians
  - then tested categorization

Ideal observer



Human performance



Speech perception case study: summary

- We covered a simple case of **classification** in a continuous-signal setting: **phoneme identification**
- We covered the principles of **rational analysis** that allow us to construct an **ideal observer** model of the process
- We used Bayesian inference to implement that model
- We explored a theoretical prediction of the implemented model and saw that experimental data confirmed the prediction