

# Speech perception

---

# Speech perception

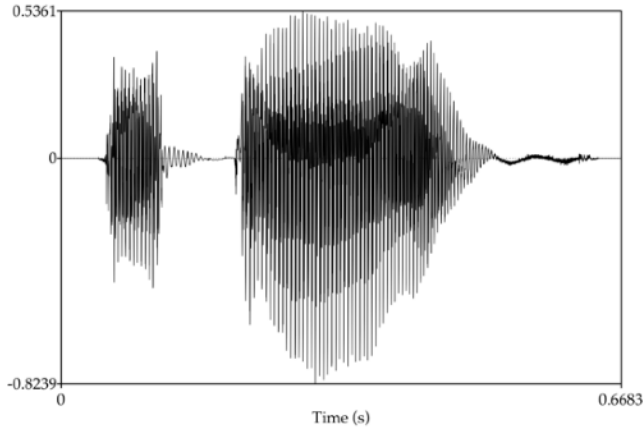
---

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

# Speech perception

---

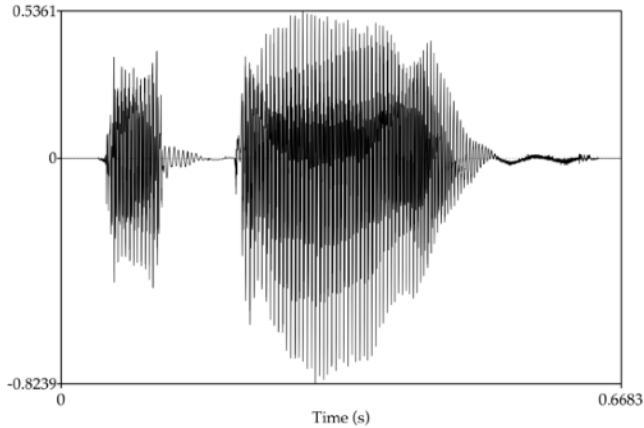
- Recognizing meaningful word sequences from sound (or sign) is a remarkable feat of human intelligence



# Speech perception

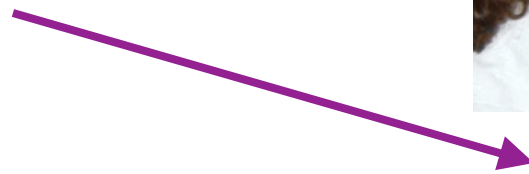
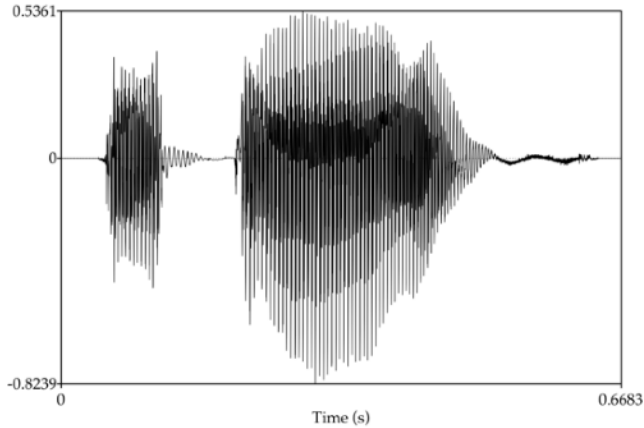
---

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



# Speech perception

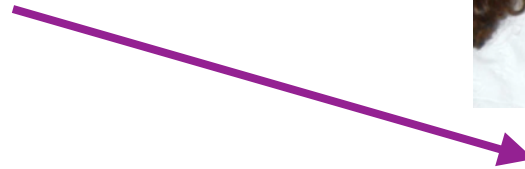
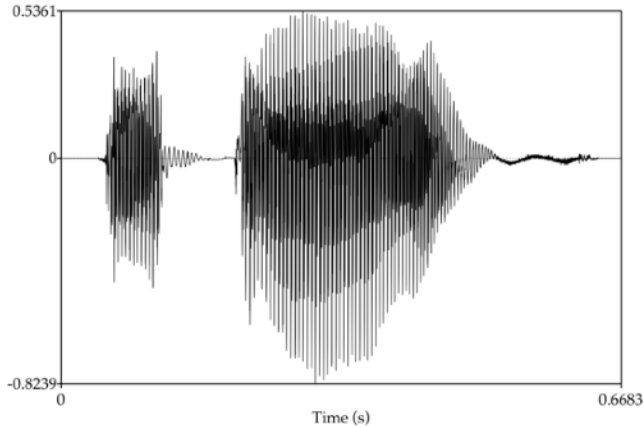
- Recognizing meaningful word sequences from sound (or sign) is a remarkable feat of human intelligence



**"above"**

# Speech perception

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

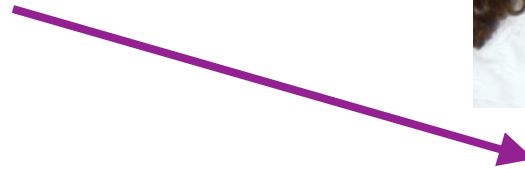
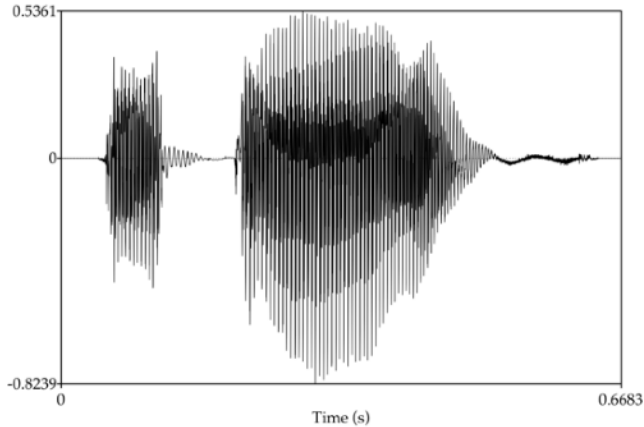


**“above”**

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

# Speech perception

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



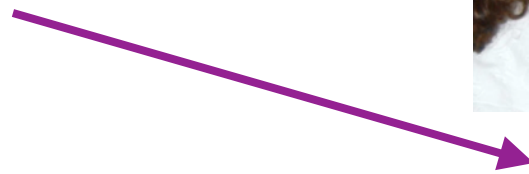
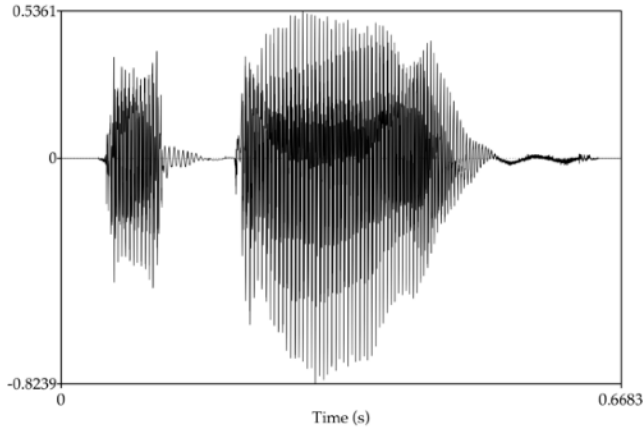
**“above”**

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



# Speech perception

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



**“above”**

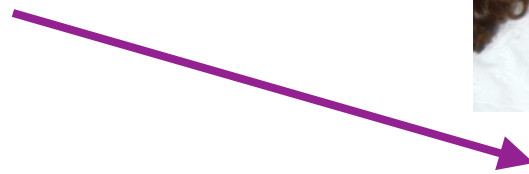
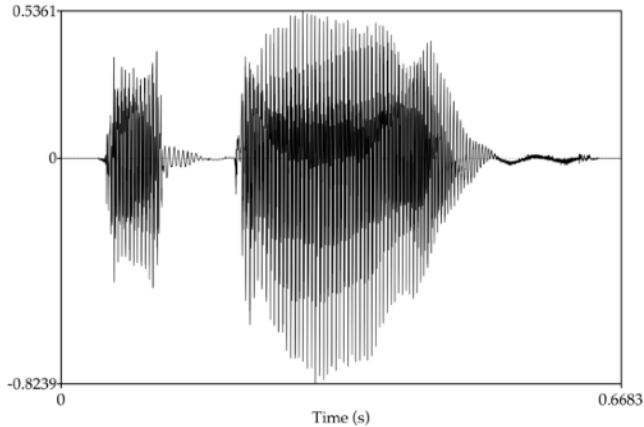
- Other species don't do this, as far as we know





# Speech perception

- 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

---

# Speech perception

---

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

# Speech perception

---

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

# 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*

# 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*

*It's not easy to wreck a nice beach*

# 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*

*It's not easy to wreck a nice beach*

# 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*

*It's not easy to wreck a nice beach*



# 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*

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

# 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*

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

# 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*

# 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*

# 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

The screenshot shows the Google Cloud Speech-to-Text demo page. At the top, there's a navigation bar with 'Google Cloud', 'Why Google', 'Solutions', 'Products', 'Pricing', and 'Getting Started'. On the right, there are search and support options. Below the navigation, the page title is 'AI and machine learning products' with 'Contact Sales' and 'Get started for' buttons. The main heading is 'Put Speech-to-Text into action'. On the left, there's a sidebar menu with categories like 'Speech-to-Text', 'Documentation', 'Use cases', and 'All features'. The 'Demo' option is selected. The main content area shows configuration options: 'Input type' (Microphone selected, File upload), 'Language' (English (United States)), 'Speaker diarization' (Off), 'Speakers' (1 speaker), and 'Punctuation' (On). There's a 'Show JSON' dropdown and a 'START NOW' button. At the bottom, there are model options: 'Default', 'Command / Search', 'Phone call', and 'Video'. The text 'It's not easy to recognize speech.' is displayed at the bottom of the interface.

# 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

---

- 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

---

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

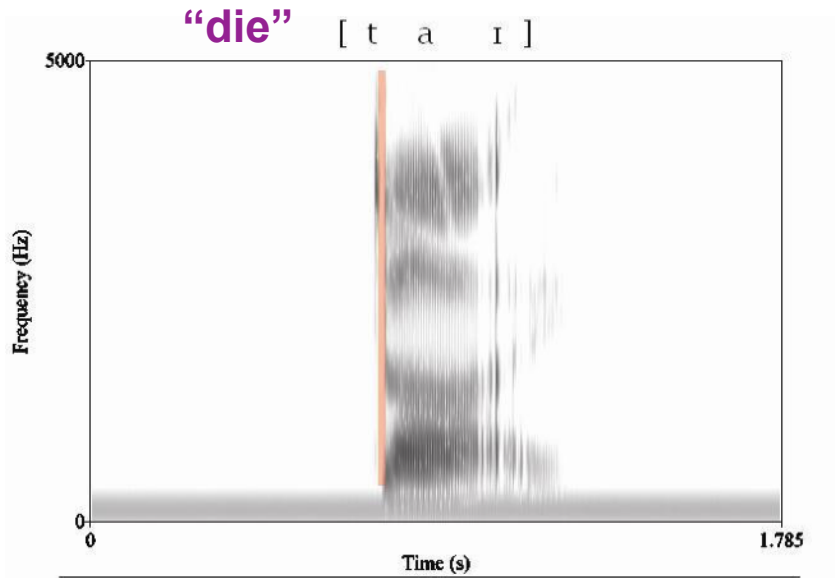
*Label each one*

*When do you hear the transition?*



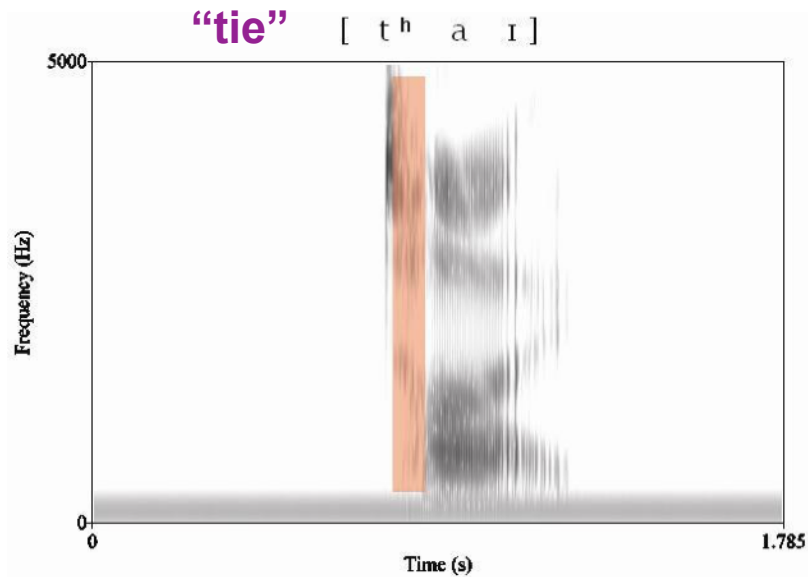
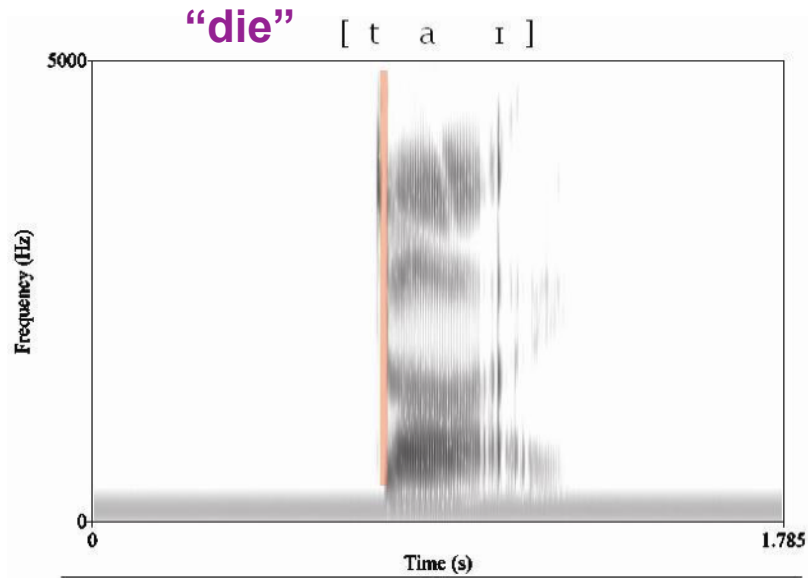
# A minimal problem in speech perception

---

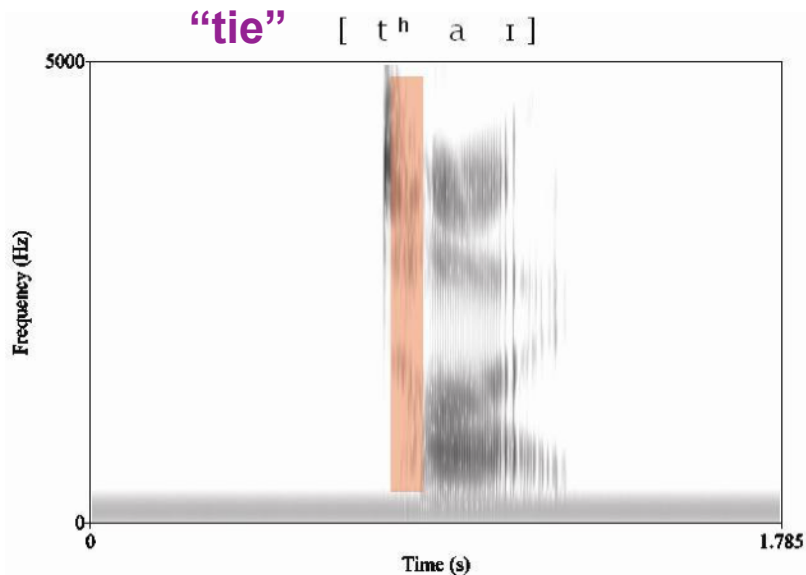
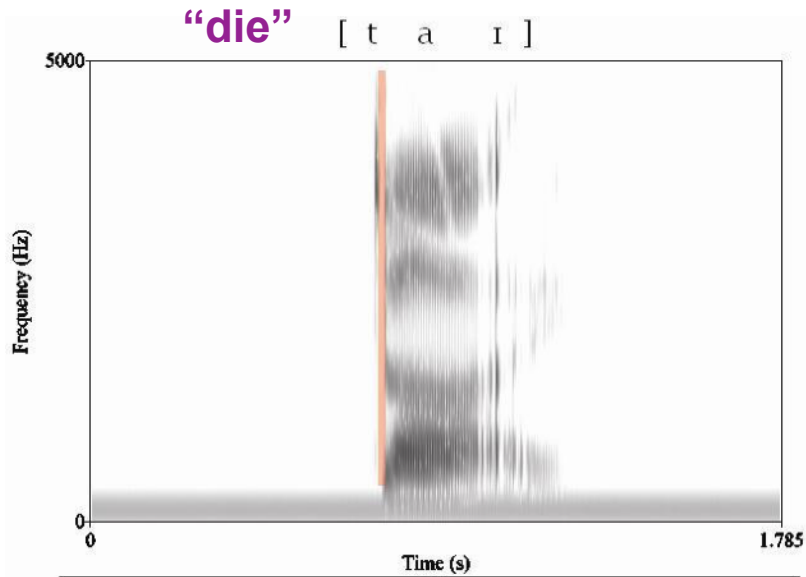


# A minimal problem in speech perception

---



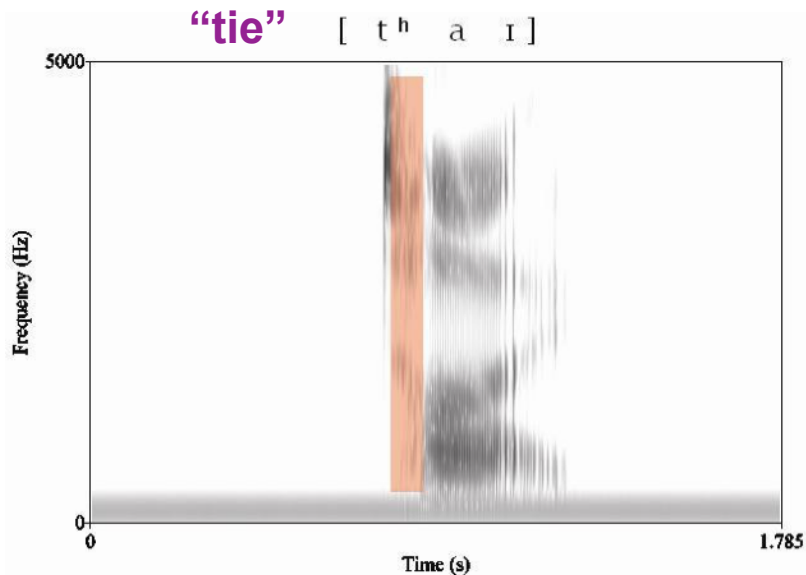
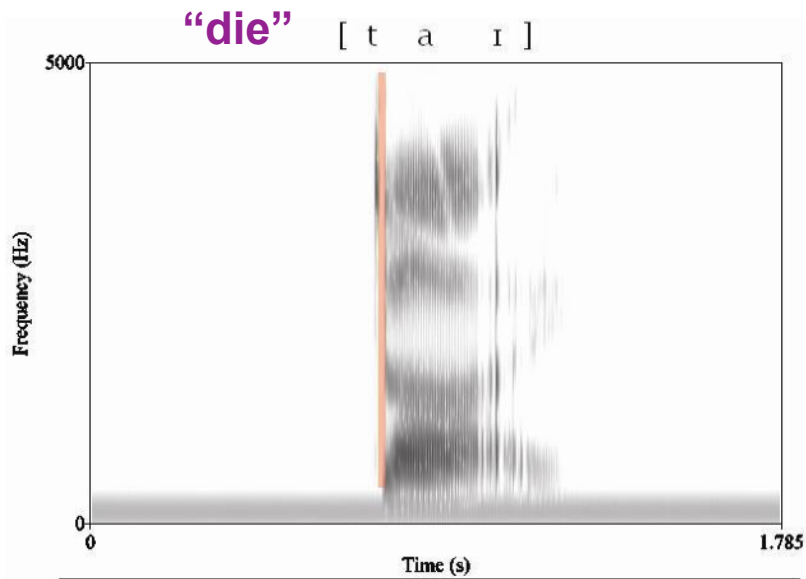
# A minimal problem in speech perception



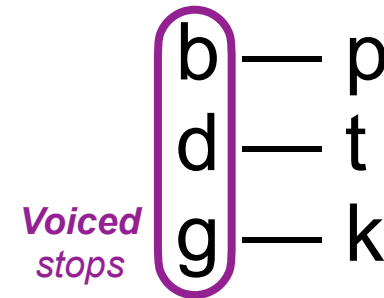
- This distinction involves a single “phonetic dimension”, **Voice Onset Time (VOT)**

b — p  
d — t  
g — k

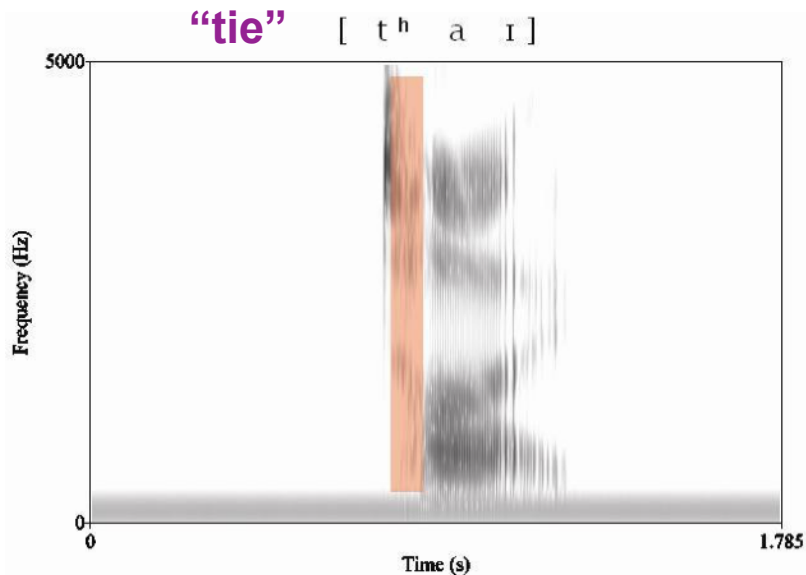
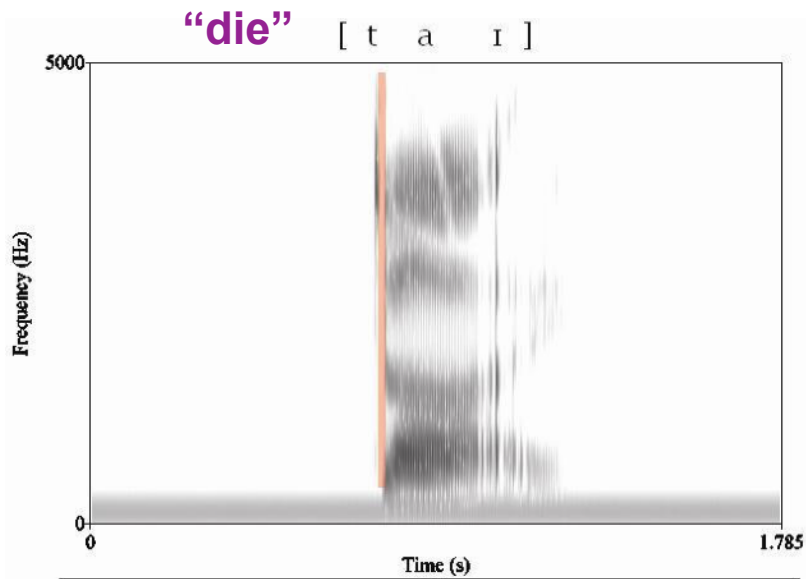
# A minimal problem in speech perception



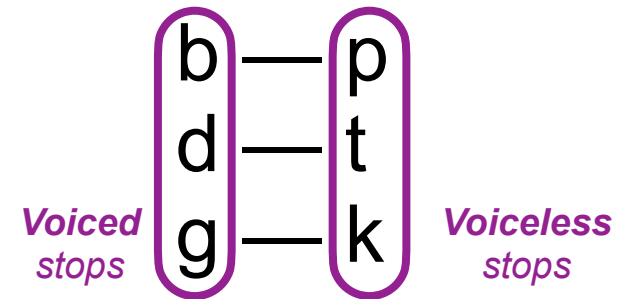
- This distinction involves a single “phonetic dimension”, **Voice Onset Time (VOT)**



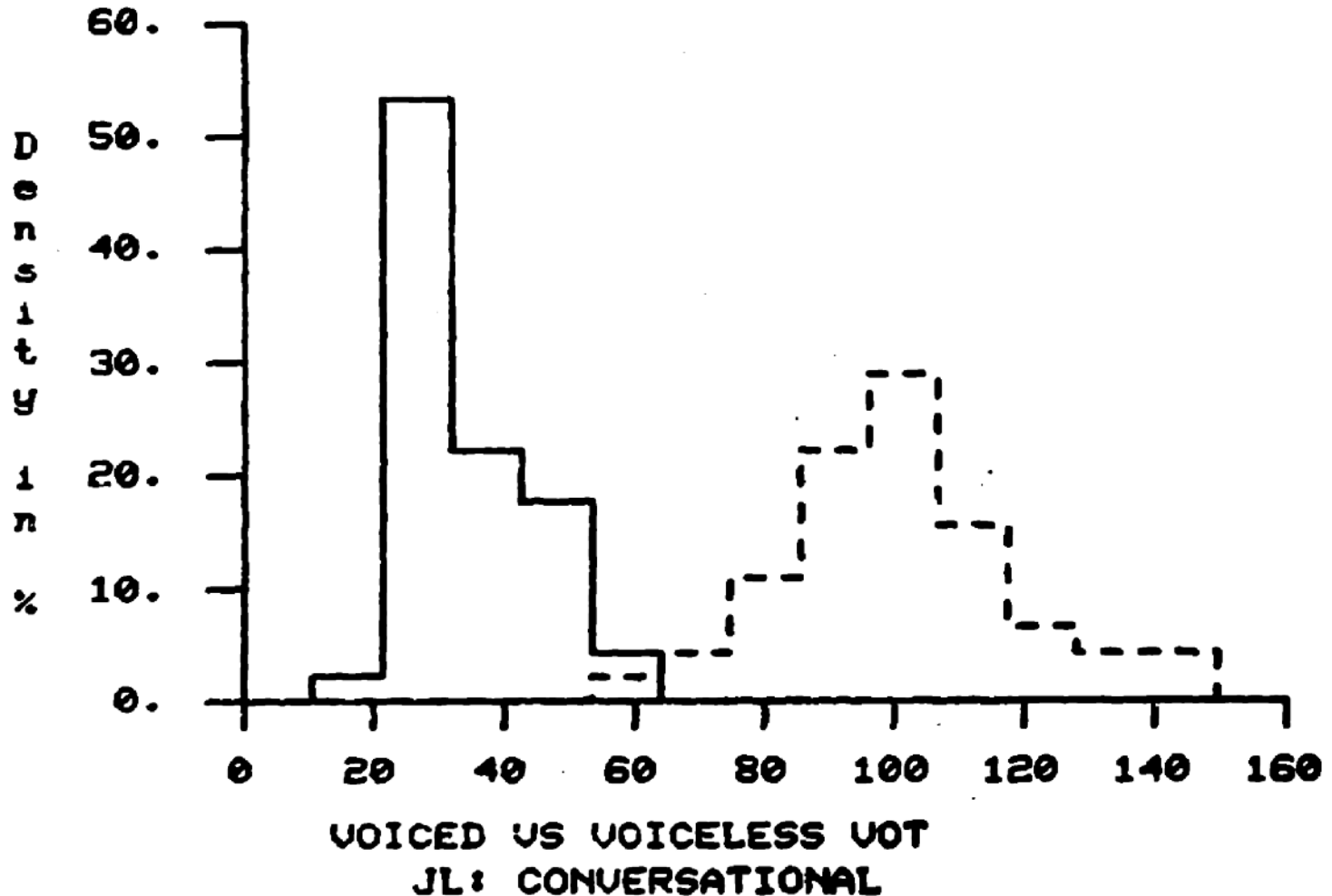
# A minimal problem in speech perception



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



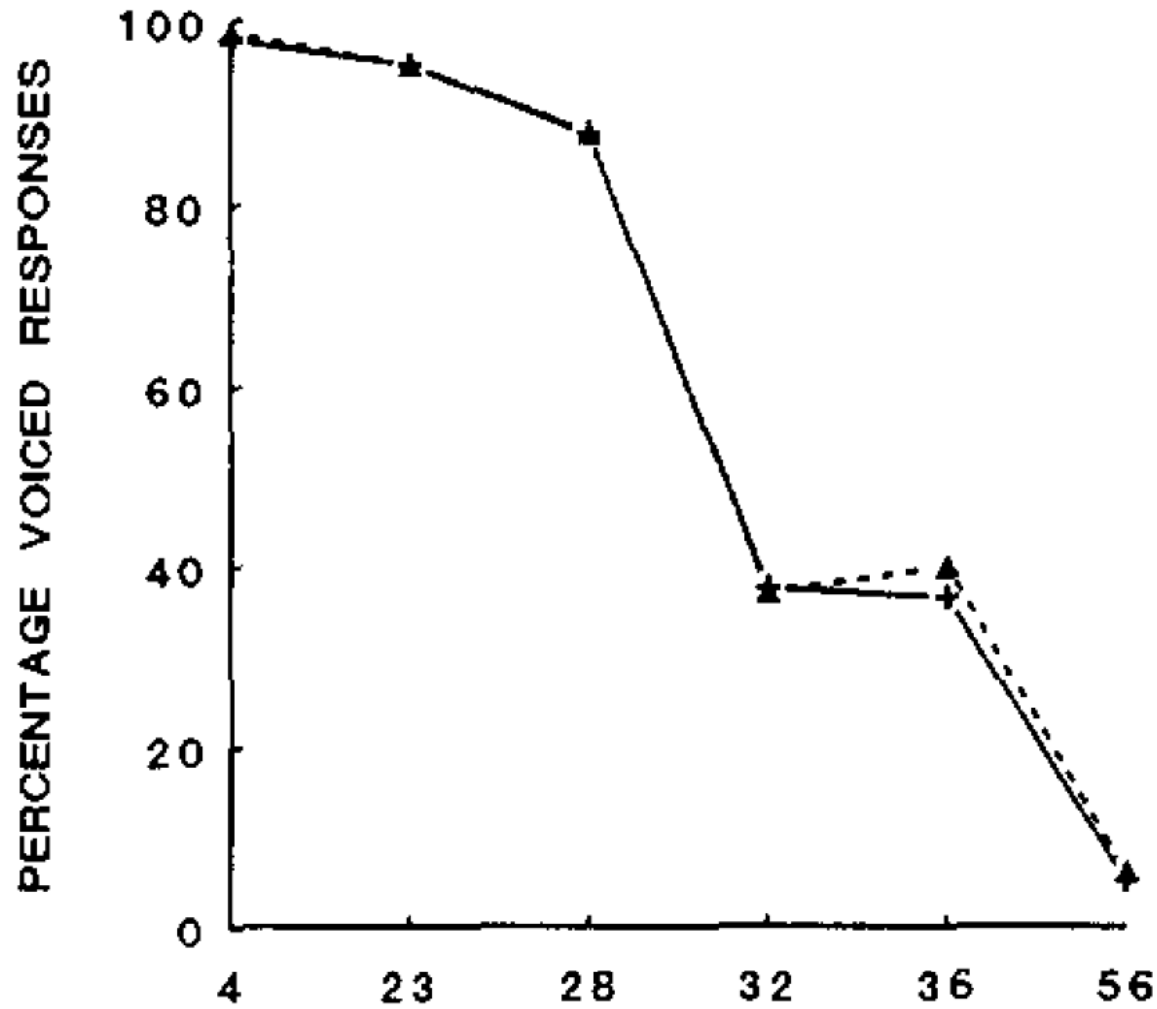
# Distribution of VOTs



(Chen, 1980)

# Human Categorization Curve

---

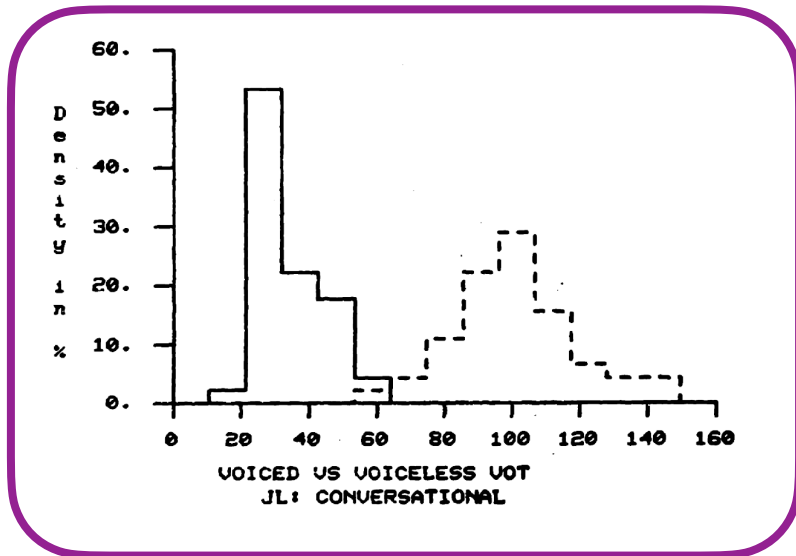


**b**

VOICE ONSET TIME (MS)

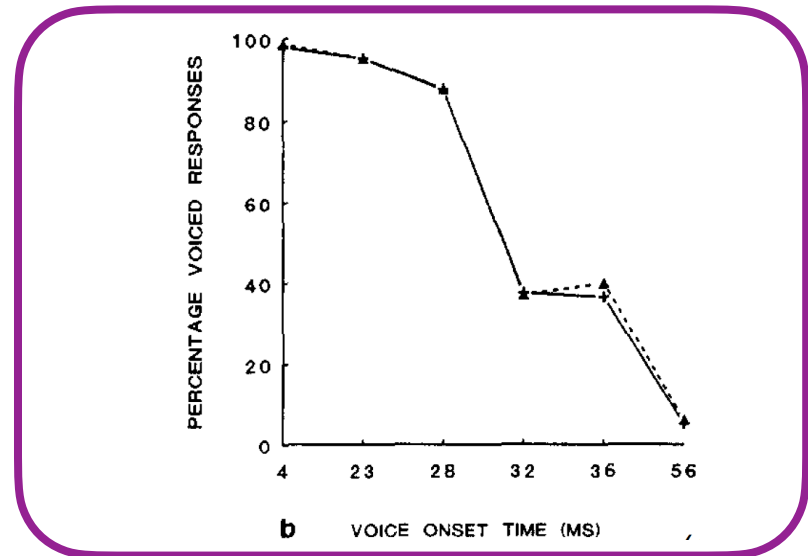
(Connine et al., 1991)

*Generative, or forward,  
distributional knowledge*



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

*Classification ("reverse-engineering")  
behavior*



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

How can we reconcile these two distributions?



# Rational analysis

---

# Rational analysis

---

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

# 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

# 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

# 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

# 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

# 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

# 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



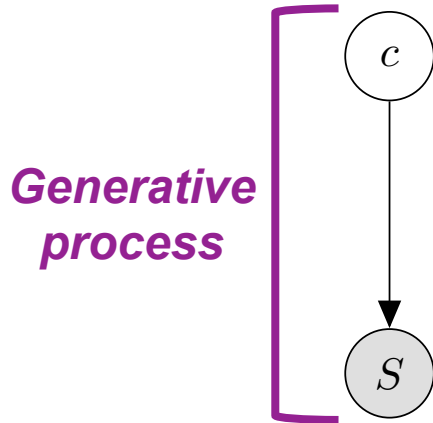
# Modeling VOT-based recognition

---

# Modeling VOT-based recognition

---

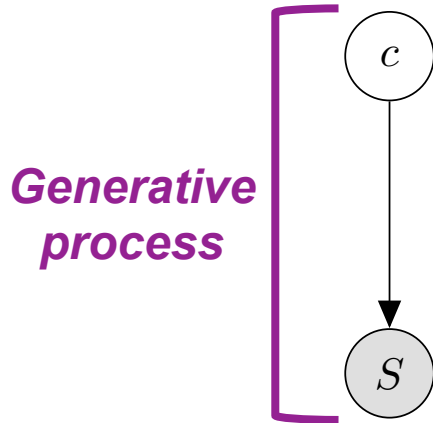
- Assume sound category  $c$  manifests as speech signal  $S$



# Modeling VOT-based recognition

---

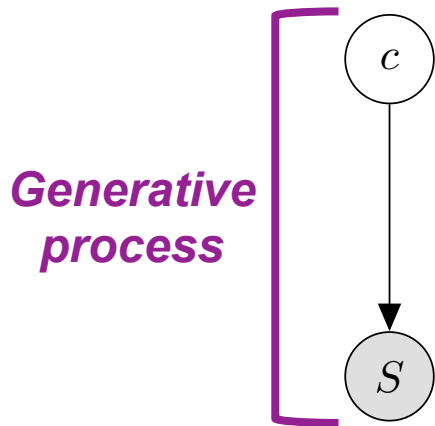
- Assume sound category  $c$  manifests as speech signal  $S$



- We can use Bayesian inference to infer  $c$  from  $S$

# Modeling VOT-based recognition

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

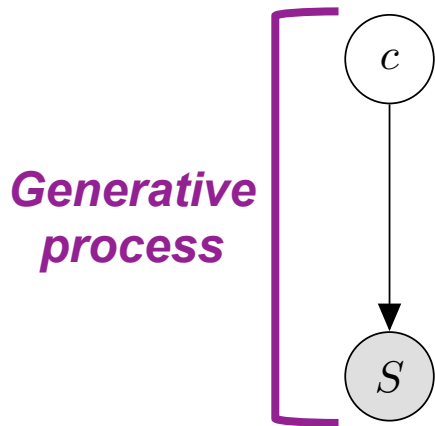


$$P(c|S) = \frac{P(S|c)P(c)}{\sum_{c'} P(S|c')P(c')}$$

- We can use Bayesian inference to infer  $c$  from  $S$

# Modeling VOT-based recognition

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

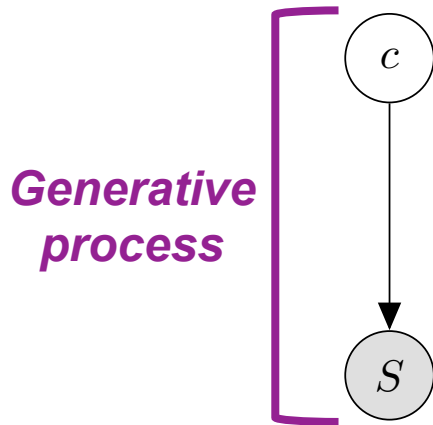


$$P(c|S) = \frac{P(S|c)P(c)}{\sum_{c'} P(S|c')P(c')}$$

- We can use Bayesian inference to infer  $c$  from  $S$
- To make this work, we need to choose:

# Modeling VOT-based recognition

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

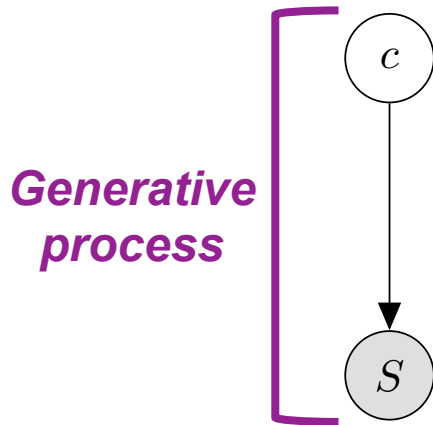


$$P(c|S) = \frac{P(S|c)P(c)}{\sum_{c'} P(S|c')P(c')}$$

- 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**  $P(S|c)$

# Modeling VOT-based recognition

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



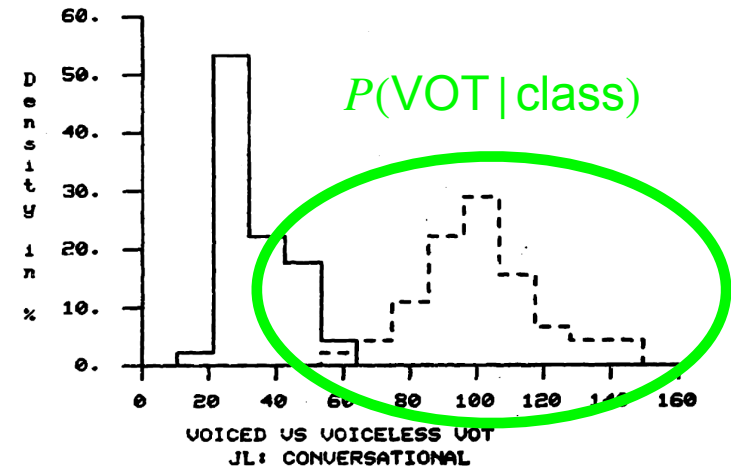
$$P(c|S) = \frac{P(S|c)P(c)}{\sum_{c'} P(S|c')P(c')}$$

- 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**

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

# Likelihood for a phonetic dimension

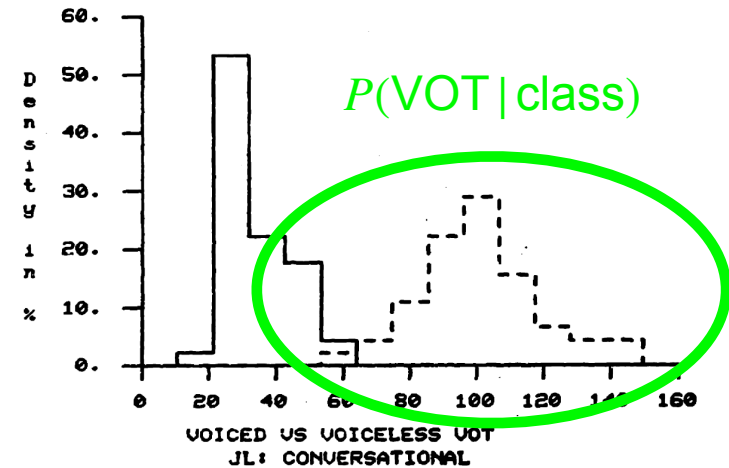
---





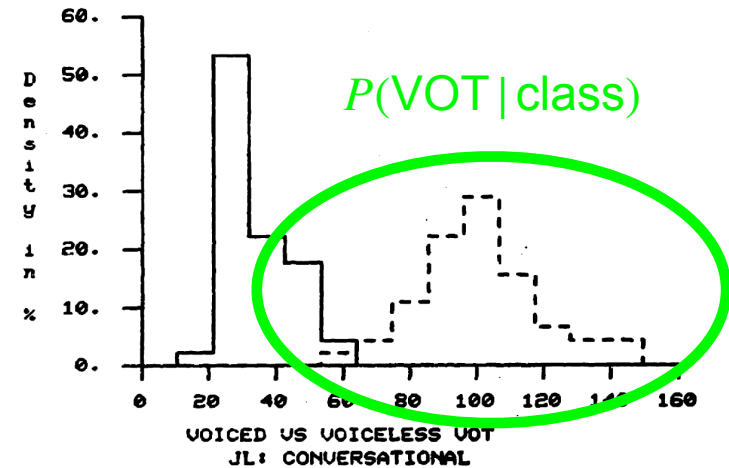
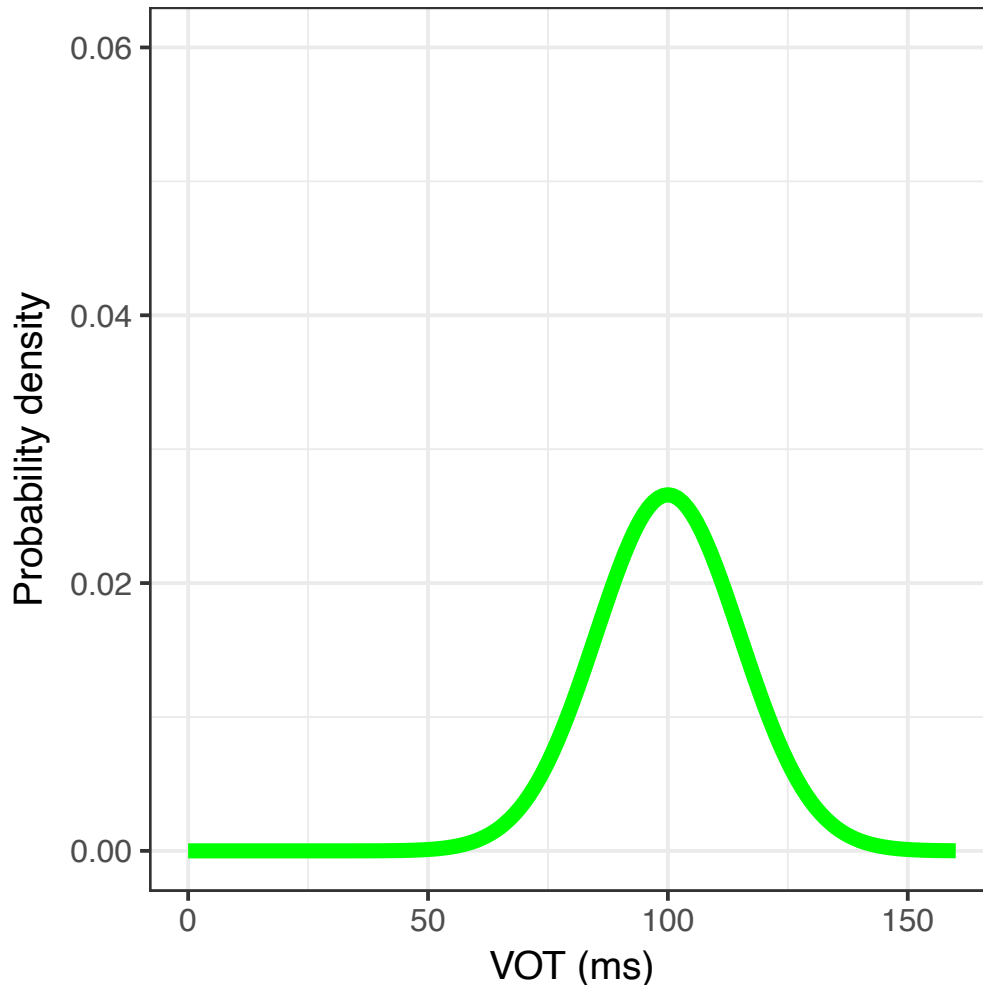
# Likelihood for a phonetic dimension

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



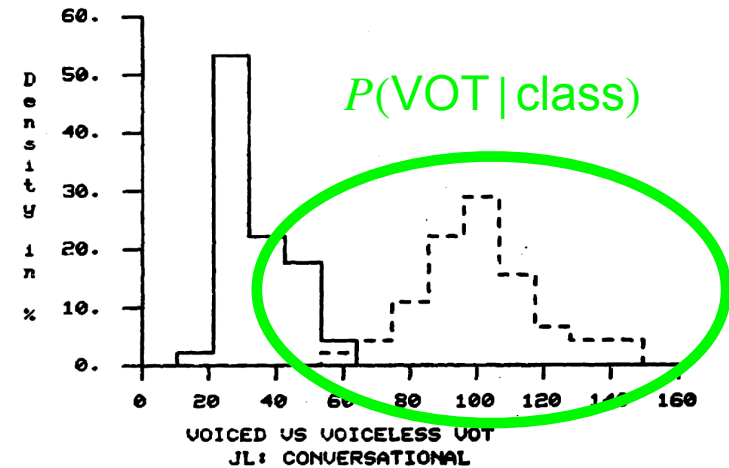
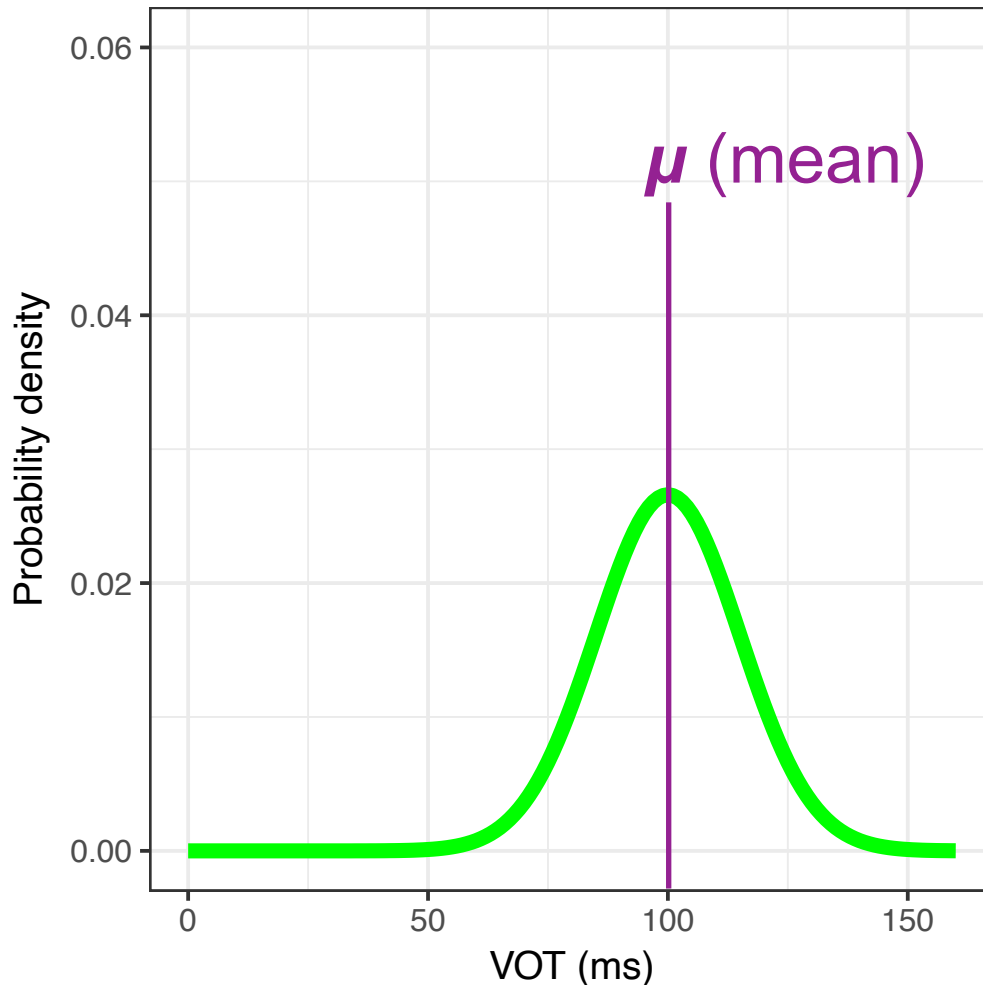
# Likelihood for a phonetic dimension

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



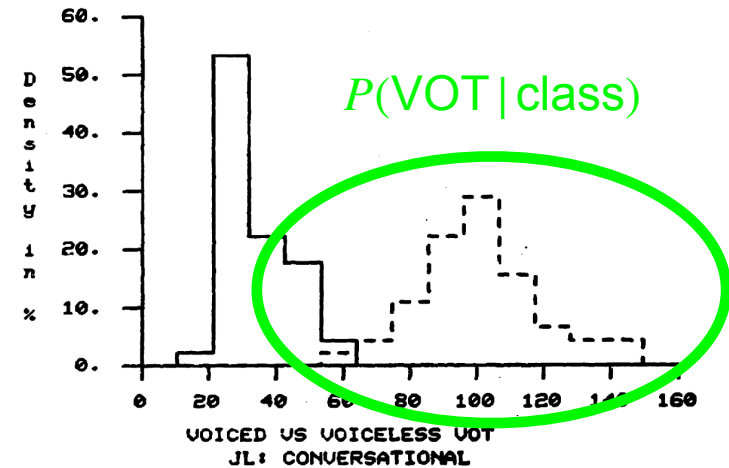
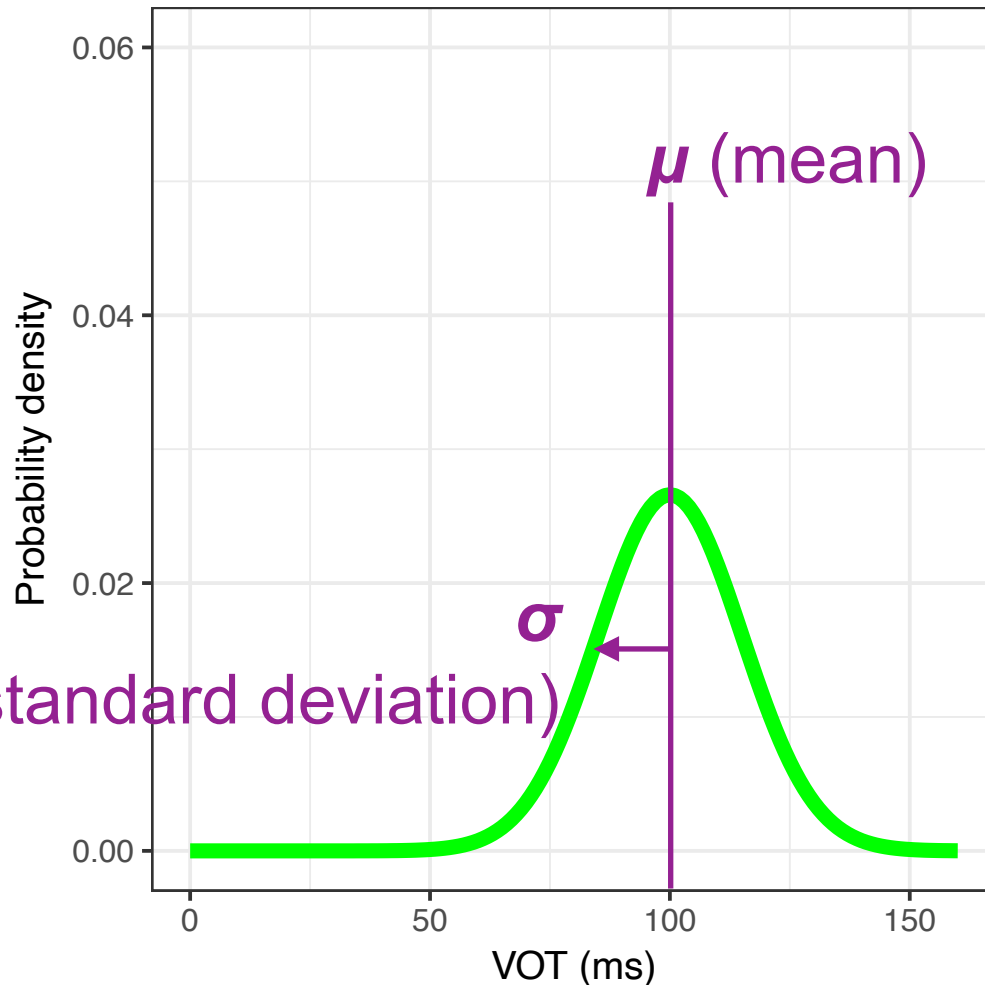
# Likelihood for a phonetic dimension

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



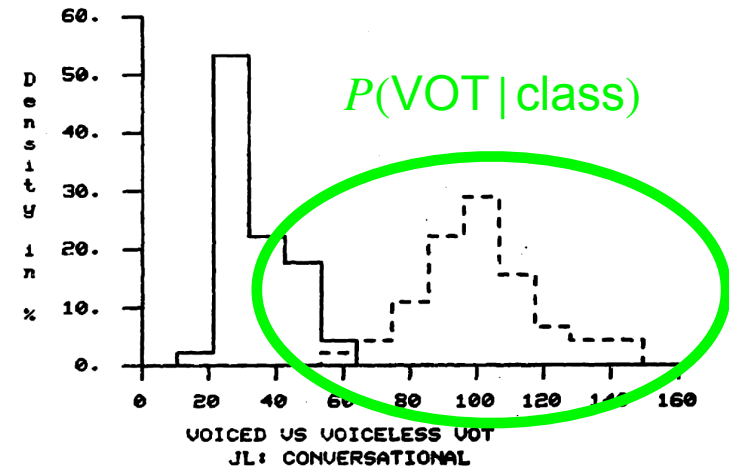
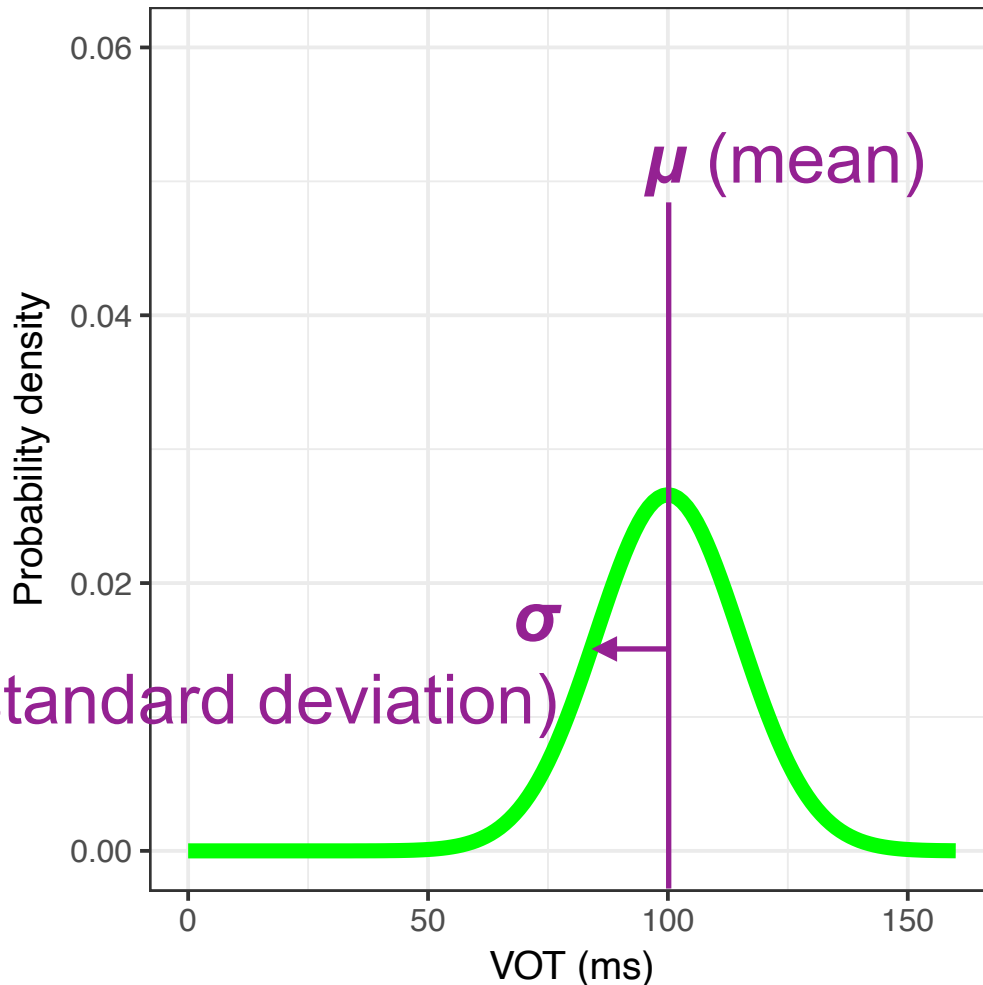
# Likelihood for a phonetic dimension

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



# Likelihood for a phonetic dimension

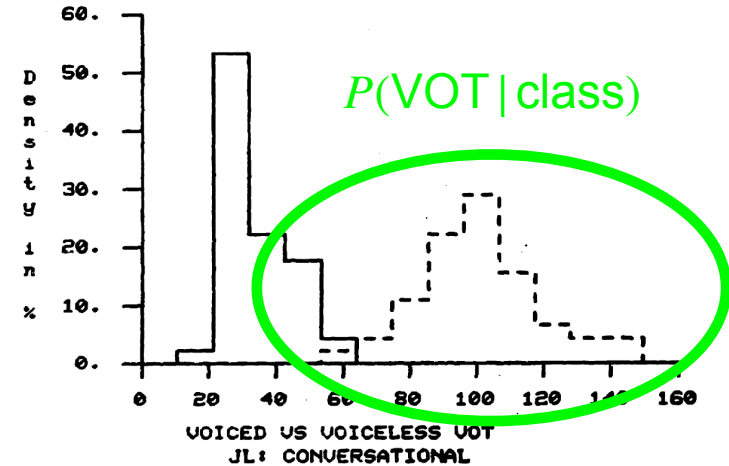
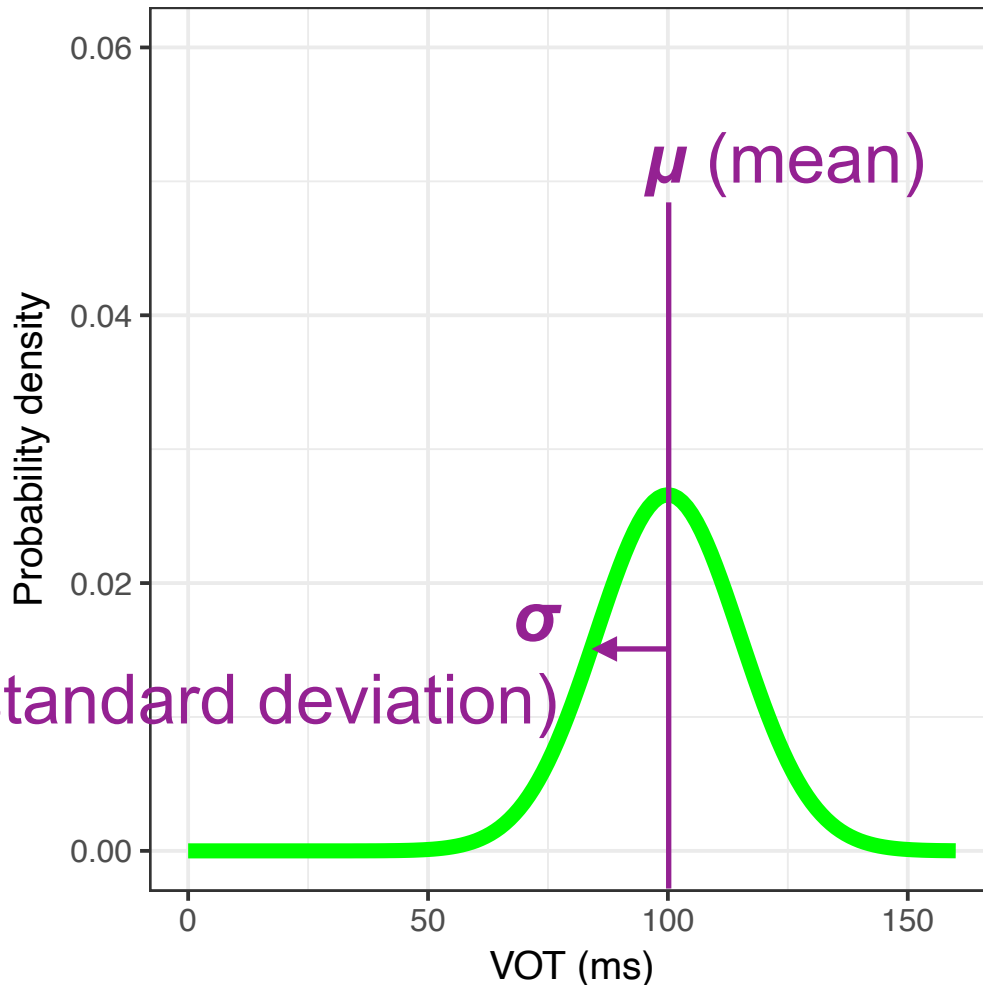
- The *normal* (a.k.a. *Gaussian*) *distribution* is a reasonable proxy



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

# Likelihood for a phonetic dimension

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

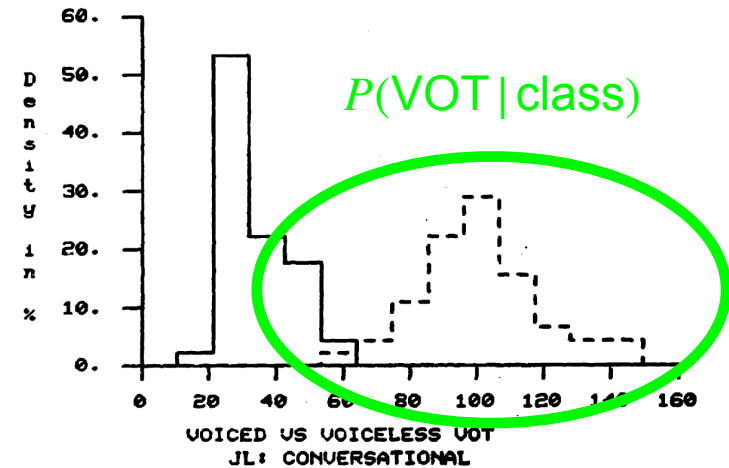
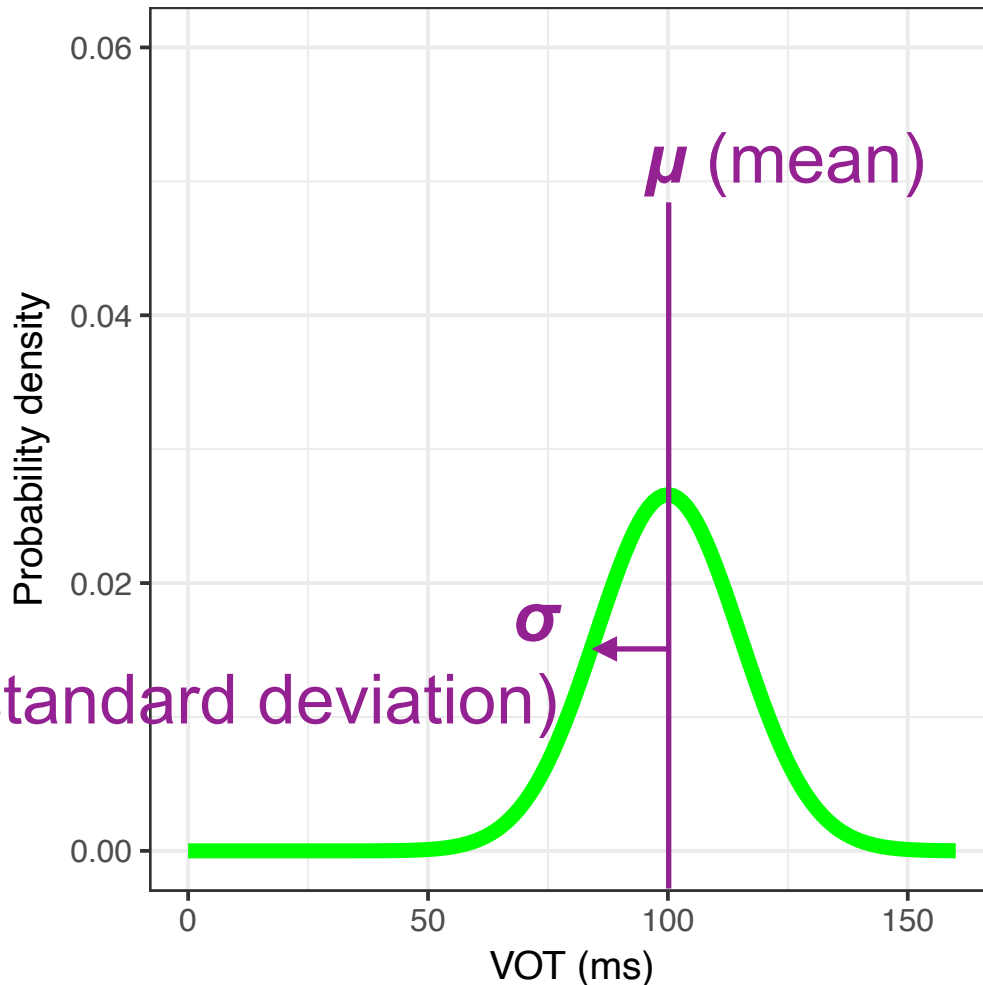


*Squared deviation  
from mean*

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

# Likelihood for a phonetic dimension

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



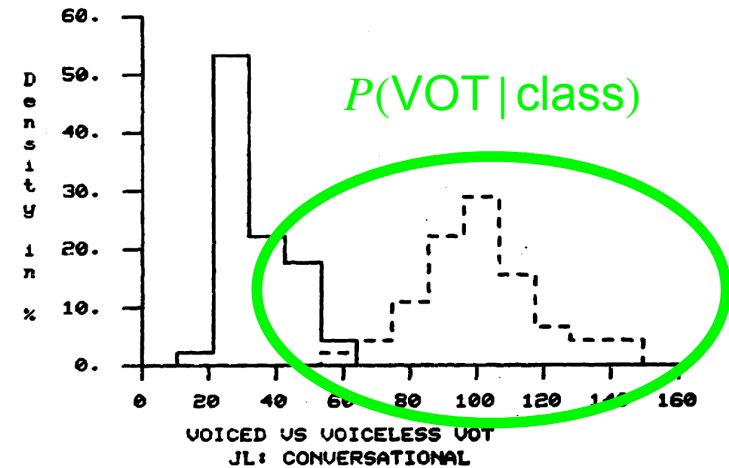
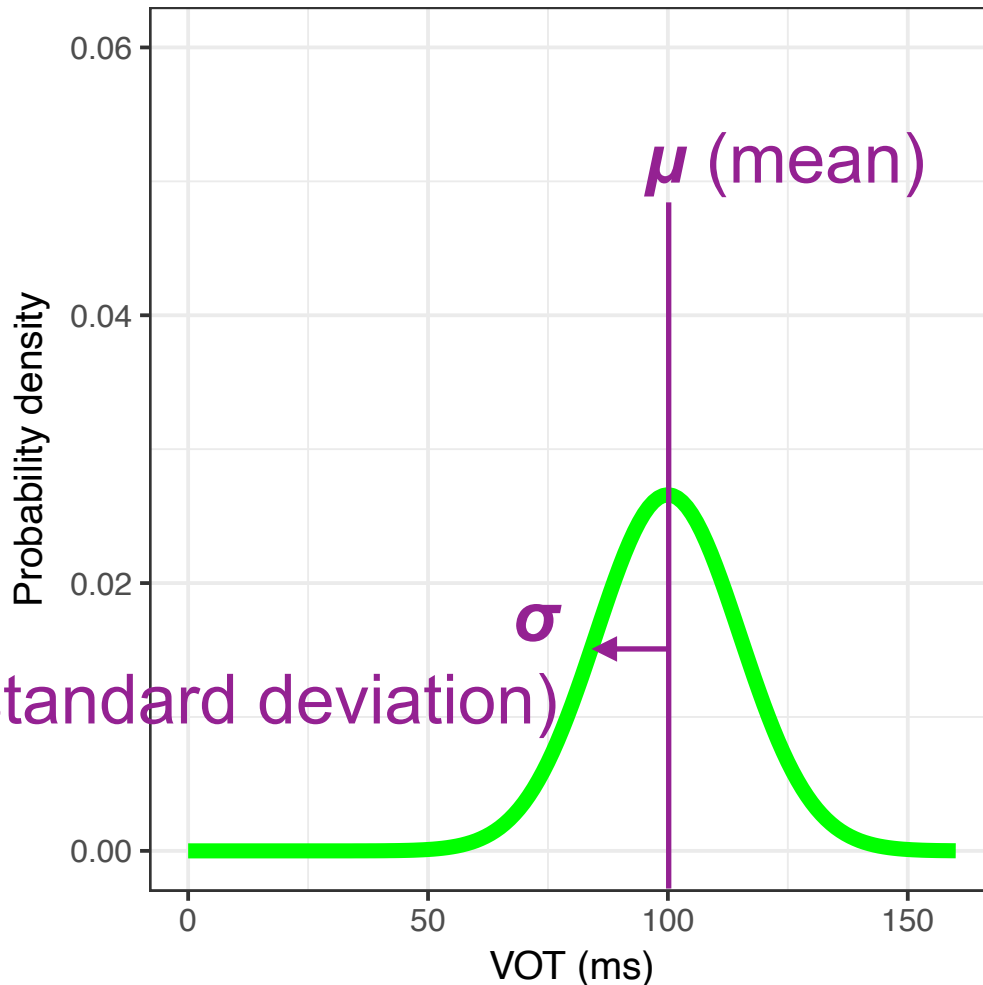
*Squared deviation  
from mean*

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

*Scaled by  
variance*

# Likelihood for a phonetic dimension

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



Squared deviation from mean

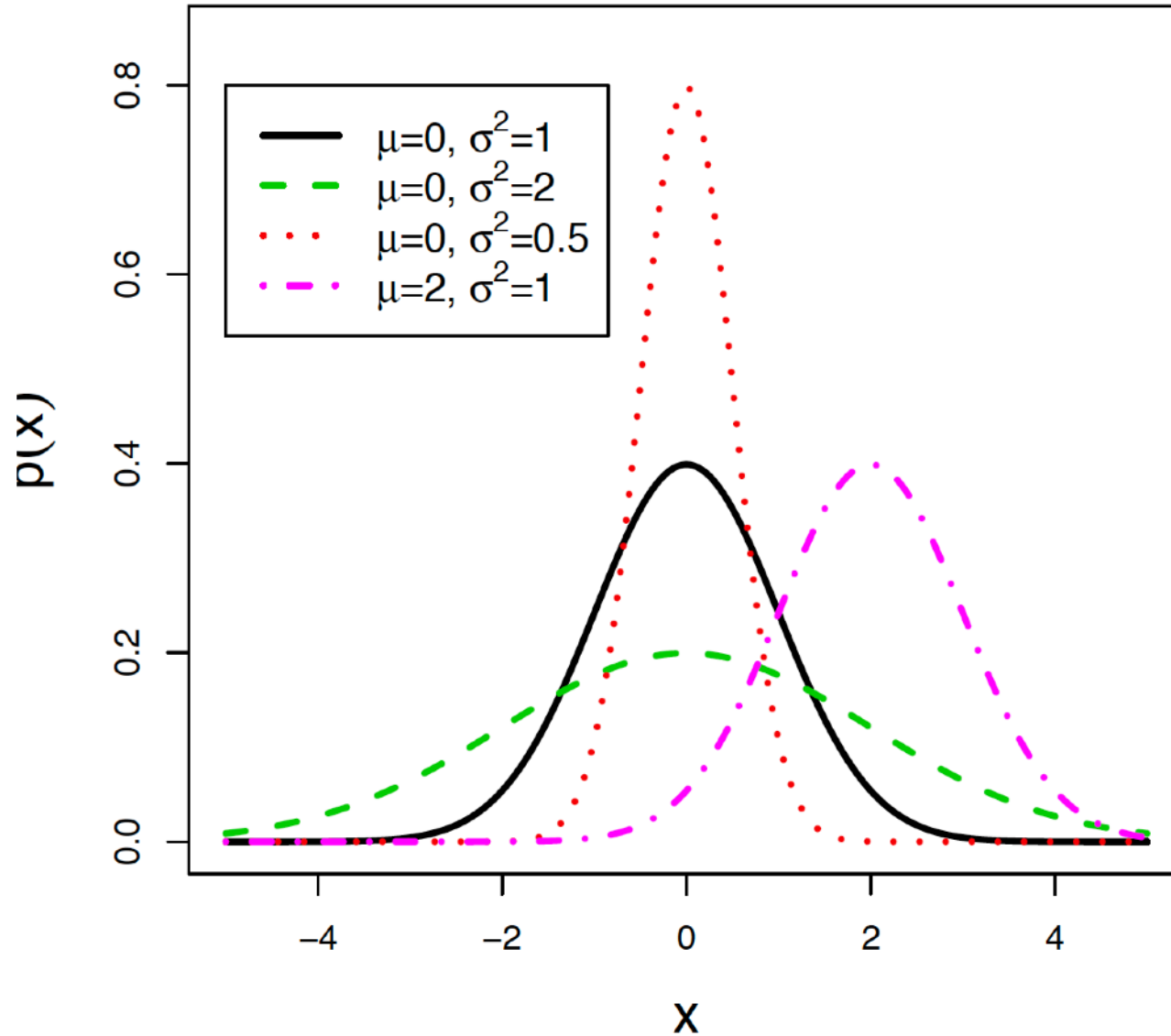
$$p(x | \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ \frac{-(x - \mu)^2}{2\sigma^2} \right]$$

(normalizing constant)

Scaled by variance



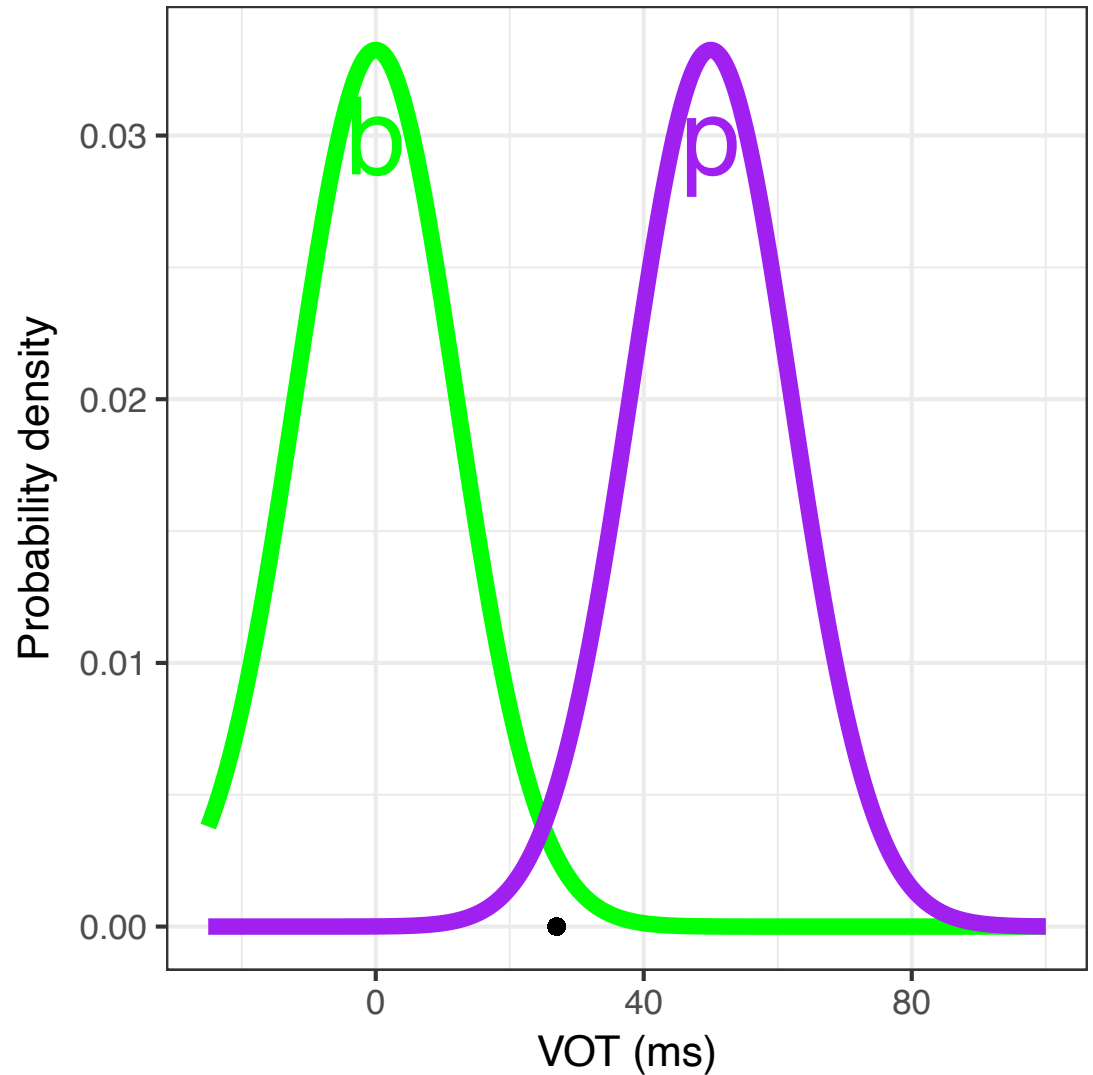
# Gaussian parameters



# Inference of phonetic category

---

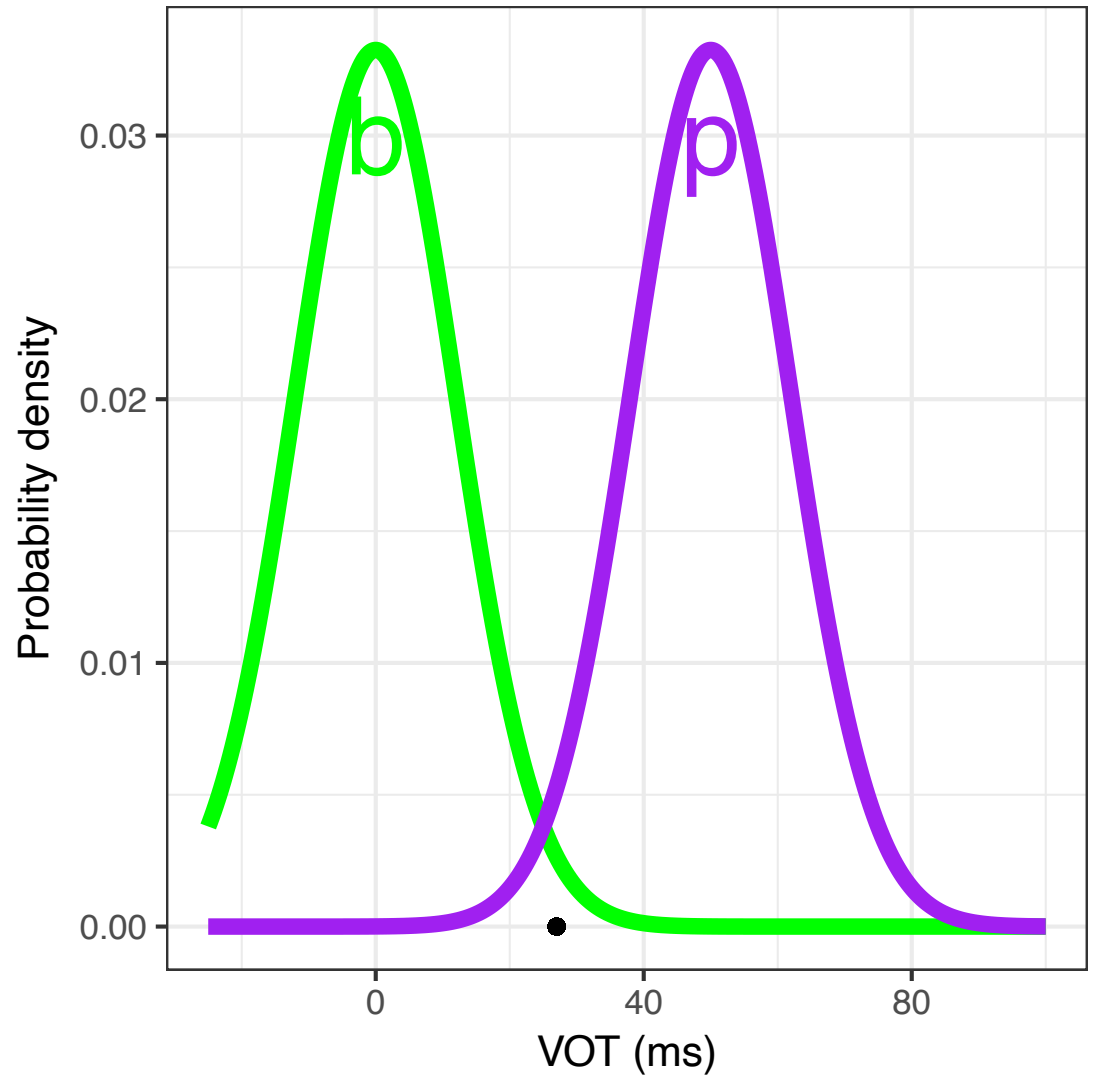
$$P(c|S) \propto P(S|c)P(c)$$



# Inference of phonetic category

---

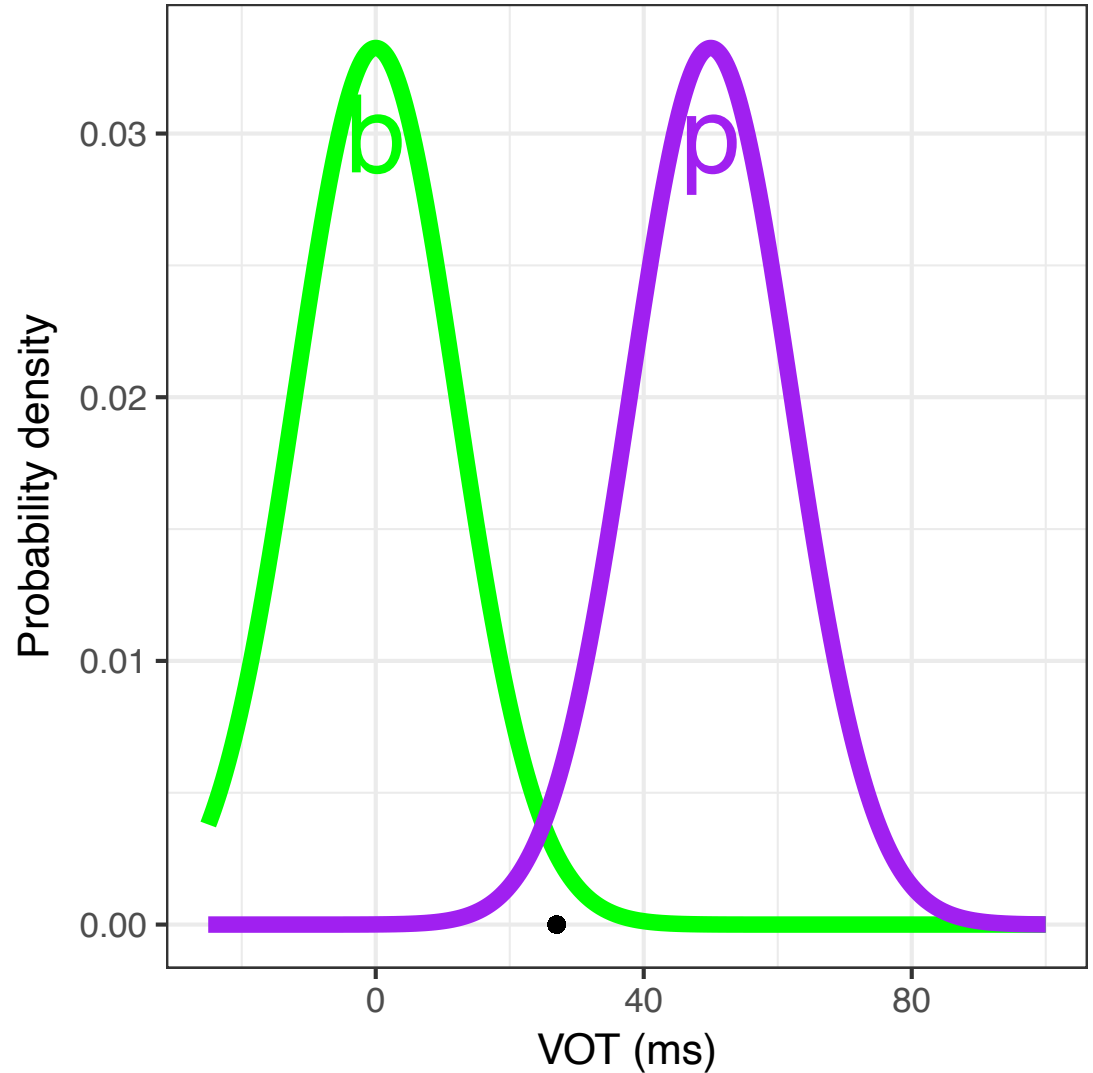
$$P(c|S) \propto P(S|c) \overline{P(c)}$$



# Inference of phonetic category

$$\underline{P(/b/) = P(/p/) = 0.5}$$

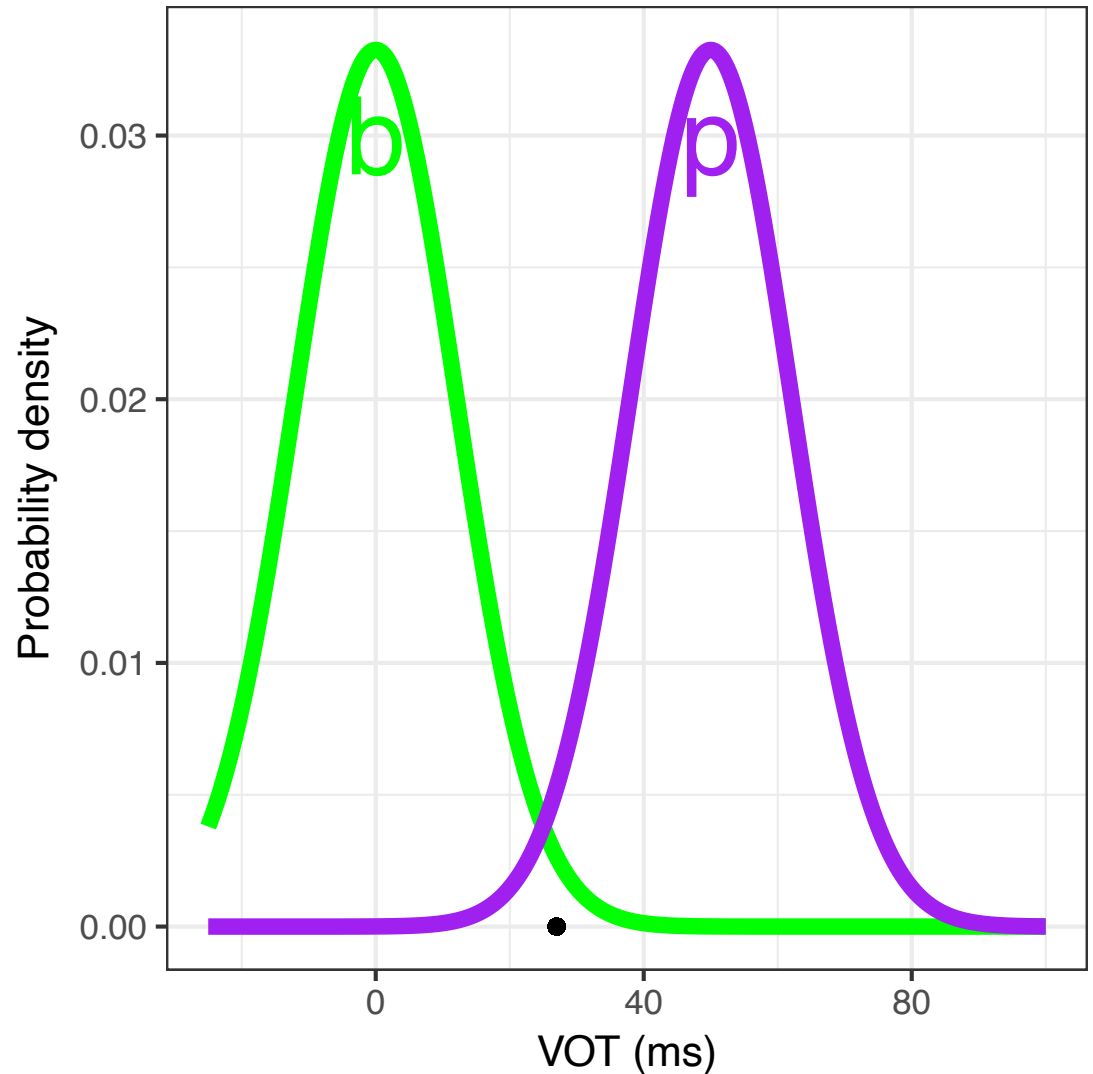
$$P(c|S) \propto P(S|c) \underline{P(c)}$$



# Inference of phonetic category

$$P(c|S) \propto \frac{P(S|c)P(c)}{P(/b/) + P(/p/)}$$

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



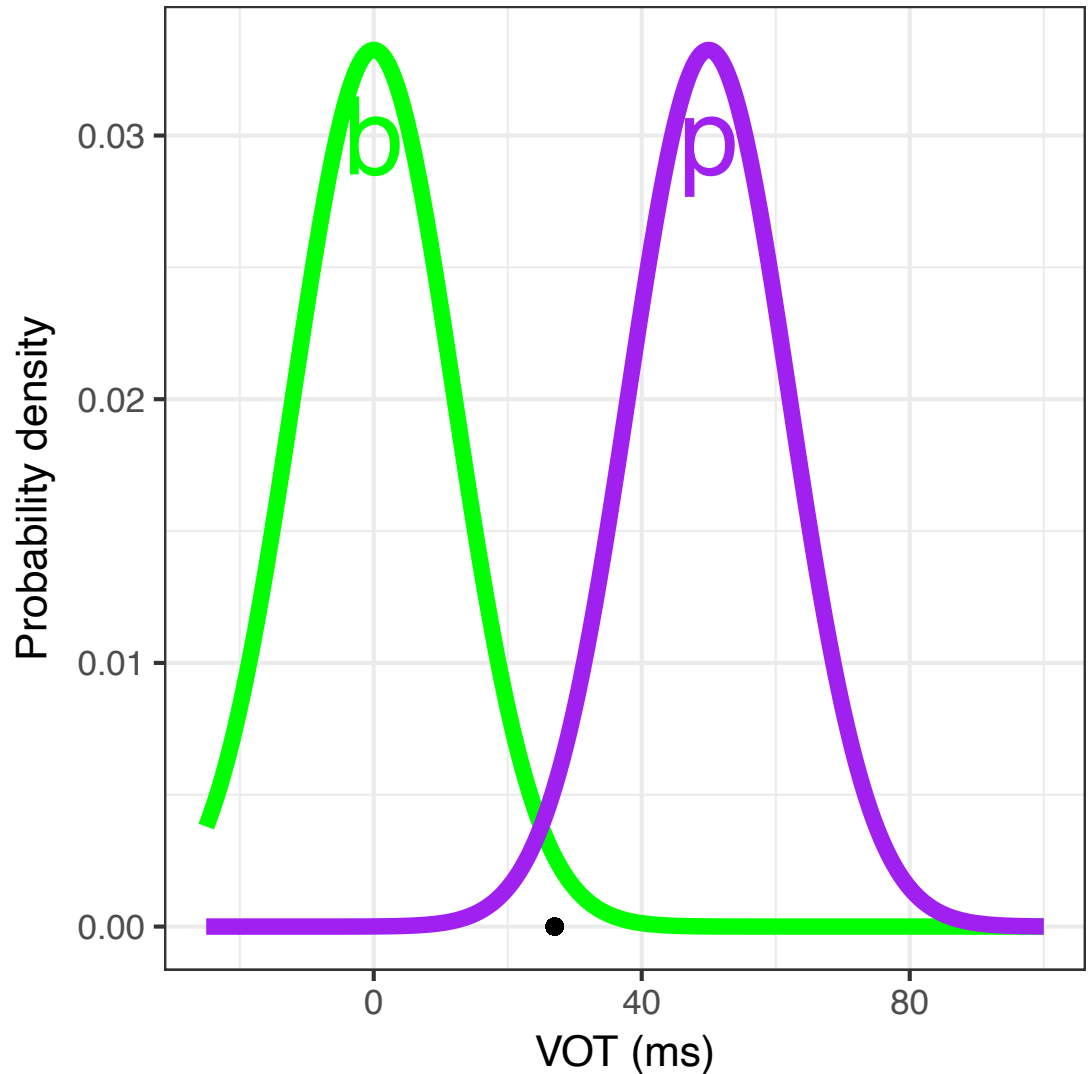
# Inference of phonetic category

$$P(c|S) \propto \frac{P(S|c)P(c)}{P(S)}$$

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

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

“S is normally distributed with mean  $\mu_c$  and variance  $\sigma^2$ ”



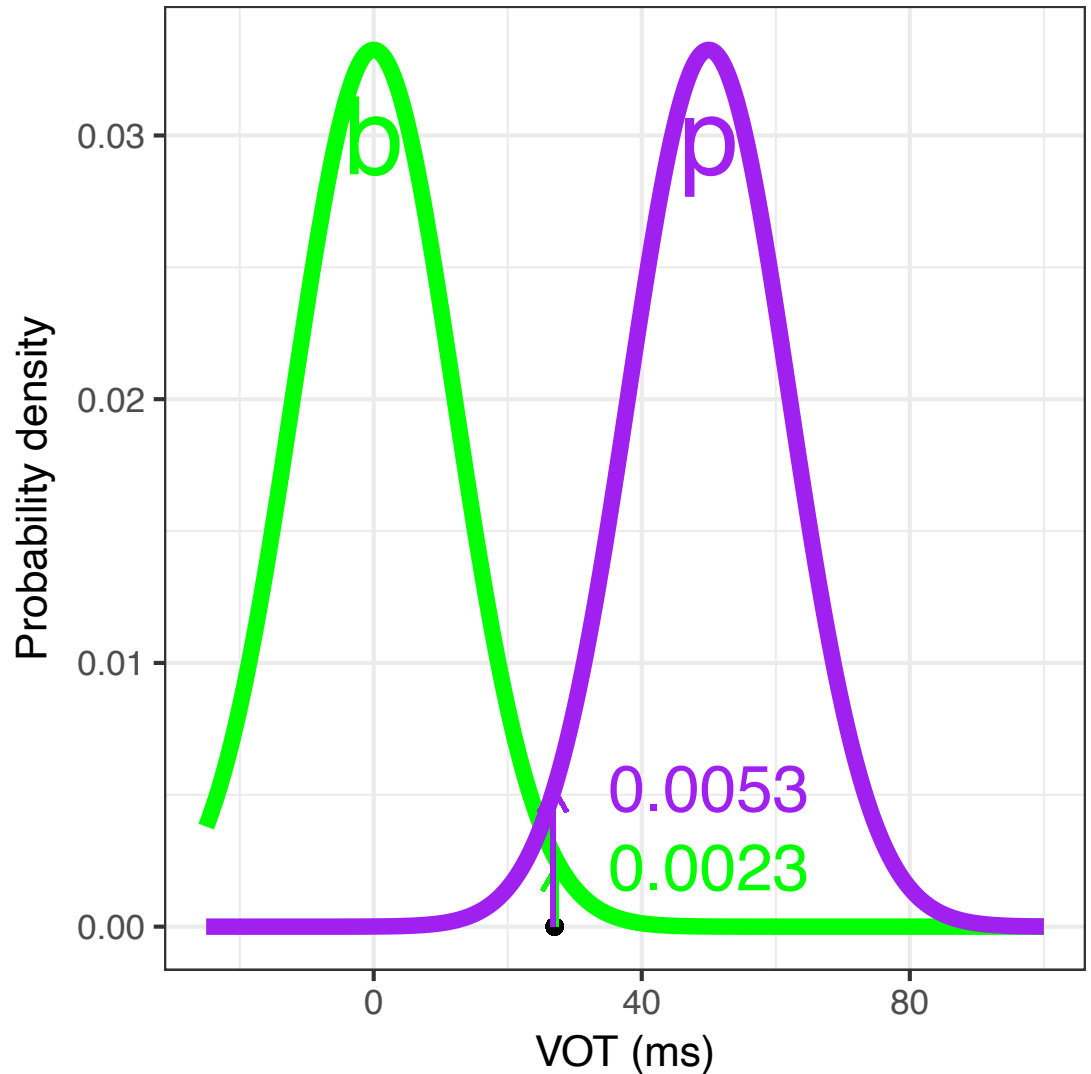
# Inference of phonetic category

$$P(c|S) \propto \frac{P(S|c)P(c)}{P(S)}$$

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

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

“S is normally distributed with mean  $\mu_c$  and variance  $\sigma^2$ ”



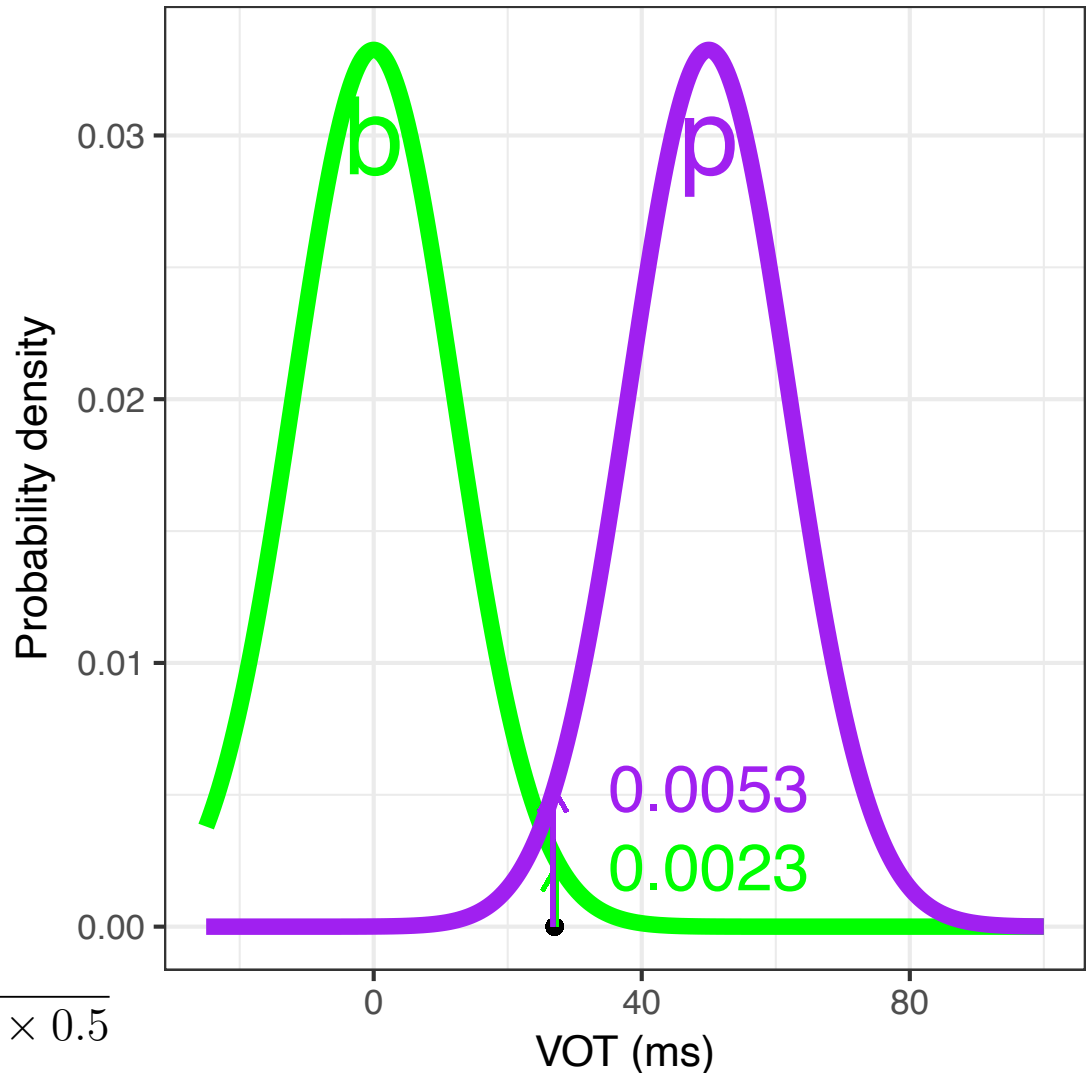
# Inference of phonetic category

$$\frac{P(/b/) = P(/p/) = 0.5}{}$$

$$P(c|S) \propto \frac{P(S|c)P(c)}{}$$

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

“S is normally distributed with mean  $\mu_c$  and variance  $\sigma^2$ ”



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



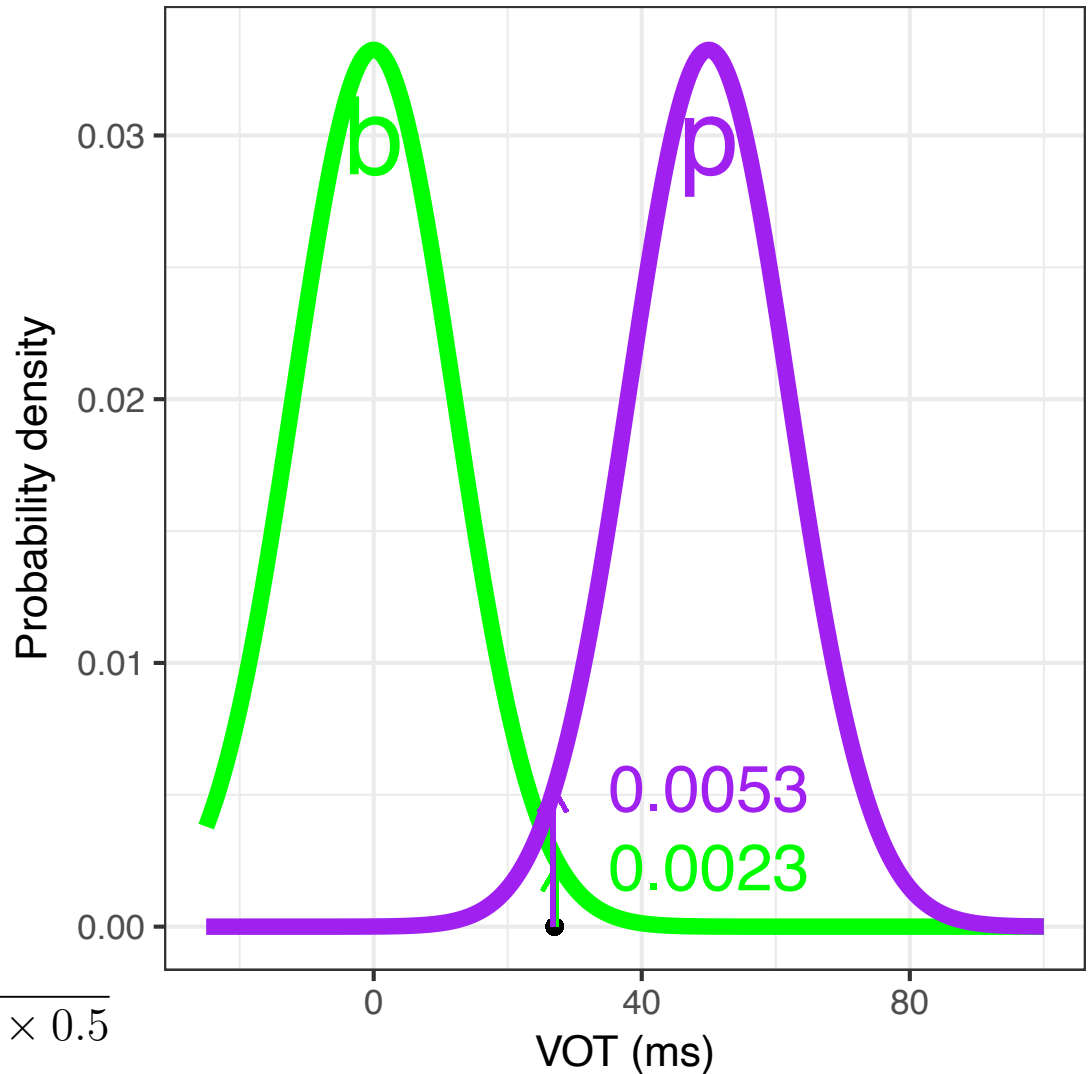
# Inference of phonetic category

$$\frac{P(/b/) = P(/p/) = 0.5}{}$$

$$P(c|S) \propto \frac{P(S|c)P(c)}{}$$

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

“S is normally distributed with mean  $\mu_c$  and variance  $\sigma^2$ ”

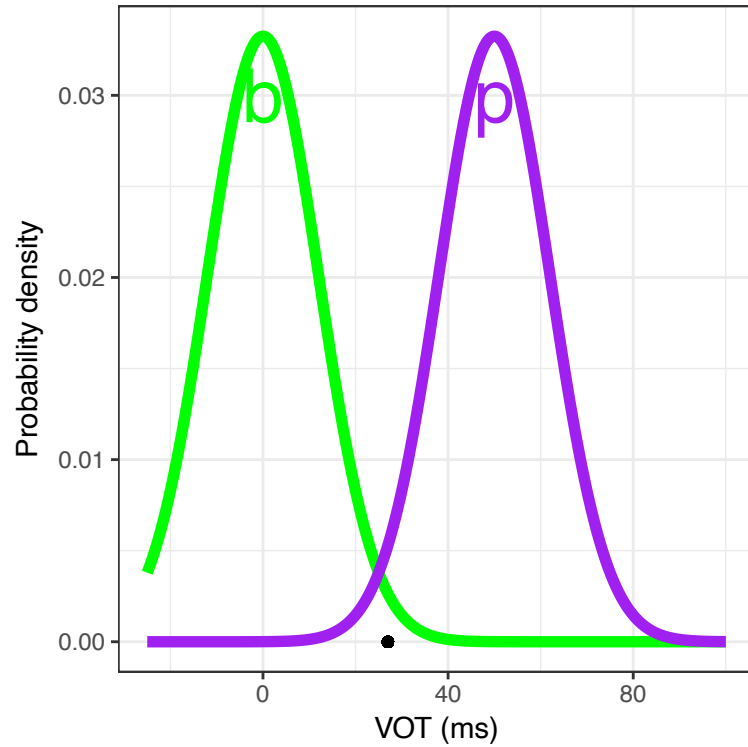


$$P(c = p|S) = \frac{0.0053 \times 0.5}{0.0053 \times 0.5 + 0.0023 \times 0.5}$$

$$= 0.697$$

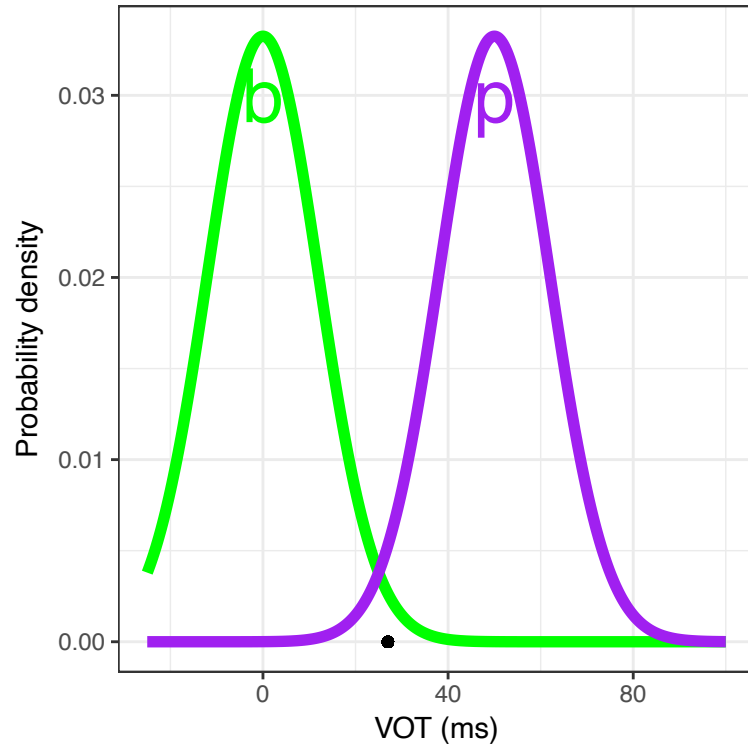
# Bayesian categorization curve

---

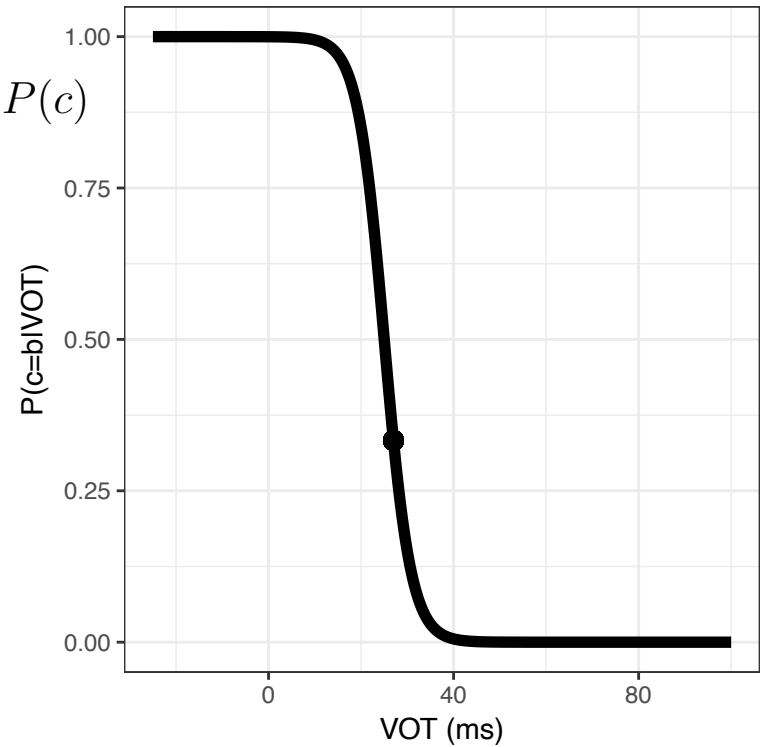


$$P(c|S) \propto P(S|c)P(c)$$

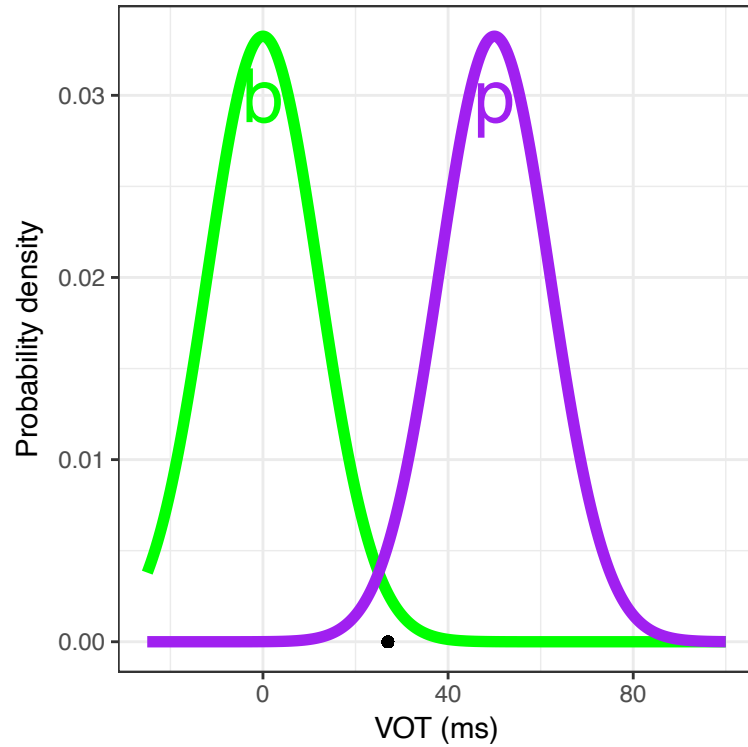
# Bayesian categorization curve



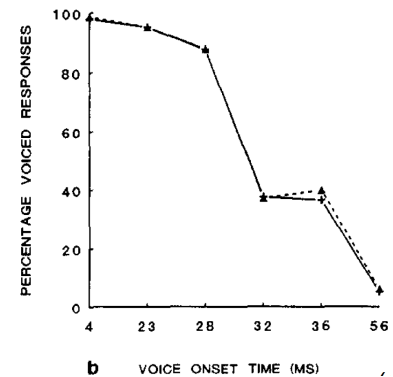
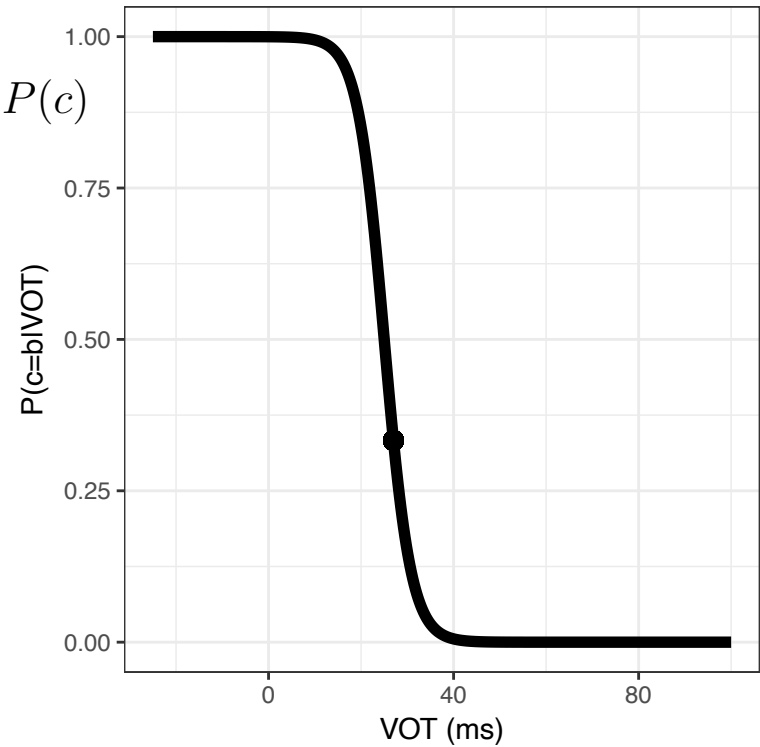
$$P(c|S) \propto P(S|c)P(c)$$



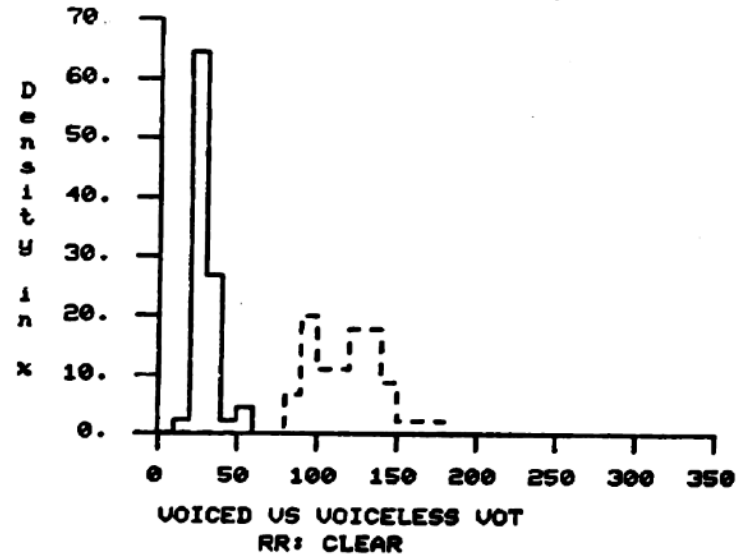
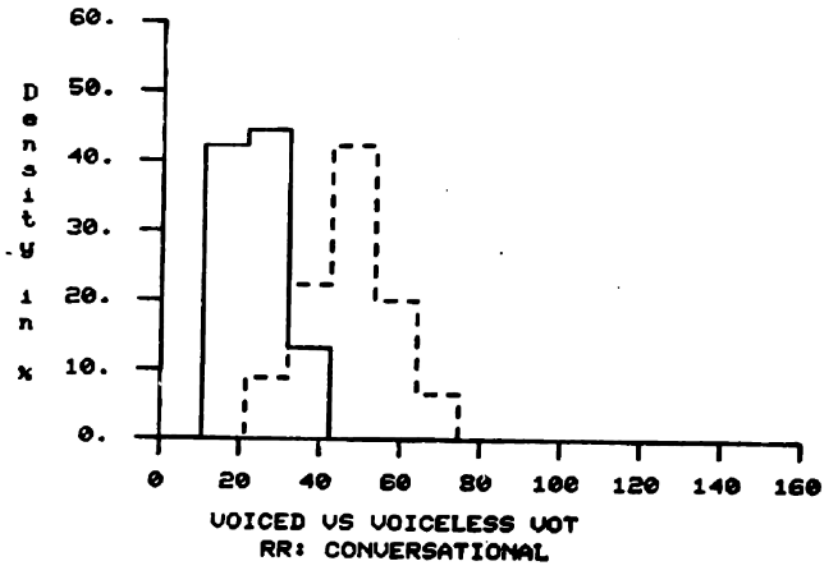
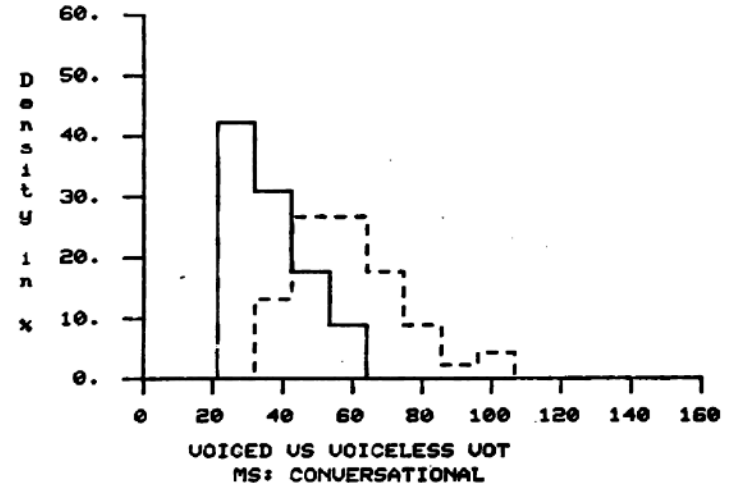
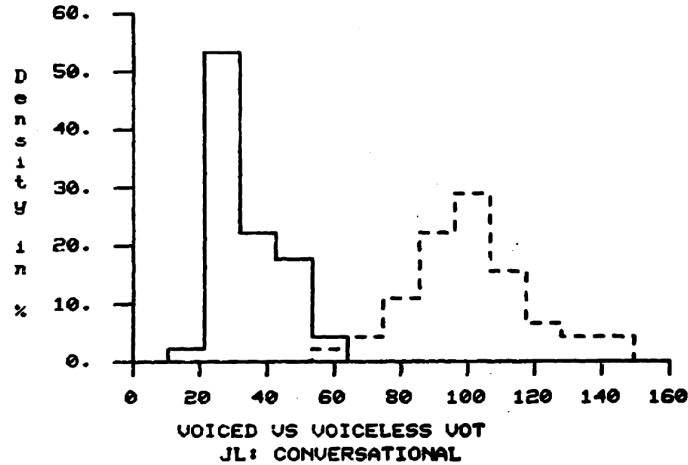
# Bayesian categorization curve



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



# Inter-speaker variability

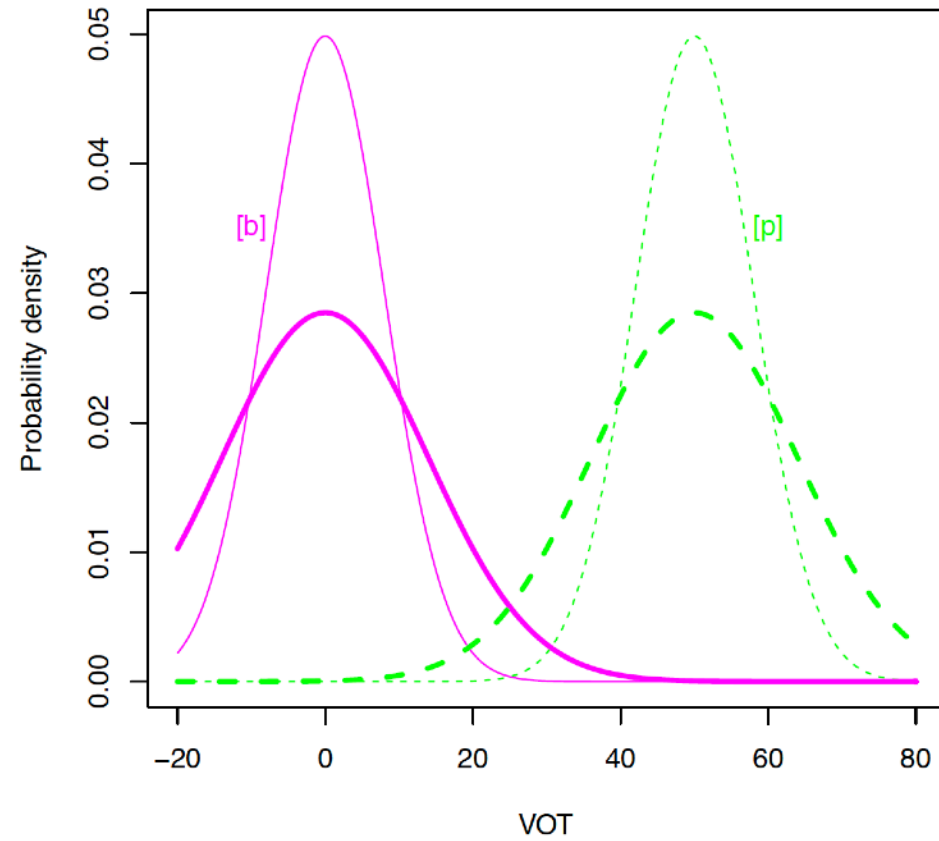


(Chen, 1980)

# Ideal observer predictions

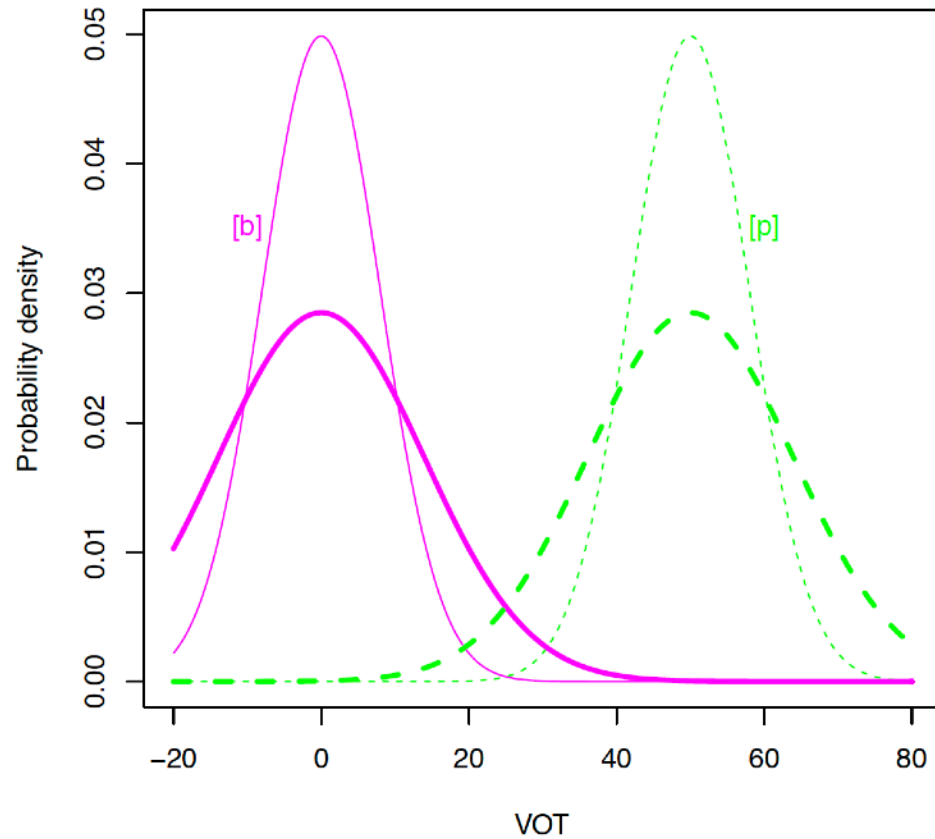
---

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

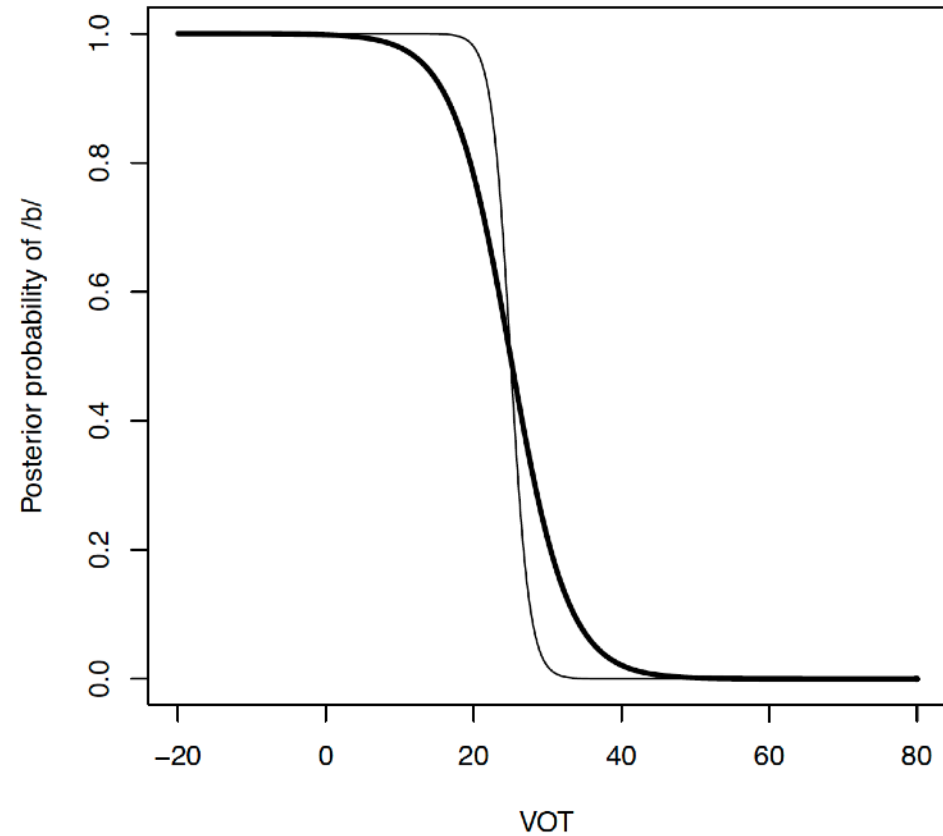


# Ideal observer predictions

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



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



# Testing effects of environment variability

---

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

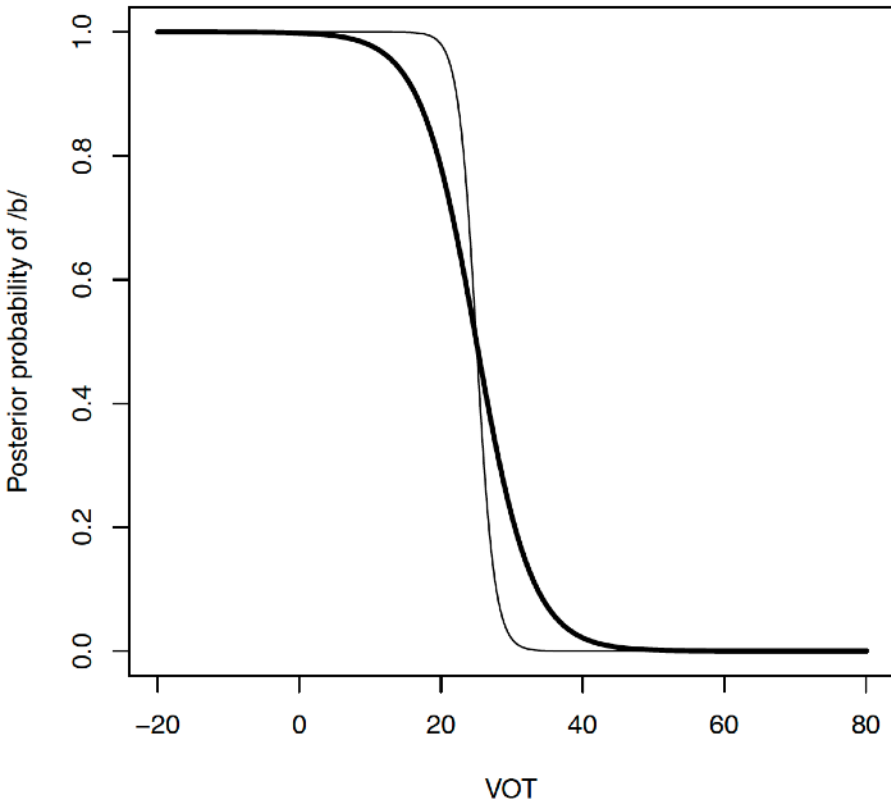


# Testing effects of environment variability

---

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

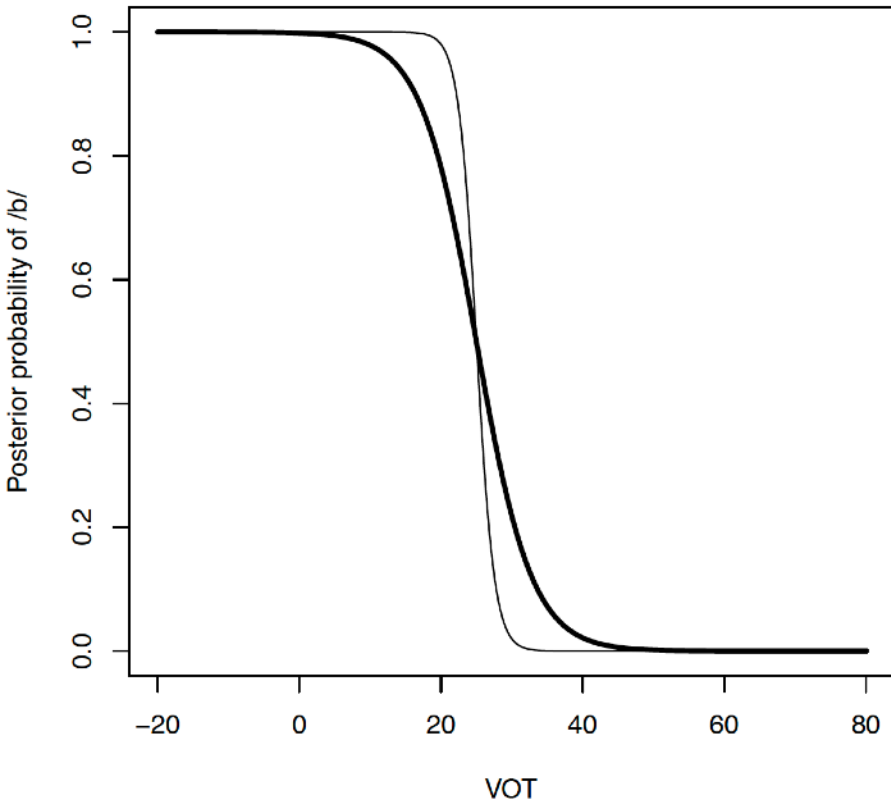
*Ideal observer*



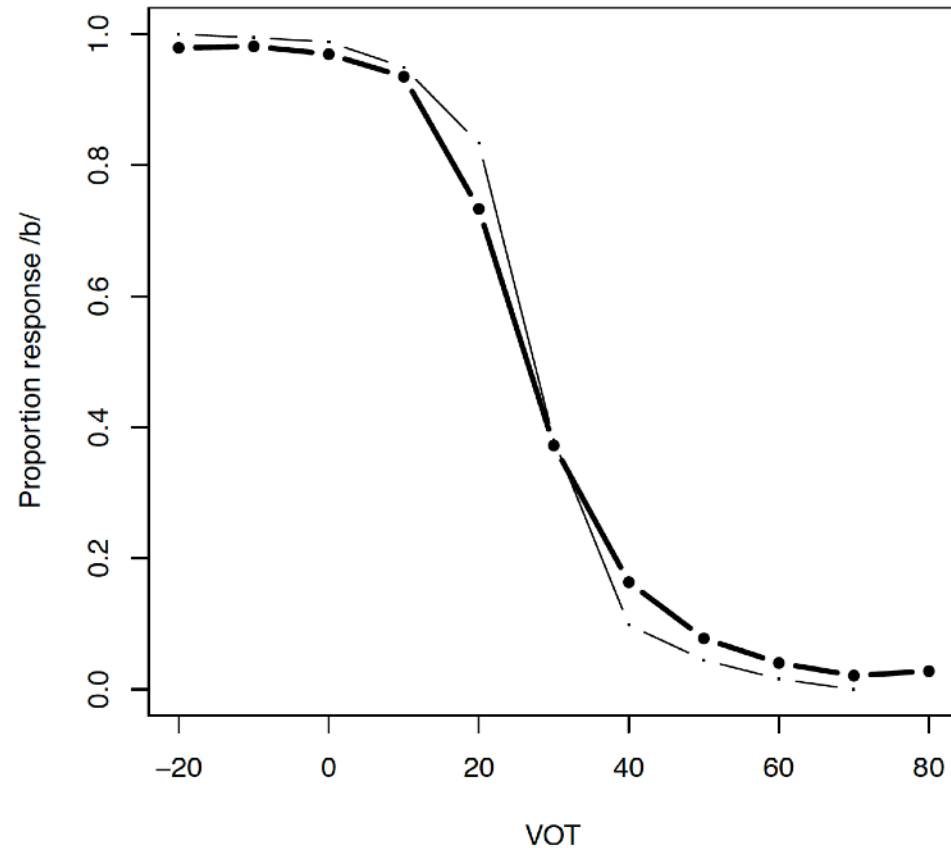
# Testing effects of environment variability

- 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

---

# Speech perception case study: summary

---

- We covered a simple case of **classification** in a continuous-signal setting: **phoneme identification**

# 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

# 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

# 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