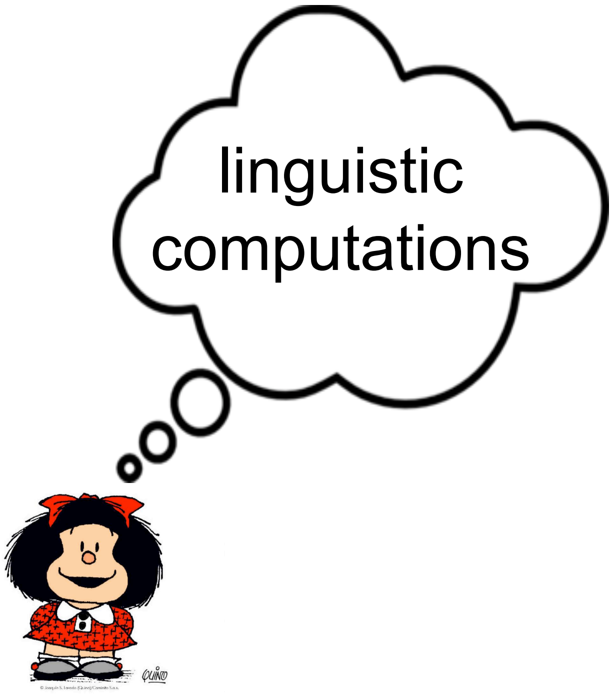
Relating Human Perceptual Data to Corpus Data through Cognitive Modeling

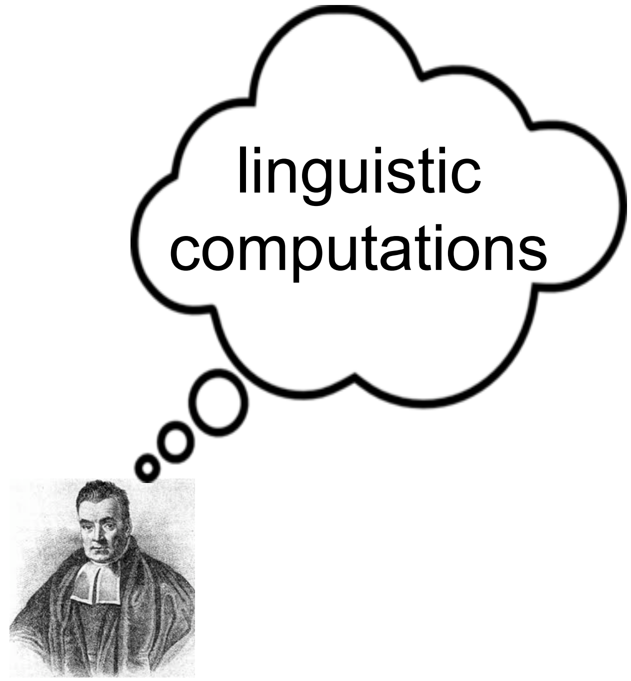
Naomi Feldman
University of Maryland

Fred Jelinek Memorial PIRE workshop
Prague, Czech Republic
July 22, 2014

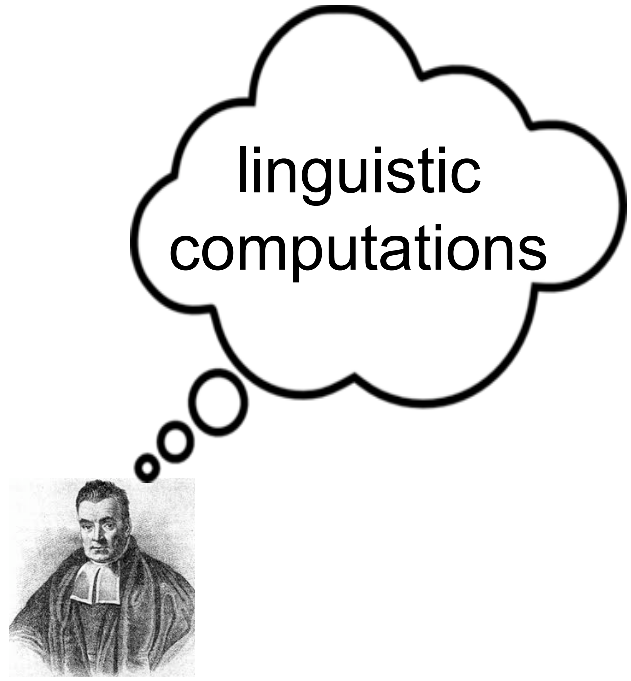
Psycholinguistics



Psycholinguistic Models



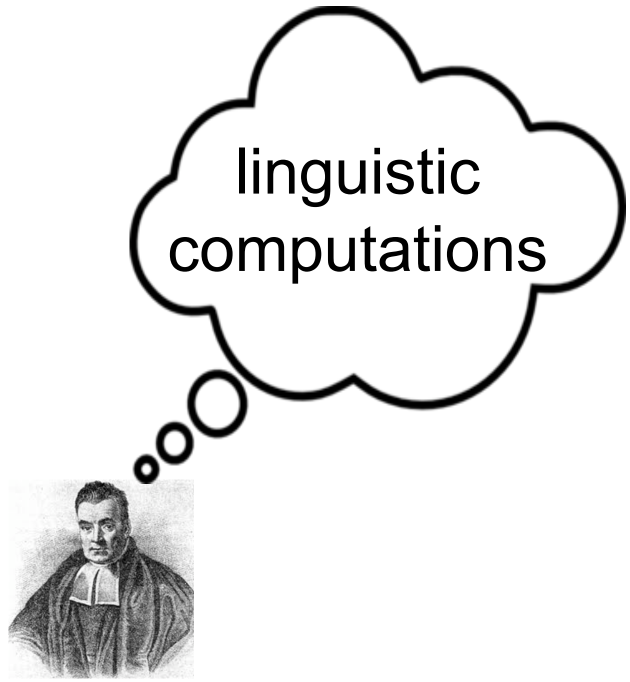
Psycholinguistic Models



Experimental data

(cognitive
psychology)

Psycholinguistic Models



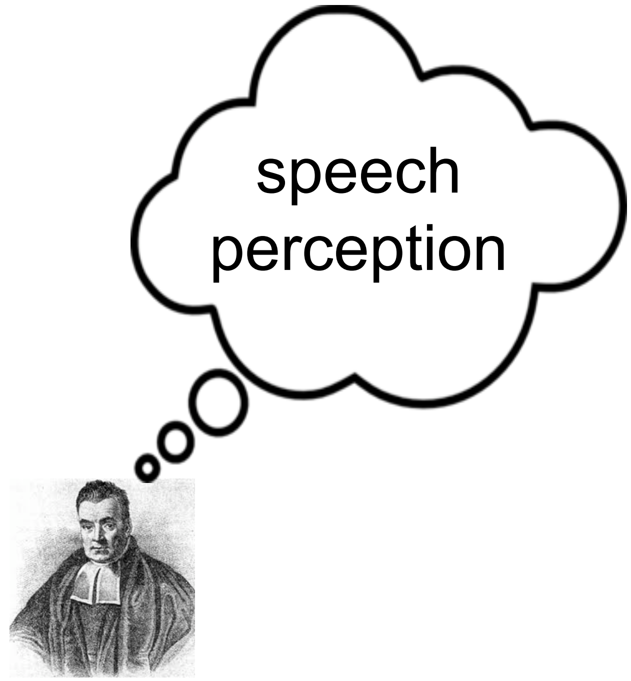
Corpus data

(computational linguistics)

Experimental data

(cognitive psychology)

Models of Speech Perception



Speech corpora

(automatic speech recognition)

Perceptual data

(cognitive psychology)

Bayesian Inference

Posterior probability

Likelihood

Prior probability

$$\boxed{p(h \mid d)} \propto p(d \mid h)\boxed{p(h)}$$

experimental data

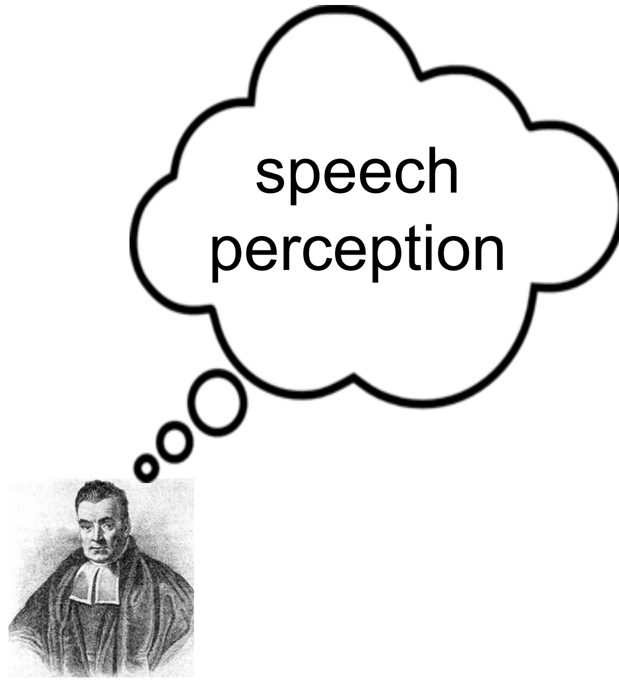
corpus data

h: hypotheses
d: data

Models Using Speech Corpora

- Word recognition in continuous speech (Scharenborg, Norris, ten Bosch, & McQueen, 2005)
- Isolated word recognition (Moore & Maier, 2007)
- Phonological generalization (Kirchner & Moore, 2010)
- Lexical decision (ten Bosch, Boves, & Ernestus, 2013)
- One-shot learning of word forms (Lake, Lee, Glass, & Tenenbaum, in press)

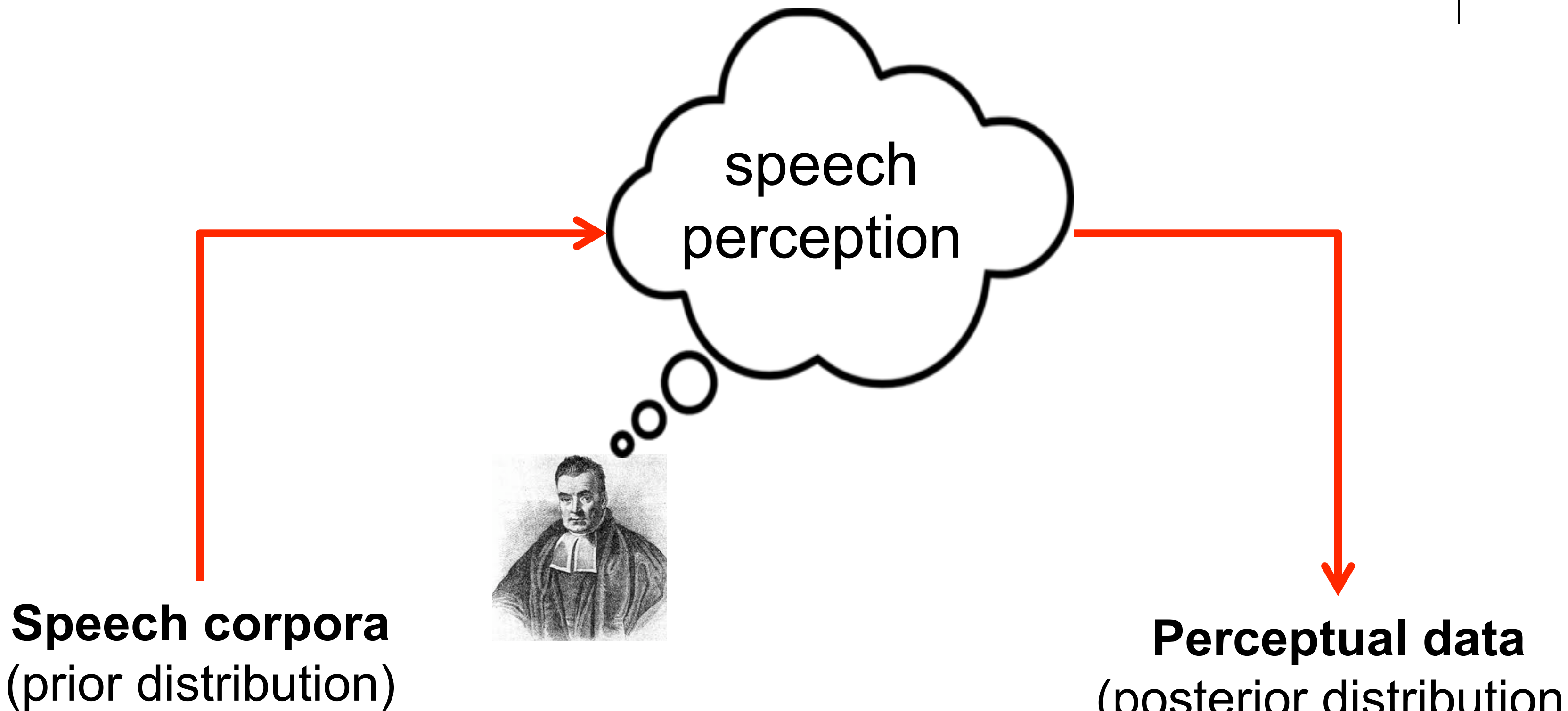
A Model of Speech Perception



Speech corpora
(prior distribution)

Perceptual data
(posterior distribution)

A Model of Speech Perception



Outline

- Behavioral data in speech perception
- Cognitive model of speech perception
- Adapting the model to speech corpora
- A case study: Speaker normalization

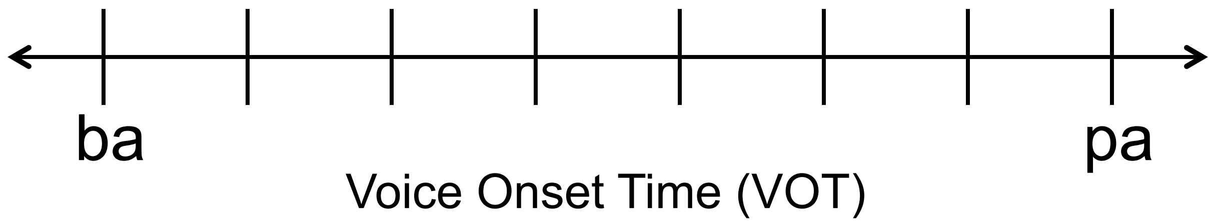
Categories Affect Perception

Sound categories

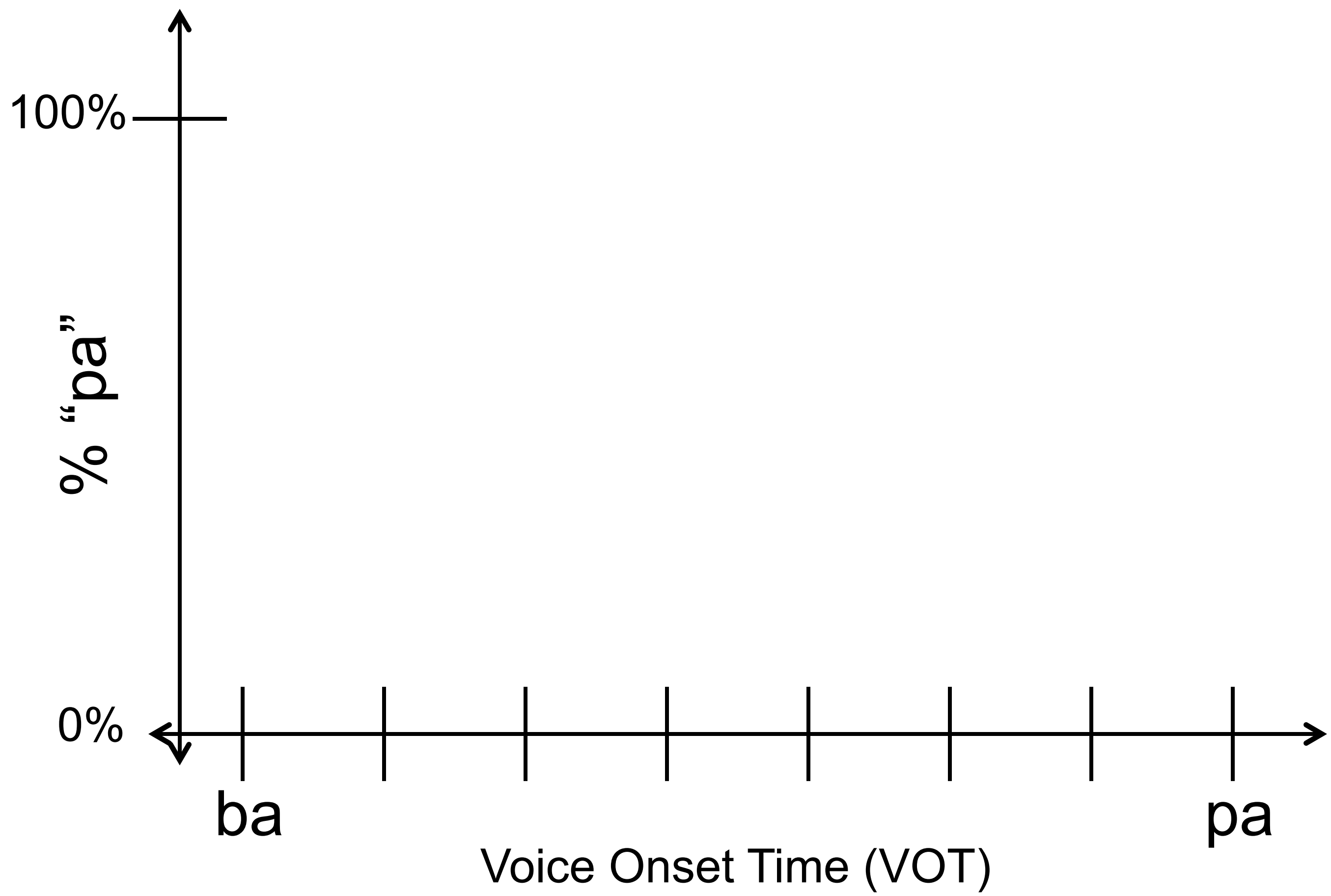
- Stop consonants (Liberman et al., 1957, 1961)
- Fricatives (Repp, 1981)
- Liquids (Miyawaki et al., 1975; Iverson et al., 2003)
- Vowels (Kuhl et al., 1992)

Parallel effects in color, face, and object perception
(Davidoff, Davies, & Roberson, 1999; Etcoff & Magee, 1992; Goldstone, Lippa, & Shiffrin, 2001)

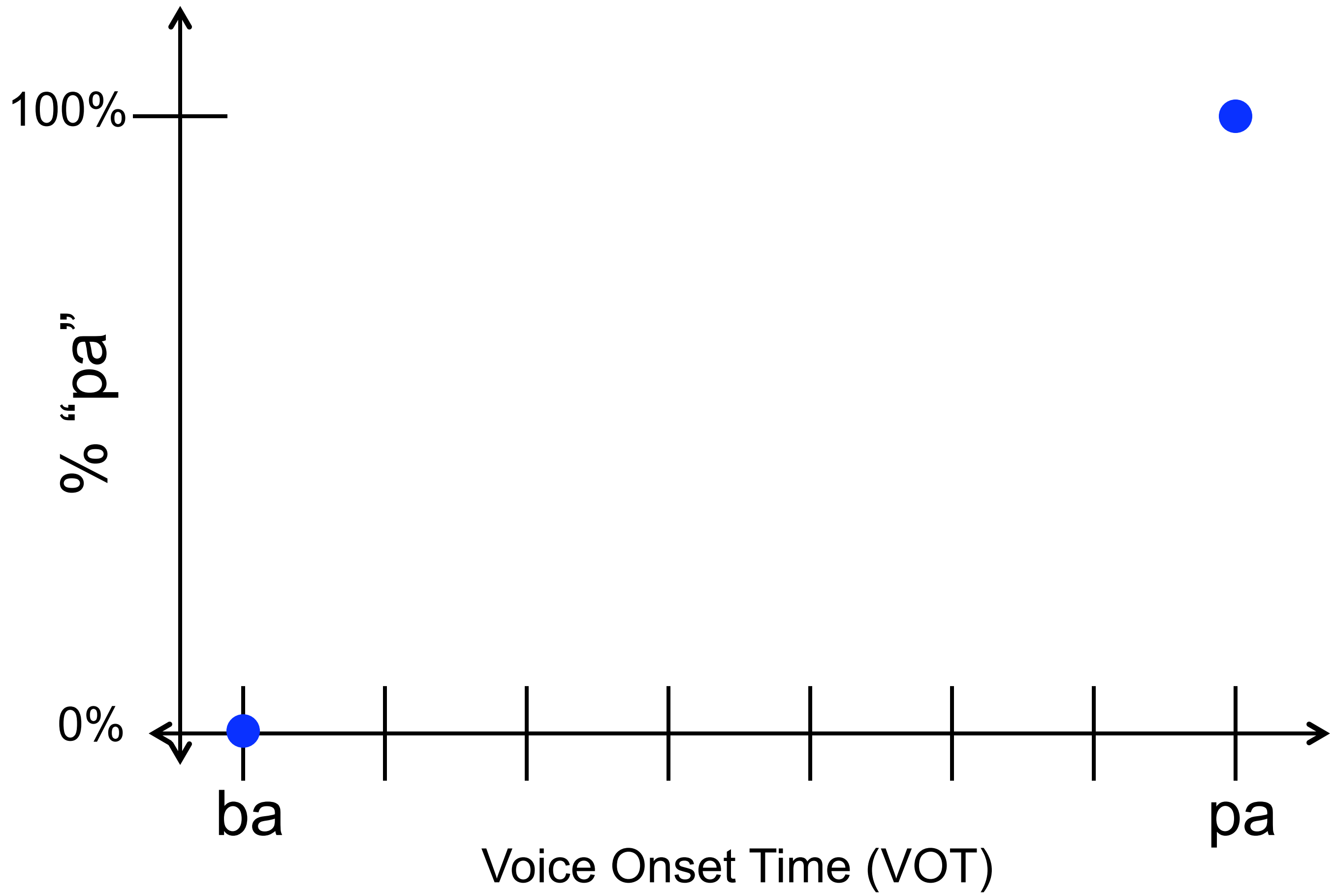
Stop Consonants



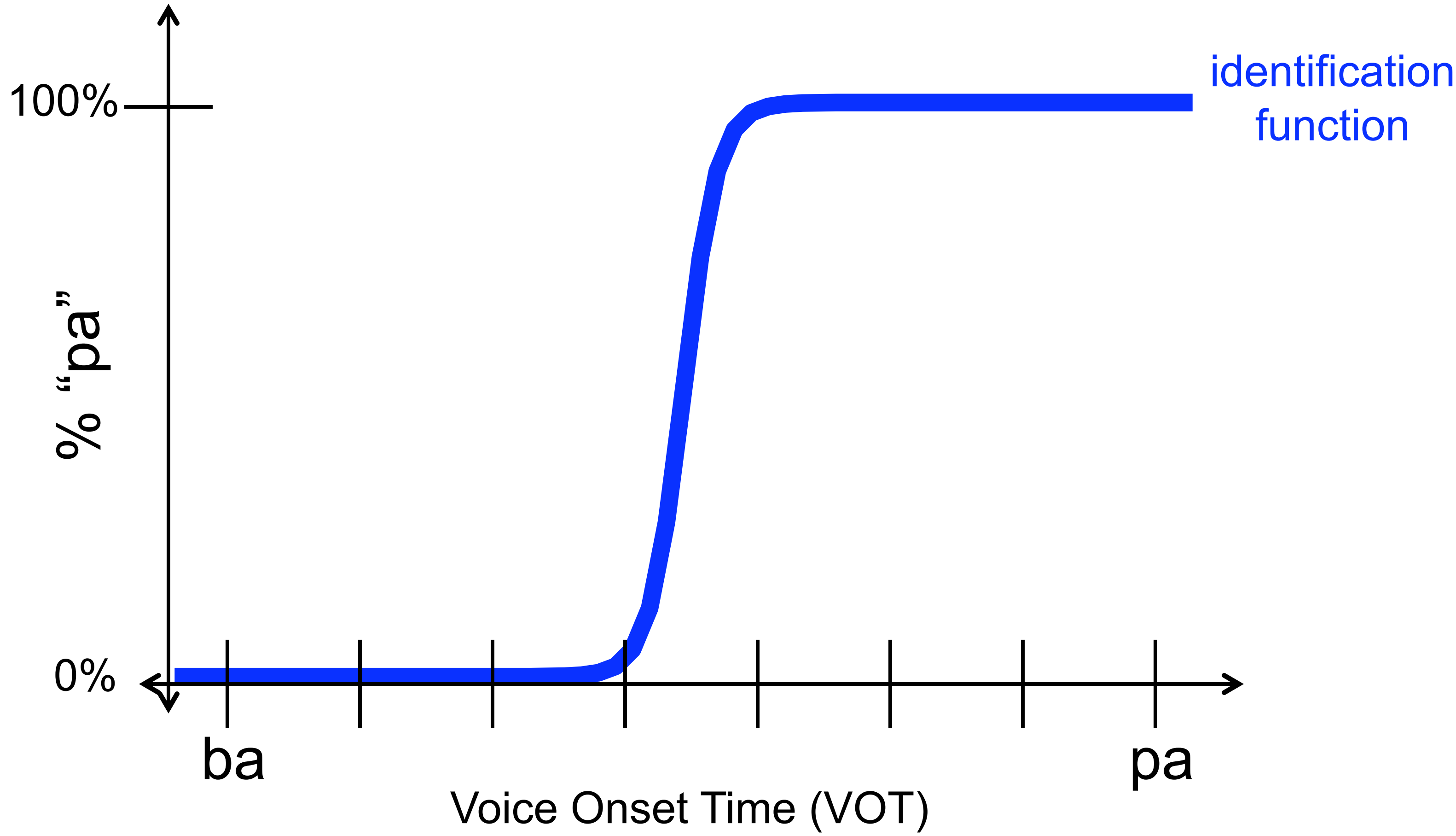
Identification Data



Identification Data

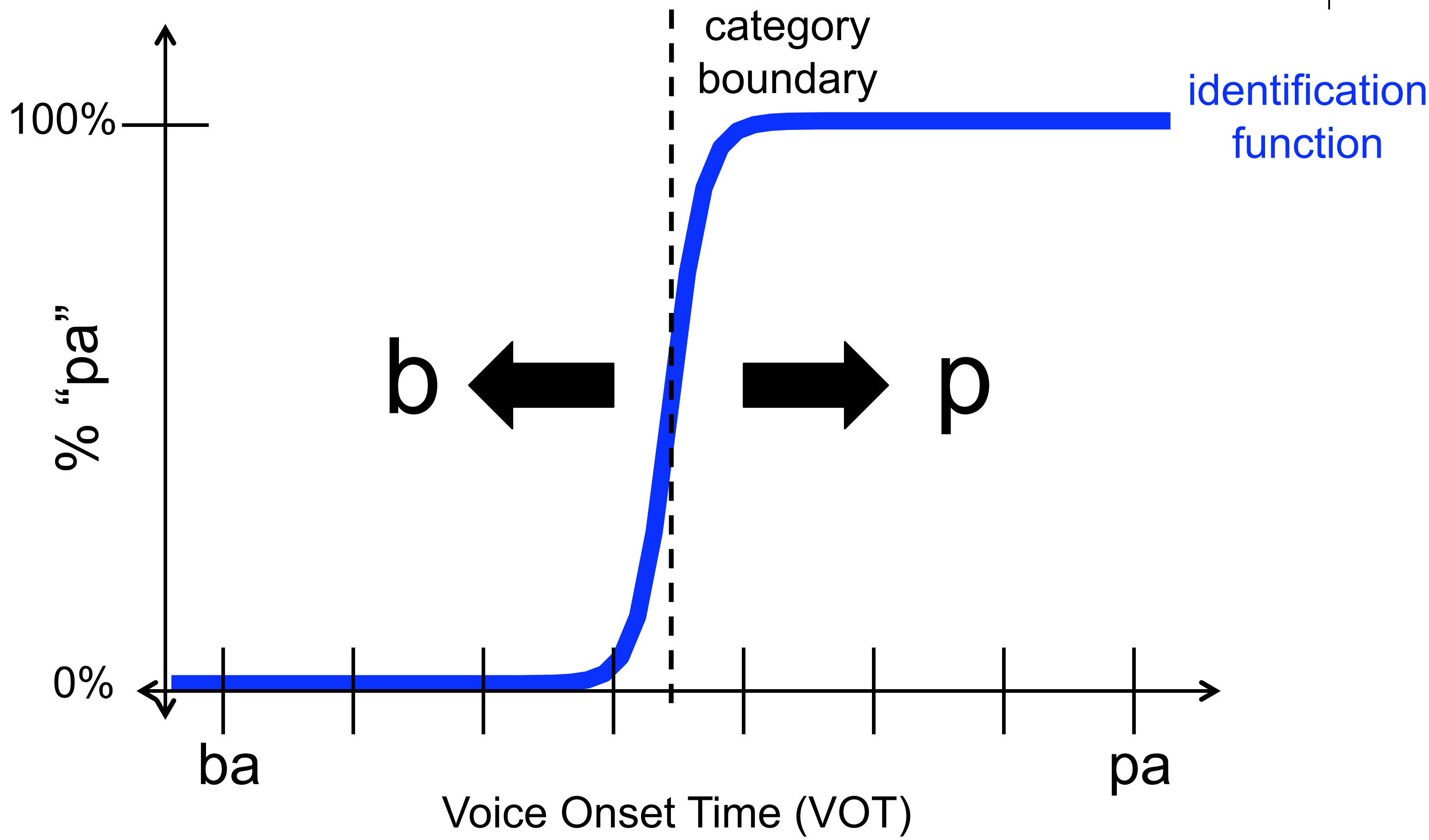


Identification Data



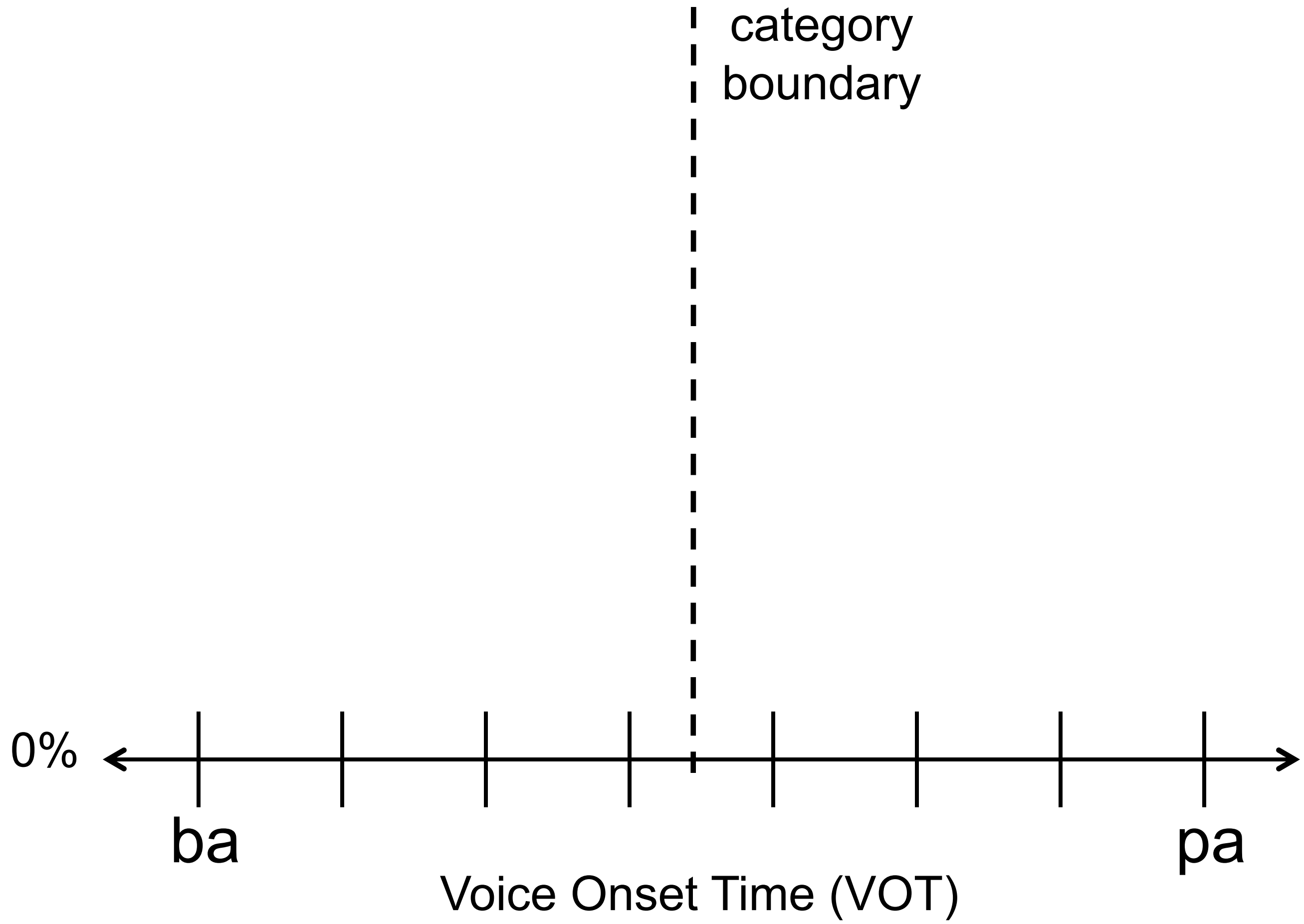
(Liberman, Harris, Hoffman, & Griffith, 1957)

Identification Data

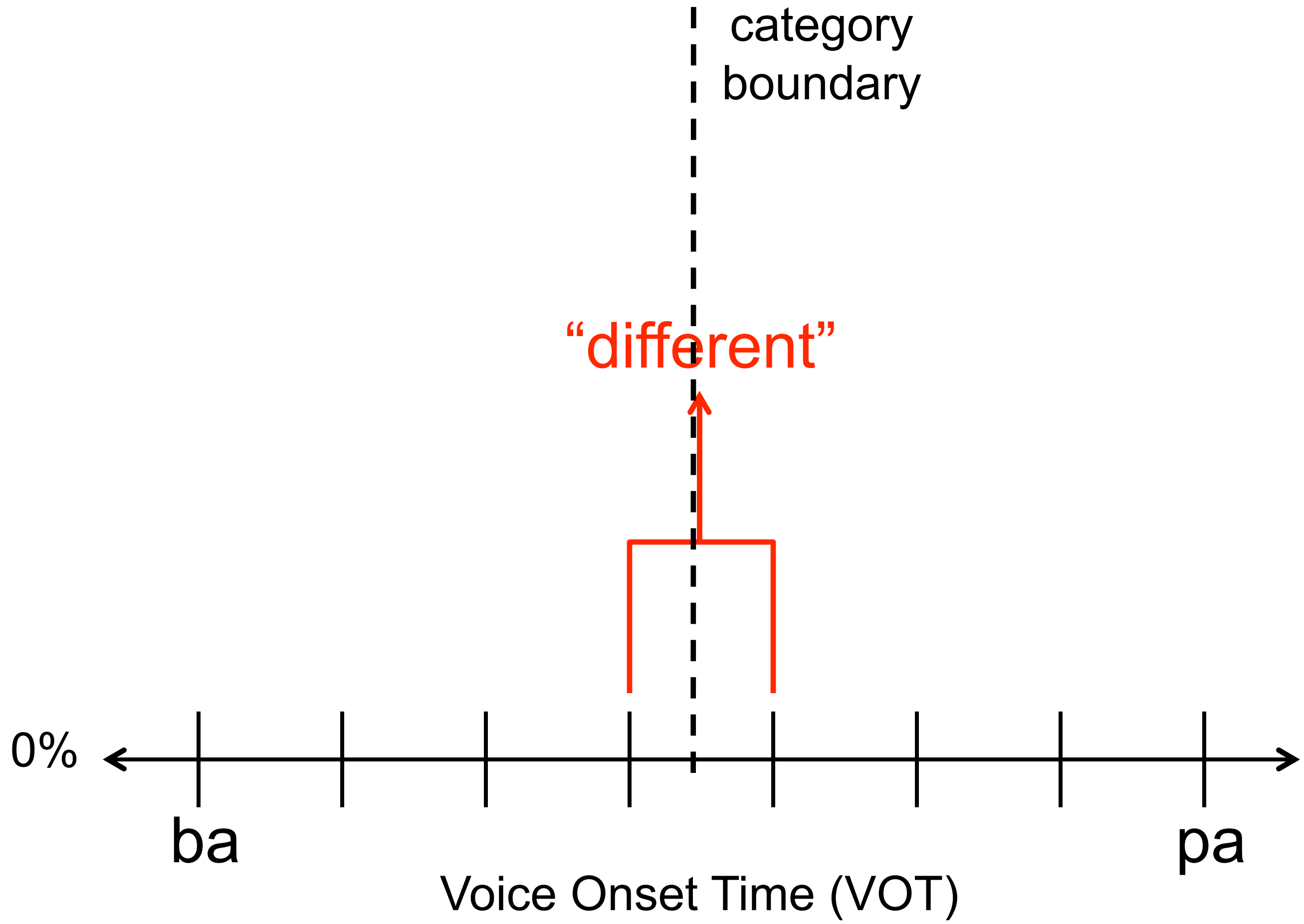


(Liberman, Harris, Hoffman, & Griffith, 1957)

Discrimination Data

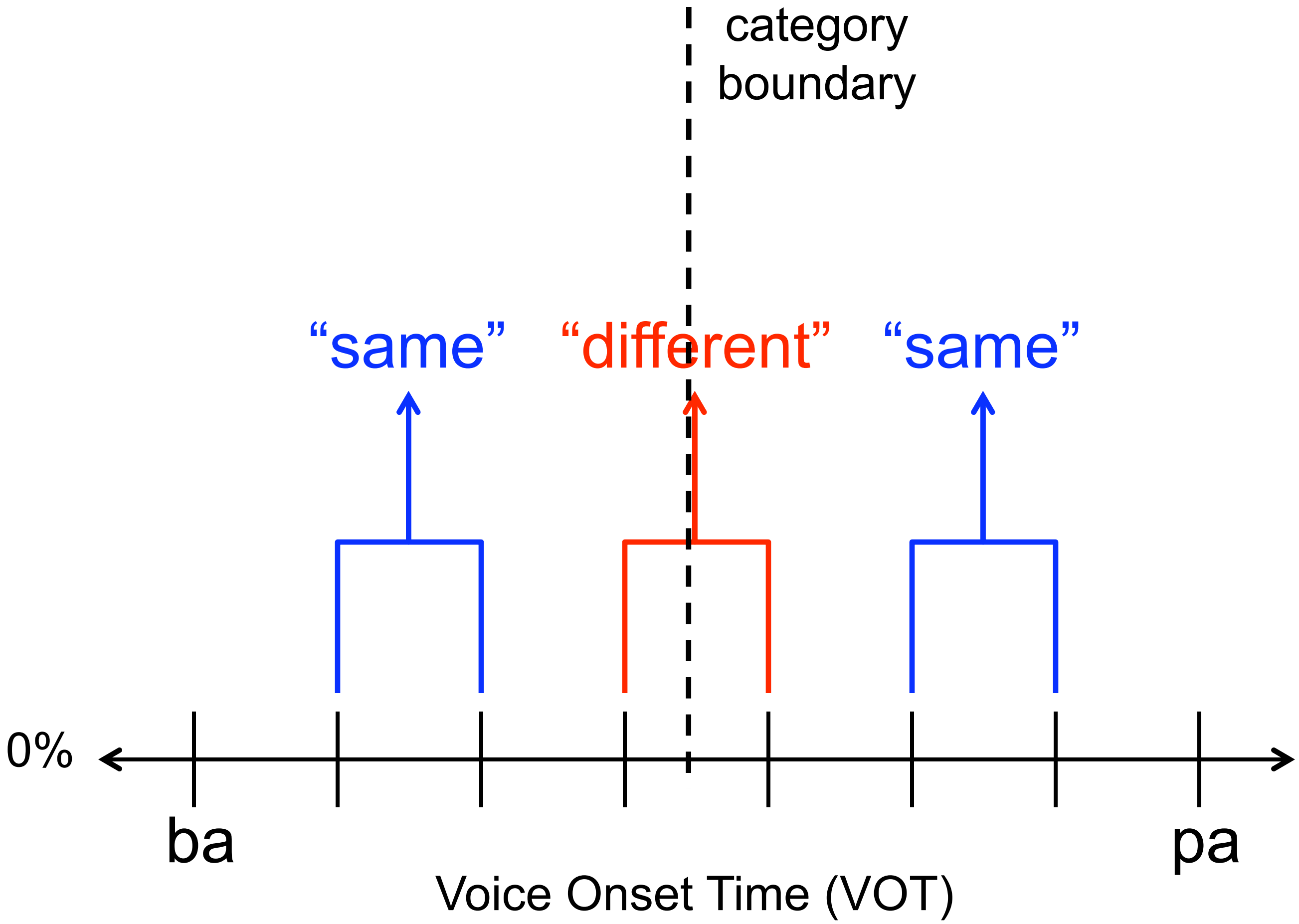


Discrimination Data



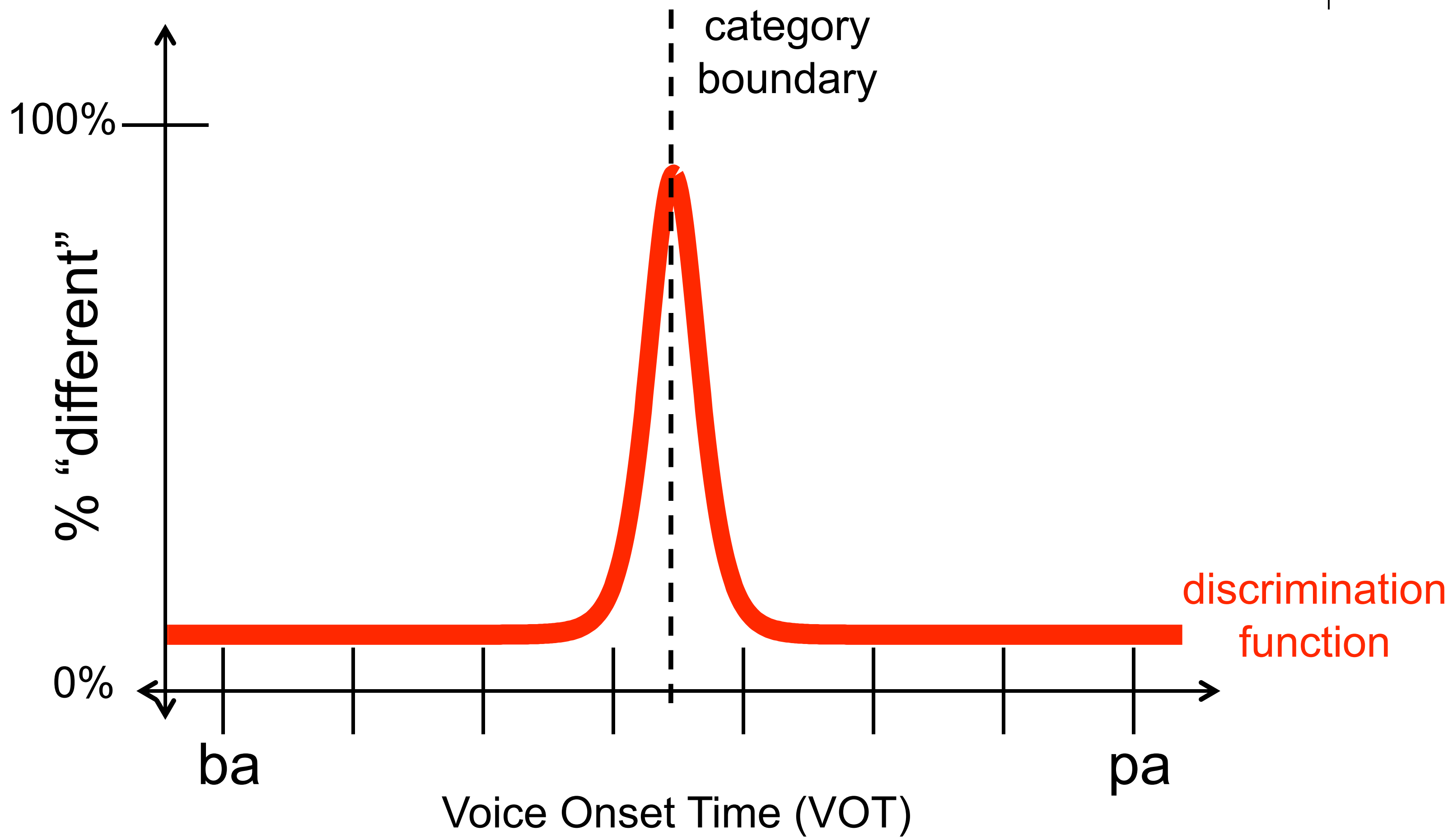
(Liberman, Harris, Hoffman, & Griffith, 1957)

Discrimination Data



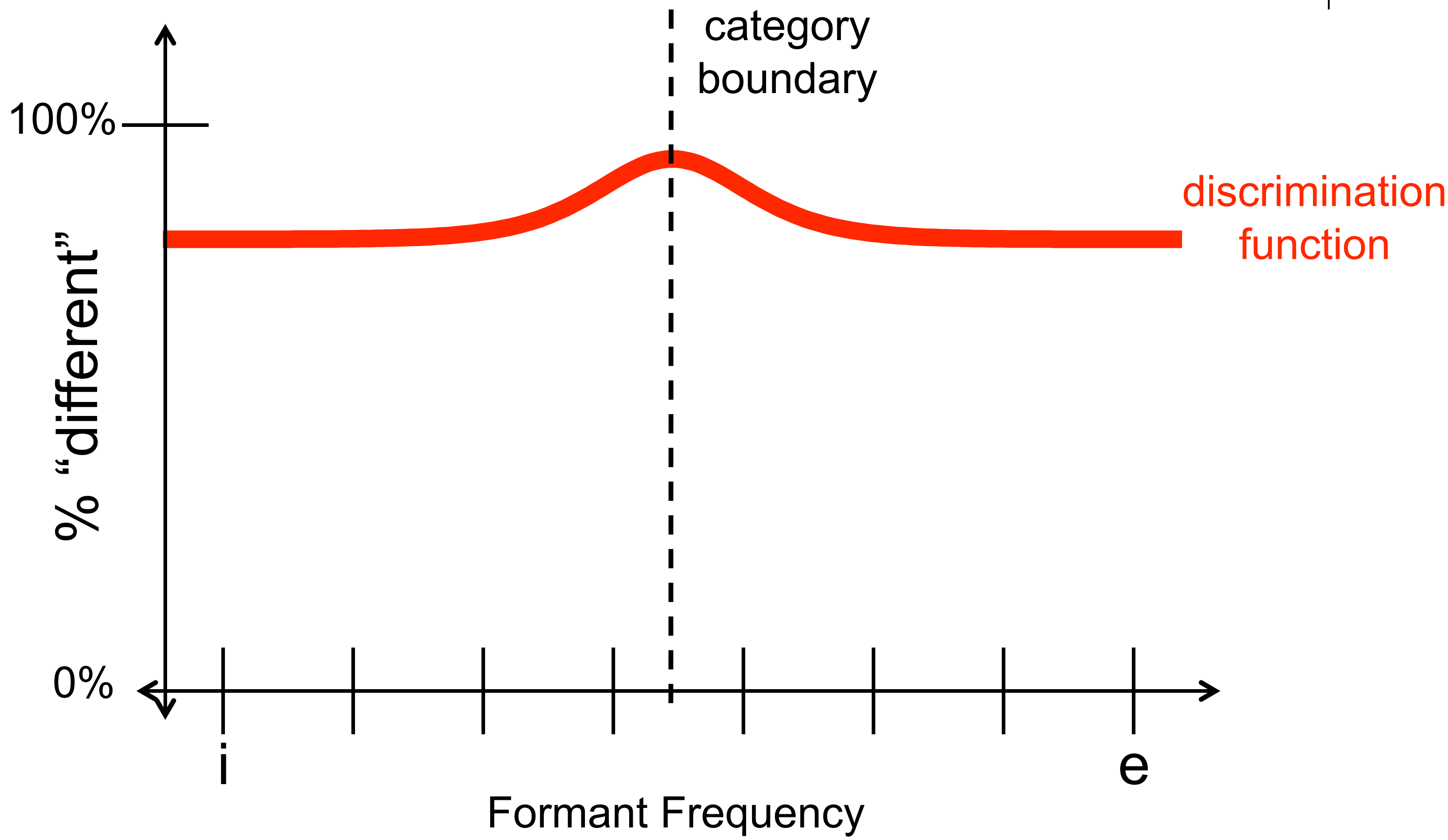
(Liberman, Harris, Hoffman, & Griffith, 1957)

Stop Consonants

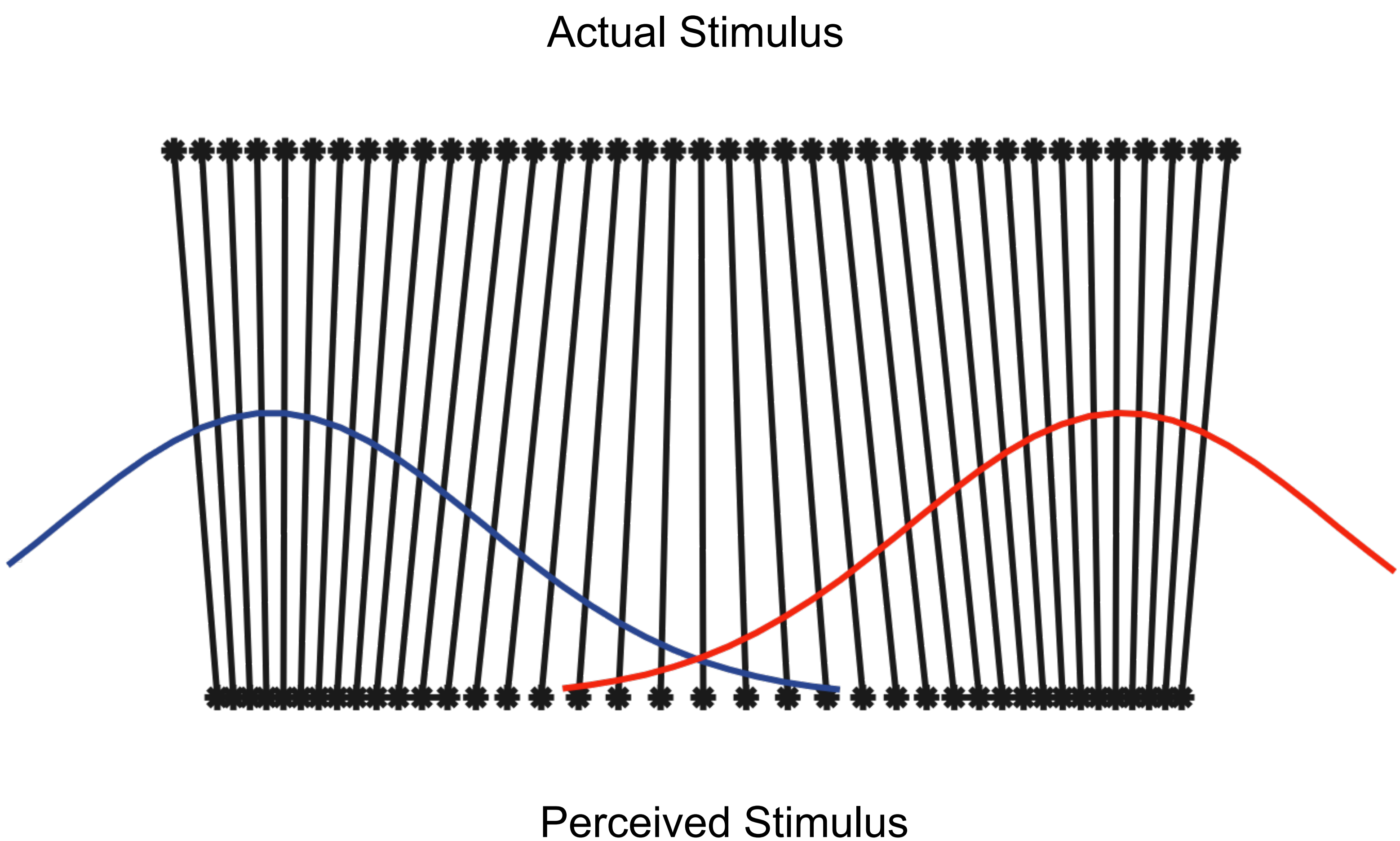


(Liberman, Harris, Hoffman, & Griffith, 1957)

Vowels


category
boundary
100%
0%
% "different"
discrimination
function
i
e
Formant Frequency

Qualitatively Similar Effects



Different Explanations

- Stop consonants: Categorical perception

  Listeners extract category information and discriminate sounds on the basis of that category information (Liberman et al., 1957)

- Vowels: Perceptual magnet effect

  Sounds are "pulled" toward phonetic category prototypes (Grieser & Kuhl, 1989; Iverson & Kuhl, 1995)

A unified explanation for strong and weak categorical effects

Outline

- Behavioral data in speech perception
- Cognitive model of speech perception


Yakov Kronrod


Emily Coppess


Tom Griffiths


James Morgan

- Adapting the model to speech corpora
- A case study: Speaker normalization

Noisy Channel Model

Noisy Channel Model

c

Speaker chooses
a phonetic category

Noisy Channel Model

C

Speaker chooses
a phonetic category

T

Speaker articulates a
“target production”

Noisy Channel Model

Noise in the
speech signal

c
Speaker chooses
a phonetic category

T
Speaker articulates a
“target production”

Noisy Channel Model

Listener hears
a speech sound

S

Noise in the
speech signal

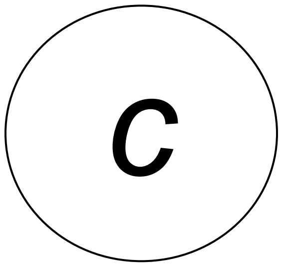
C

Speaker chooses
a phonetic category

T

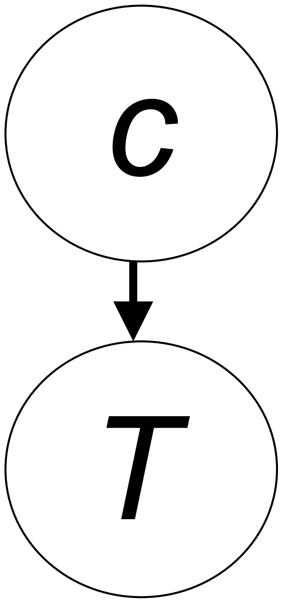
Speaker articulates a
"target production"

Generative Model



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

Generative Model

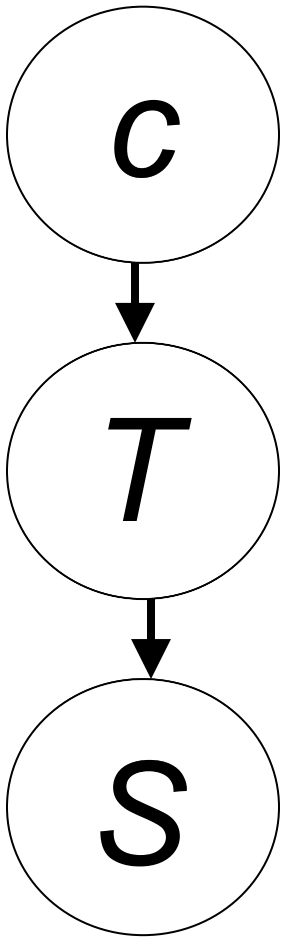


Choose a category c with probability $p(c)$

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

$$p(T \mid c) = N\left(\mu_c, \sigma_c^{\;2}\right)$$

Generative Model



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

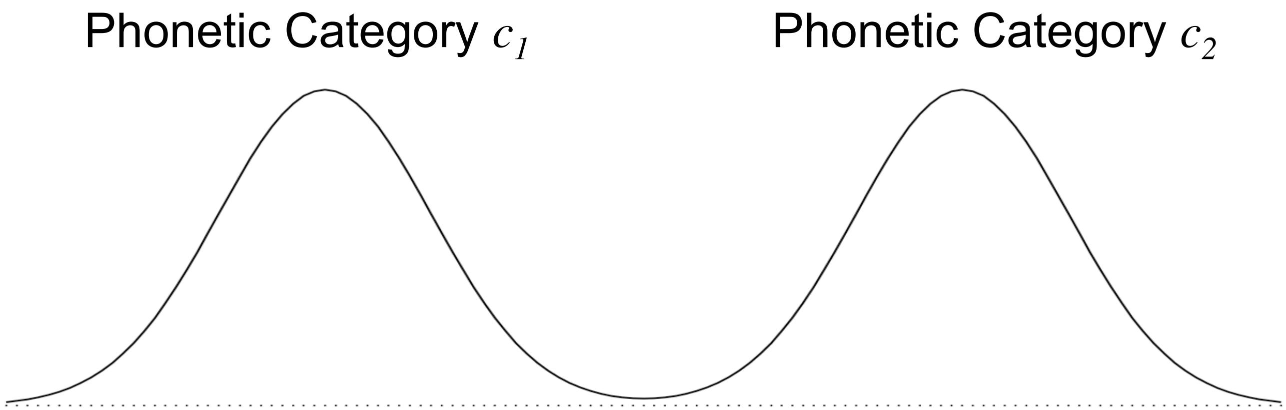
Articulate a target production T with probability $p(T|c)$

$$p(T \mid c) = N\left(\mu_c, \sigma_c^{\ 2}\right)$$

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

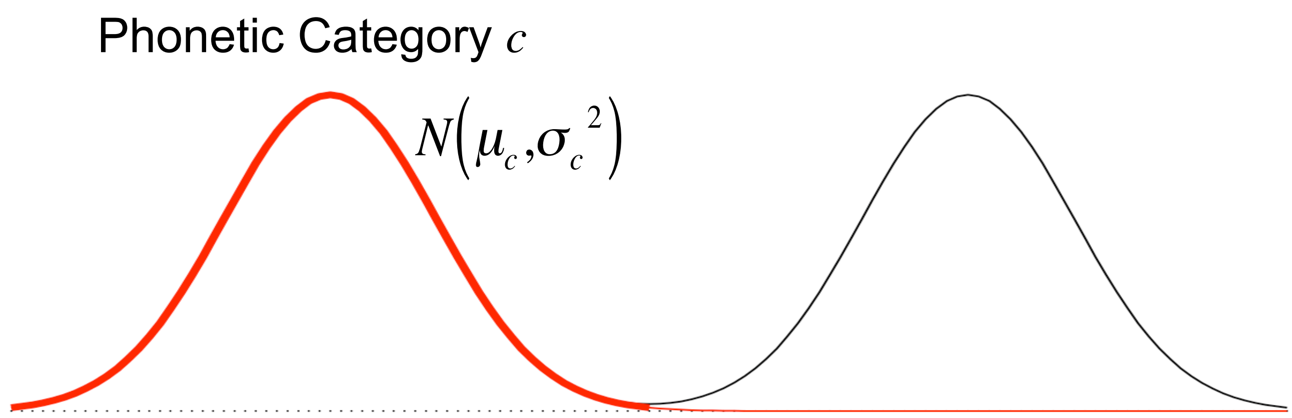
$$p(S \mid T) = N\left(T, \sigma_S^{\ 2}\right)$$

Generative Model



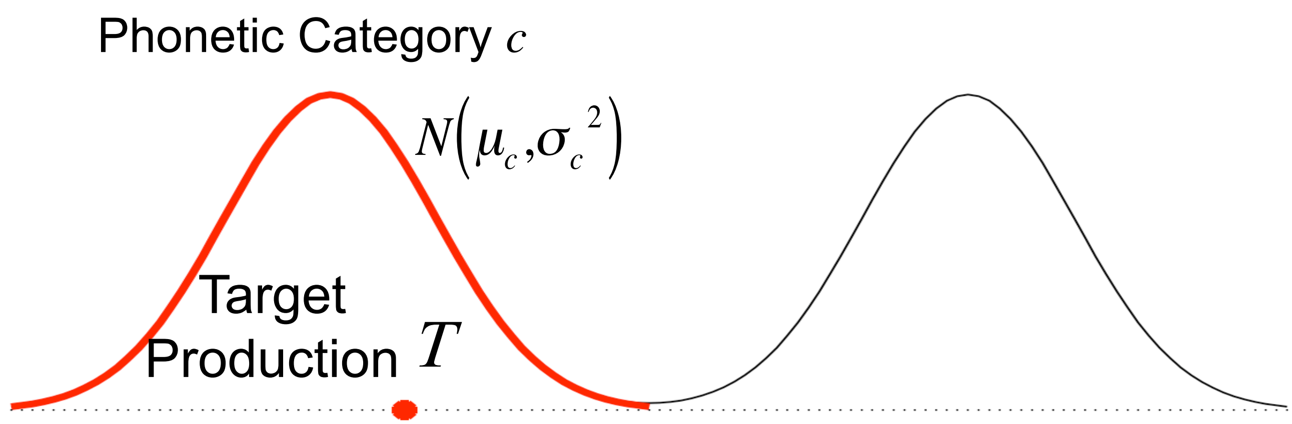
(Feldman, Griffiths, & Morgan, 2009)

Generative Model



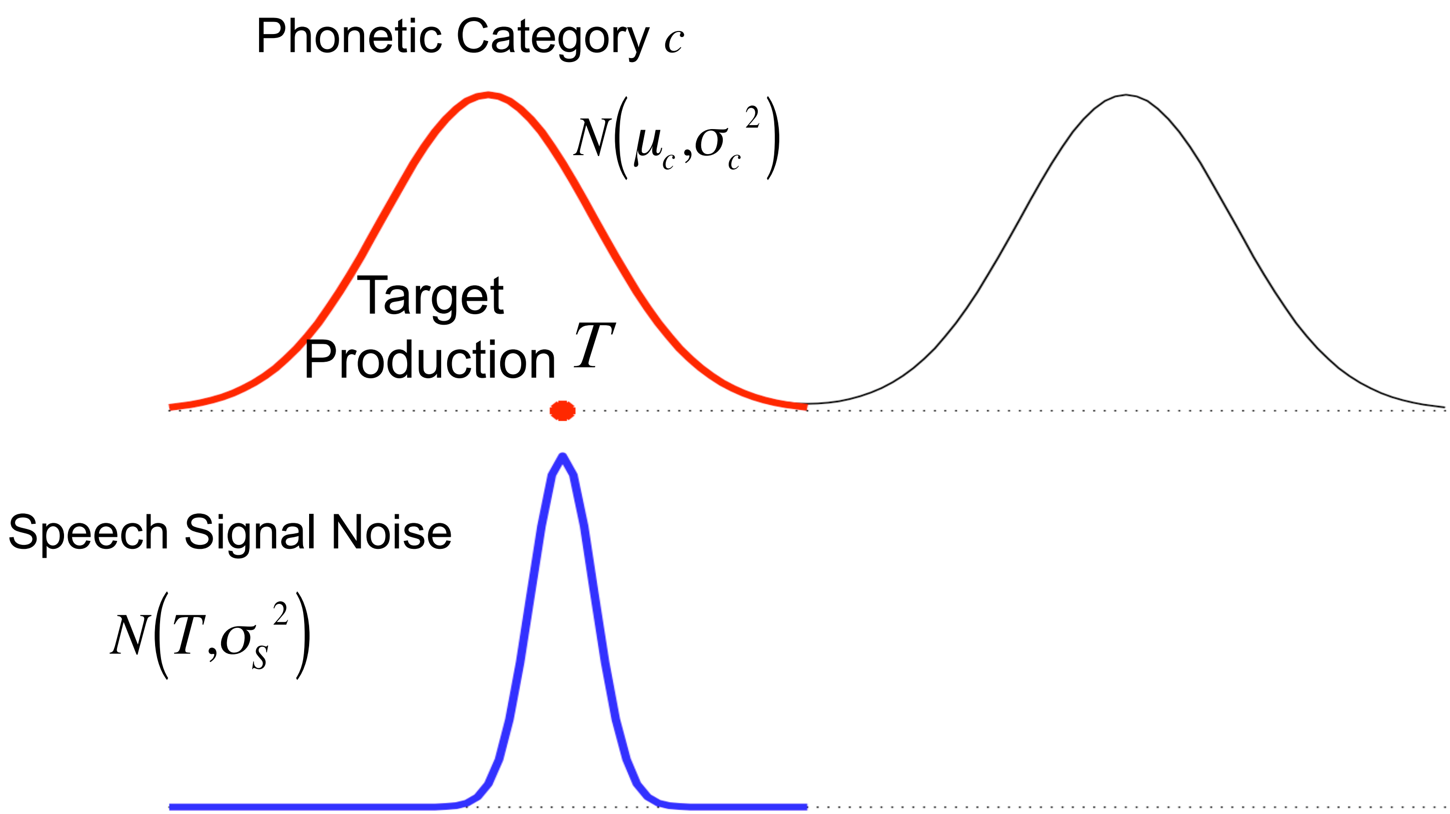
(Feldman, Griffiths, & Morgan, 2009)

Generative Model



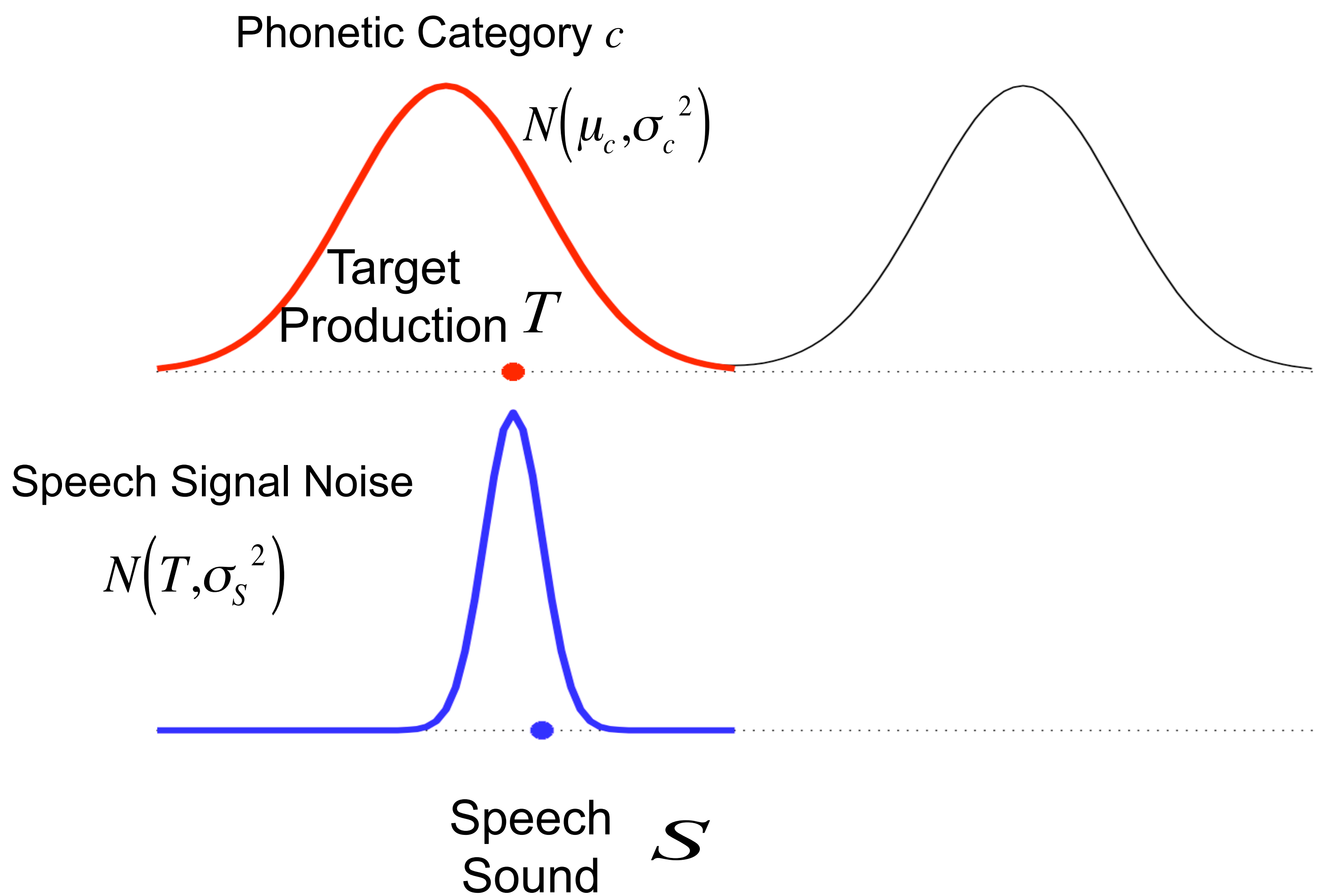
(Feldman, Griffiths, & Morgan, 2009)

Generative Model



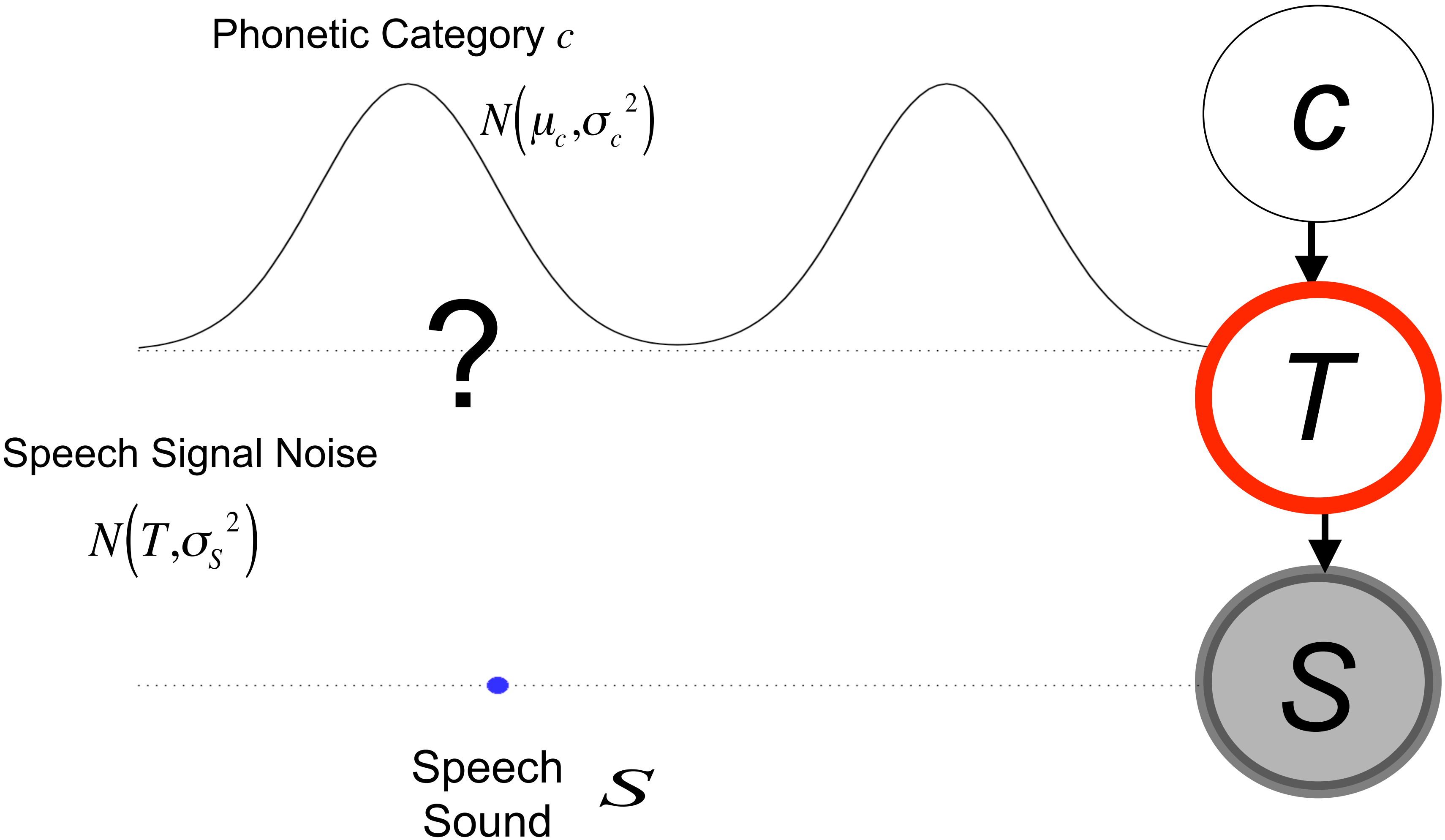
(Feldman, Griffiths, & Morgan, 2009)

Generative Model



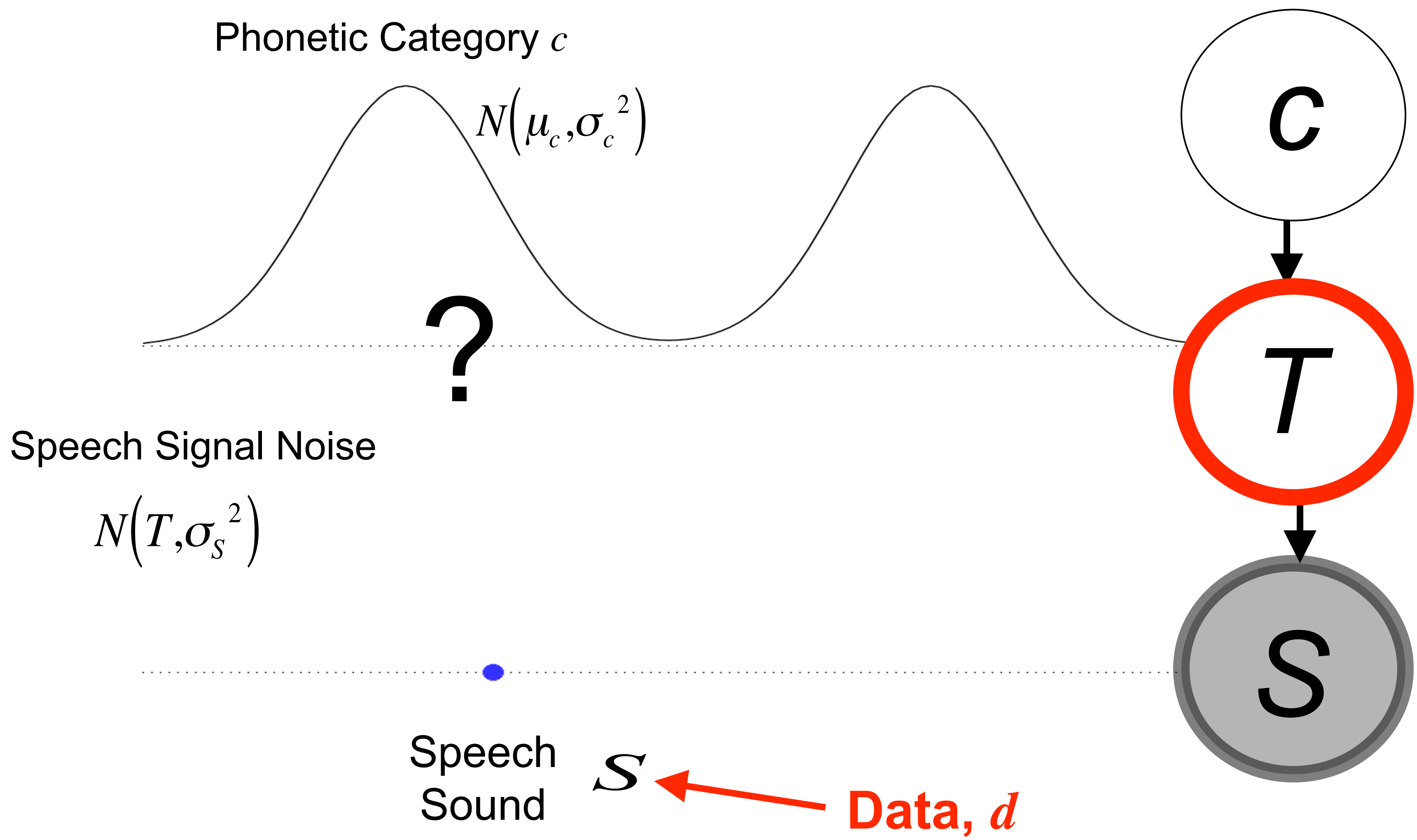
(Feldman, Griffiths, & Morgan, 2009)

Bayes for Discrimination



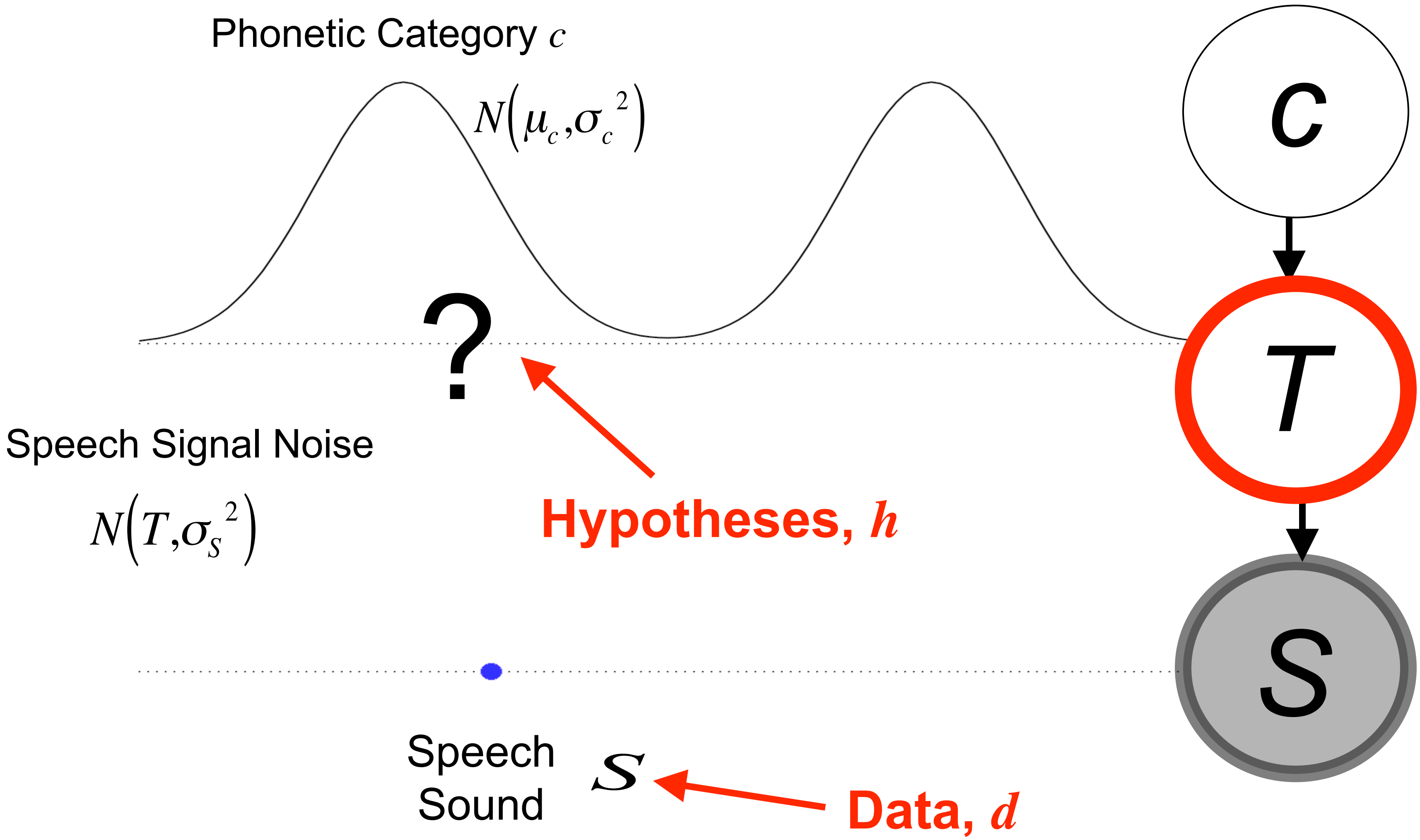
(Feldman, Griffiths, & Morgan, 2009)

Bayes for Discrimination



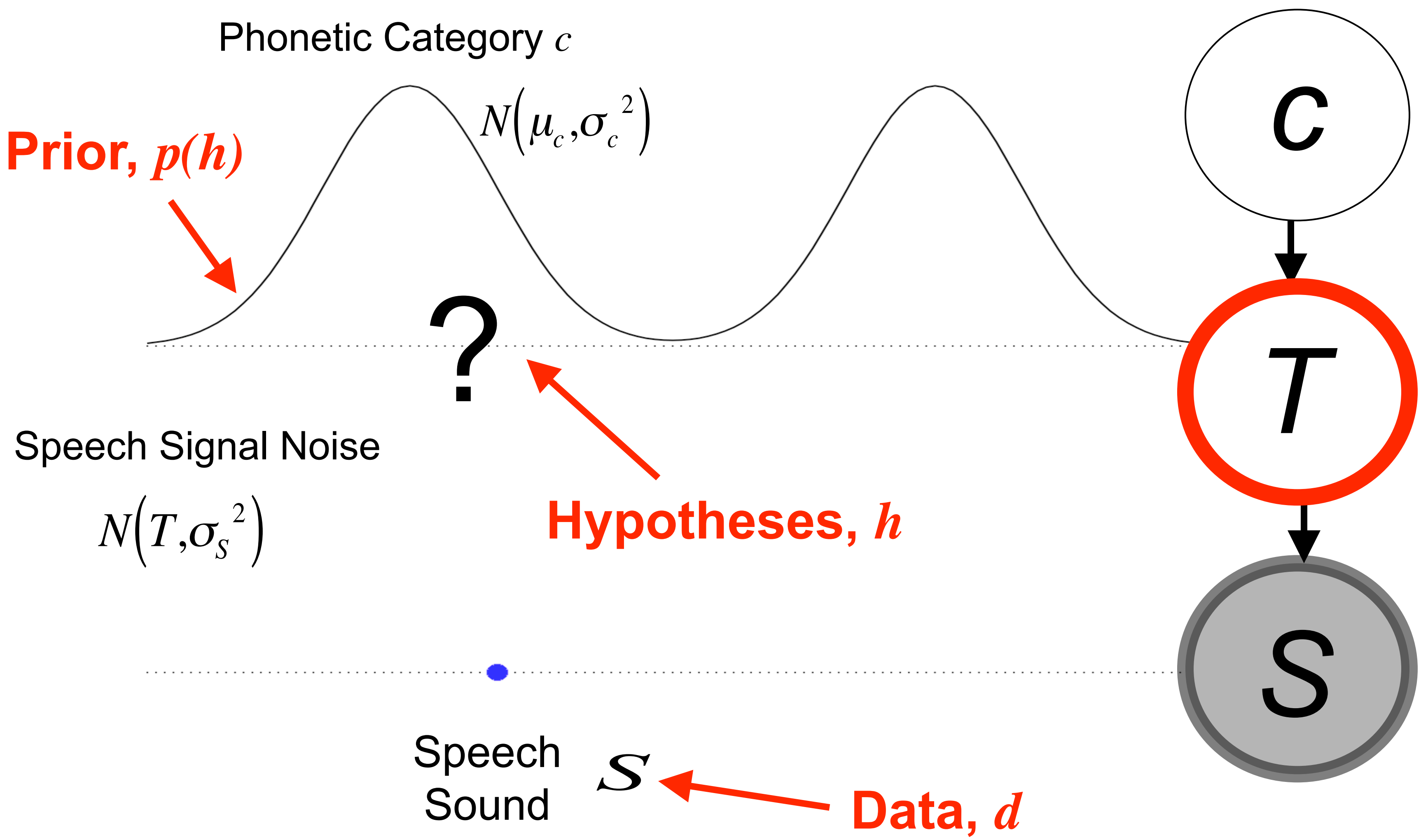
(Feldman, Griffiths, & Morgan, 2009)

Bayes for Discrimination



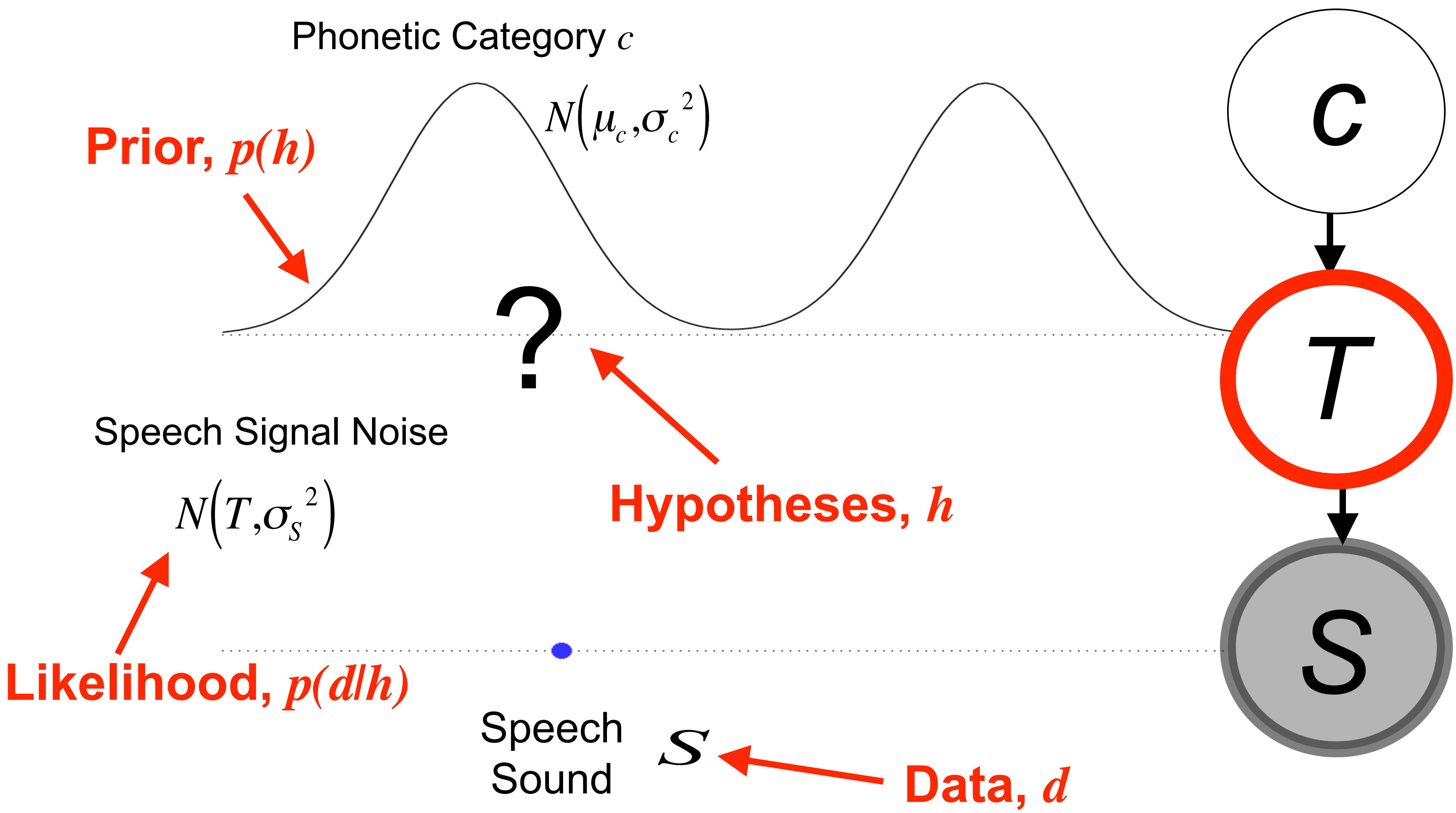
(Feldman, Griffiths, & Morgan, 2009)

Bayes for Discrimination



(Feldman, Griffiths, & Morgan, 2009)

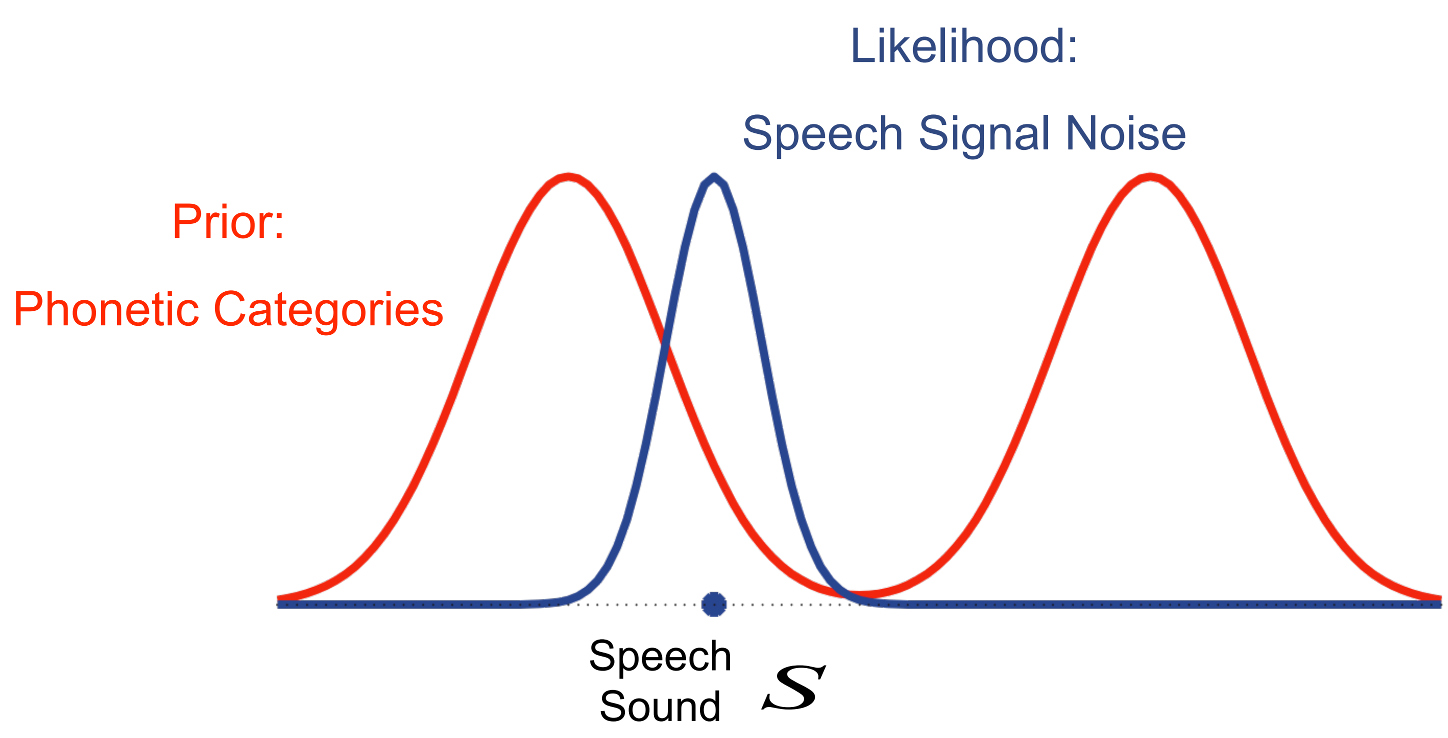
Bayes for Discrimination



(Feldman, Griffiths, & Morgan, 2009)

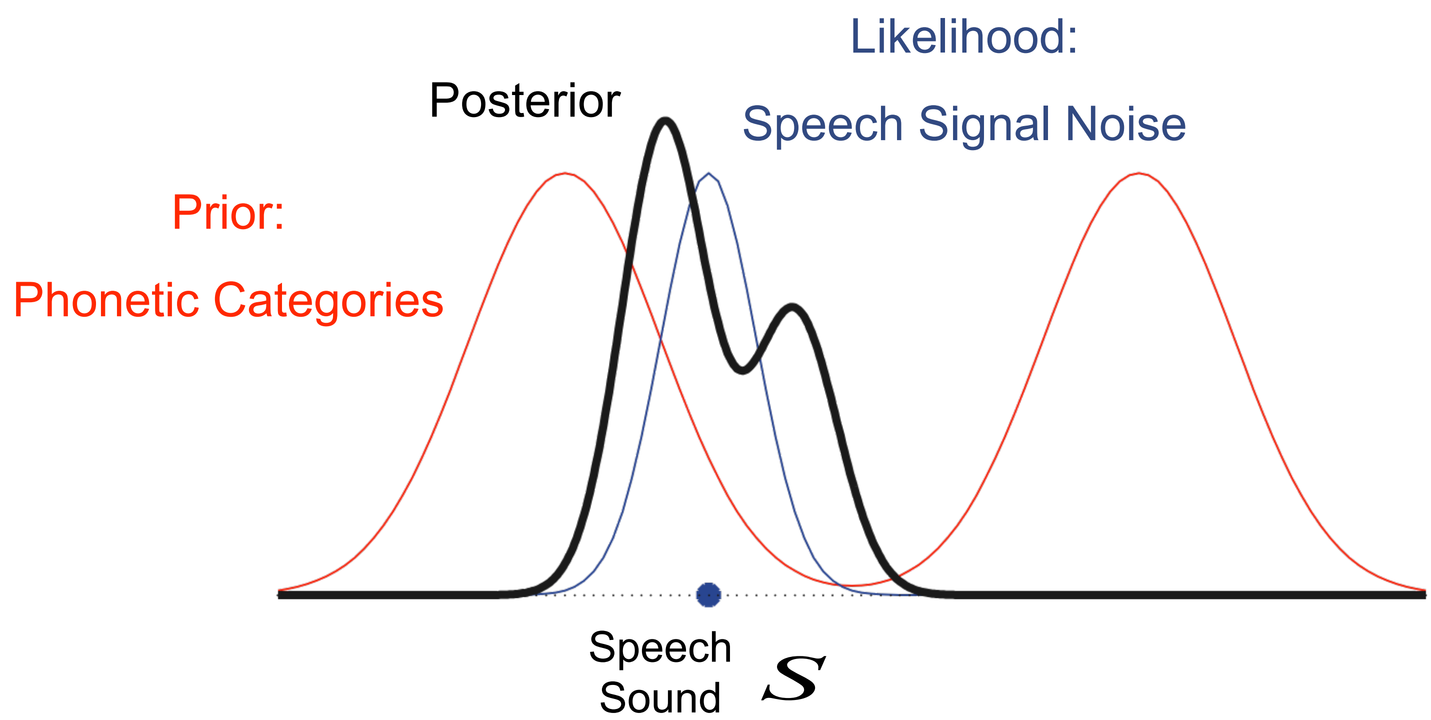
Inferring the Speaker's Target

Bayes' rule: $$p(T \mid S) \propto p(S \mid T)\,p(T)$$



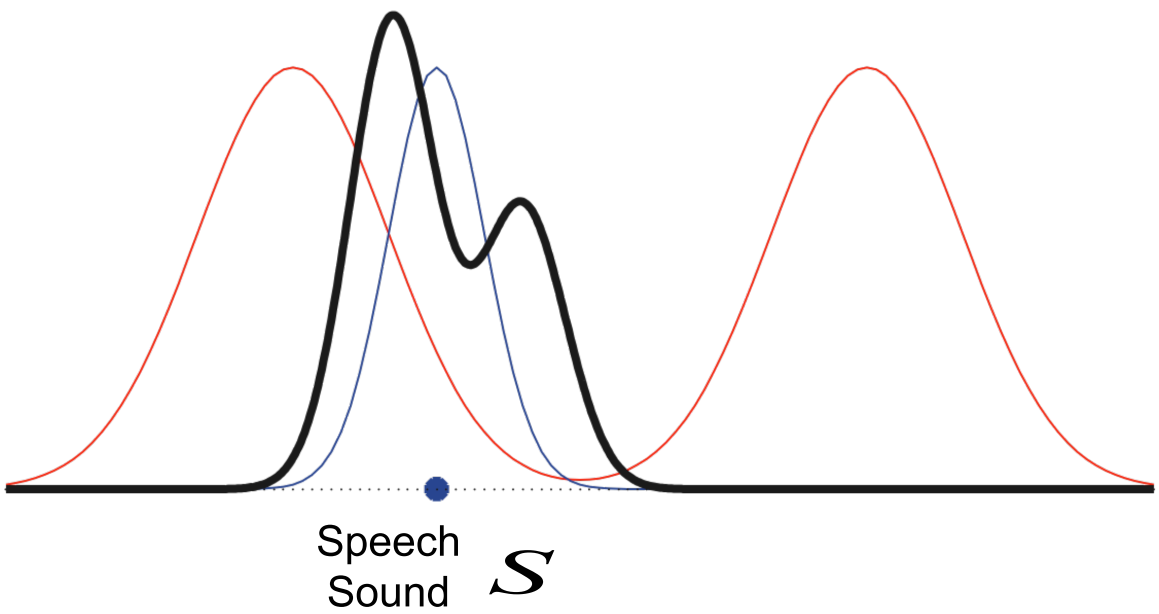
Inferring the Speaker's Target

Bayes' rule: $$p(T \mid S) \propto p(S \mid T)p(T)$$



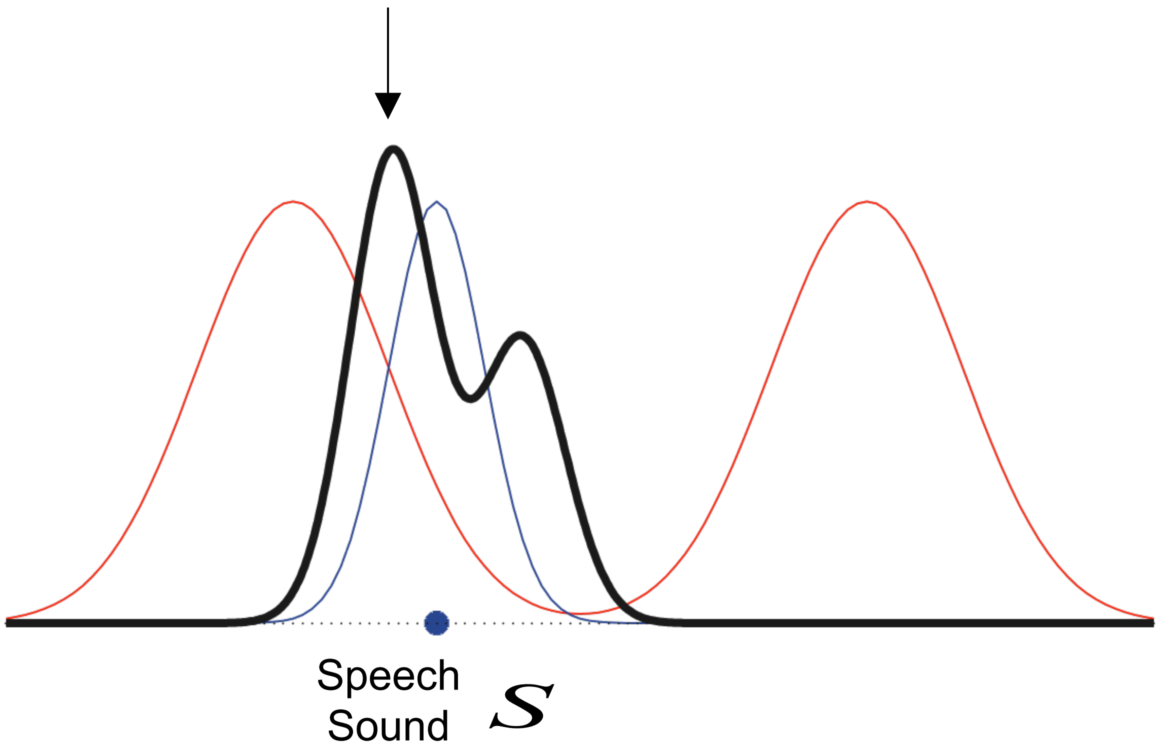
Inferring the Speaker's Target

$$E[T \mid S, c] = \frac{\sigma_c^2 S + \sigma_S^2 \mu_c}{\sigma_c^2 + \sigma_S^2}$$



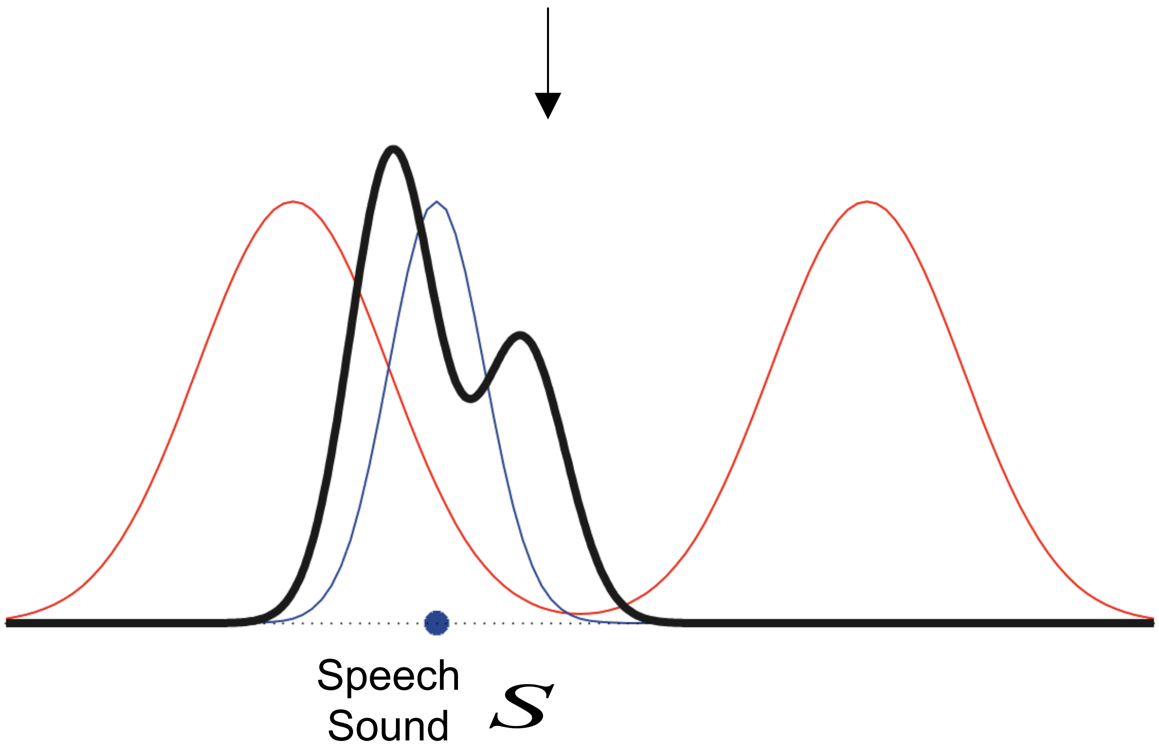
Inferring the Speaker's Target

$$E\left[T \mid S,c\right] = \frac{\sigma_c{}^2 \boxed{S} + \sigma_S{}^2 \boxed{\mu_c}}{\sigma_c{}^2 + \sigma_S{}^2}$$



Inferring the Speaker's Target

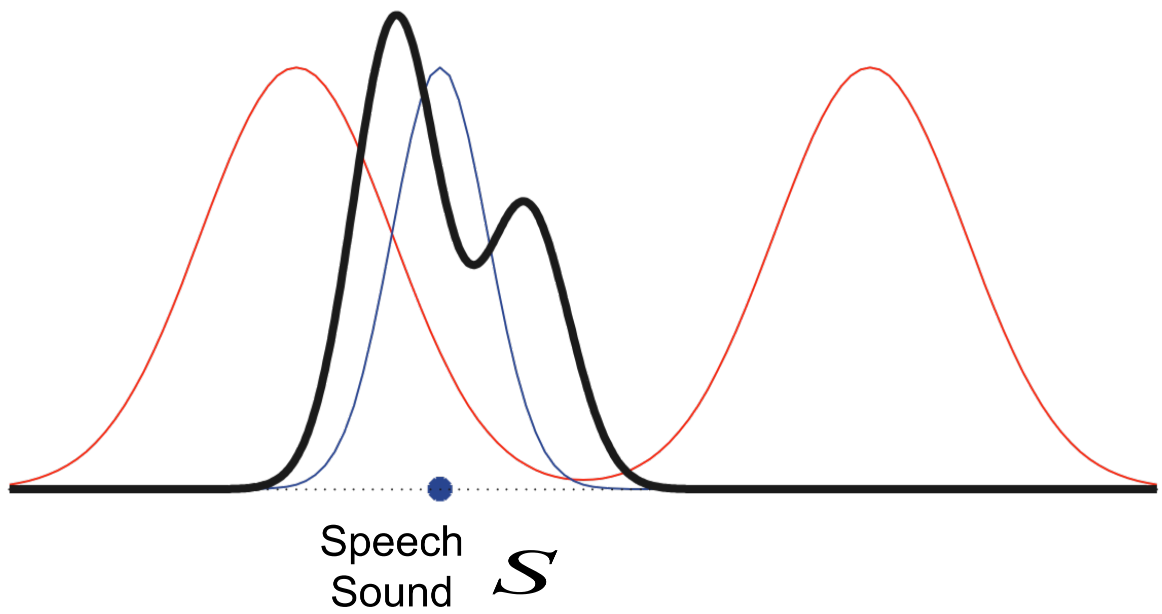
$$E\left[T \mid S, c\right] = \frac{\sigma_c^{\,2} S + \sigma_S^{\,2} \mu_c}{\sigma_c^{\,2} + \sigma_S^{\,2}}$$



Inferring the Speaker's Target

Sum over phonetic categories:

$$E[T \mid S] = \sum_c \boxed{p(c \mid S)} \frac{\sigma_c^{\,2} S + \sigma_S^{\,2} \mu_c}{\sigma_c^{\,2} + \sigma_S^{\,2}}$$



Qualitative Predictions

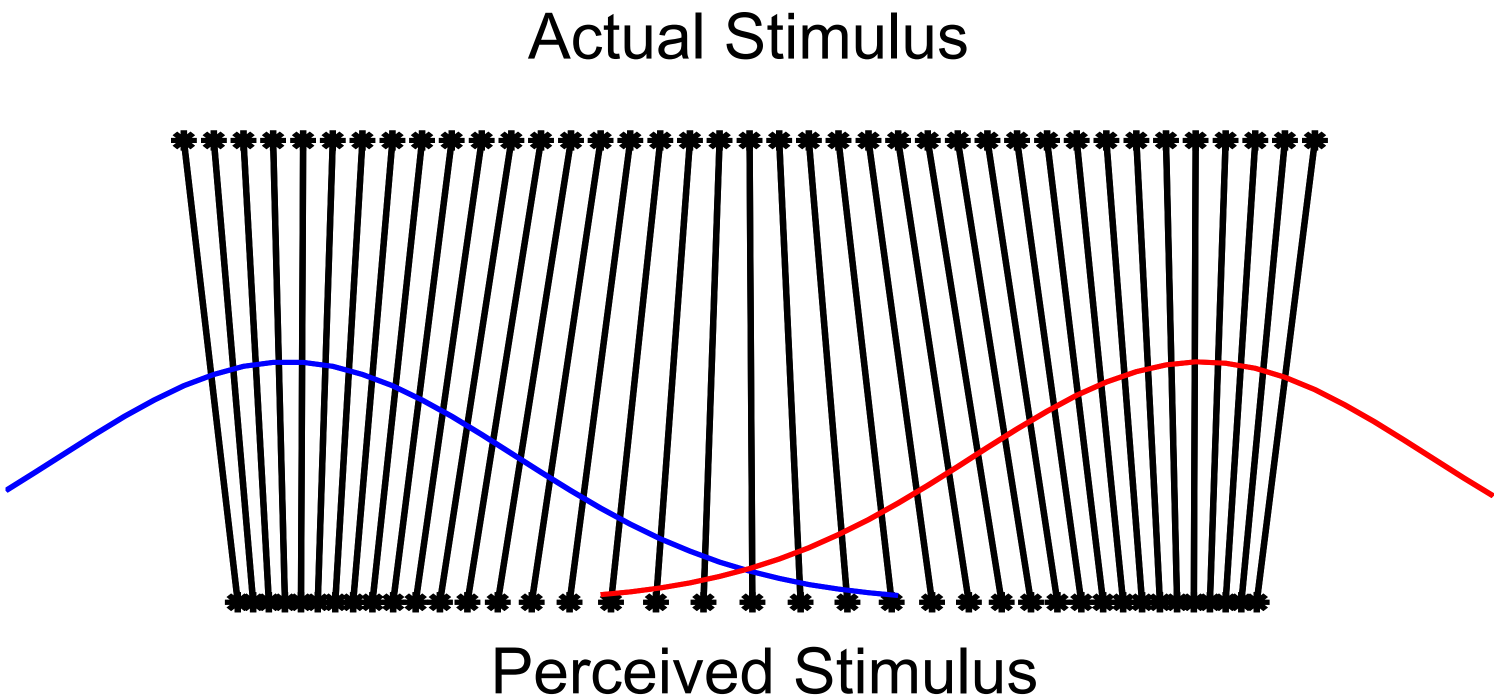
- Perception of unambiguous speech sounds is pulled toward the phonetic category mean.

Qualitative Predictions

- Perception of unambiguous speech sounds is pulled toward the phonetic category mean.

- Speech sounds between two categories are pulled simultaneously toward both category means, each category cancelling out the other's effect.

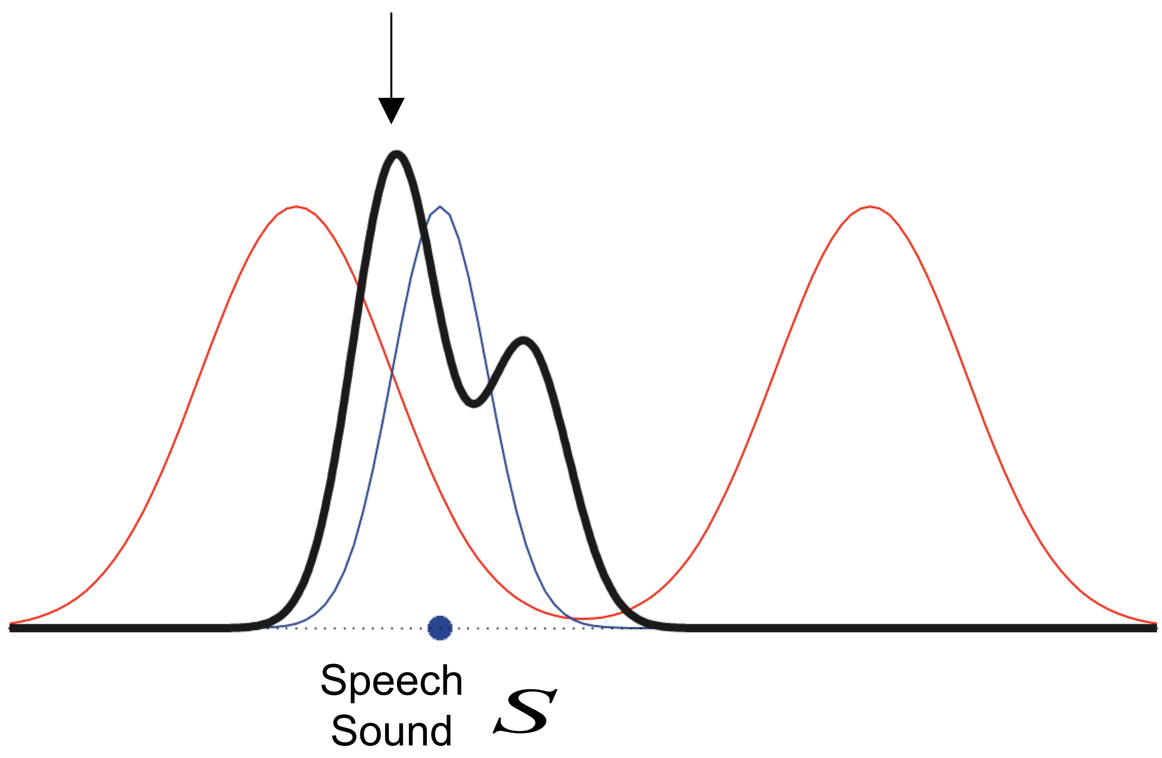
Qualitative Predictions



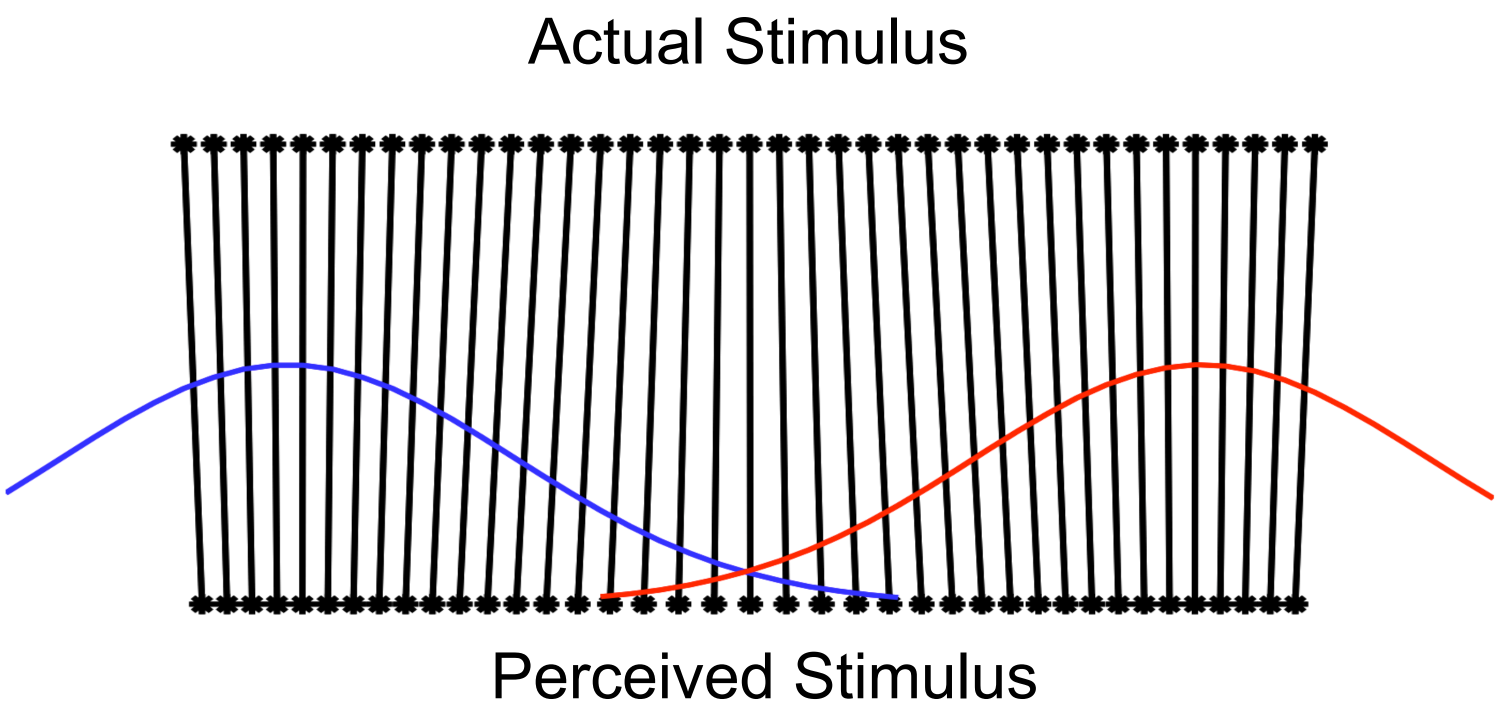
Categorical effects arise because listeners use their knowledge of phonetic categories to optimally infer a speaker's "target production" under conditions of uncertainty.

Meaningful and Noise Variance

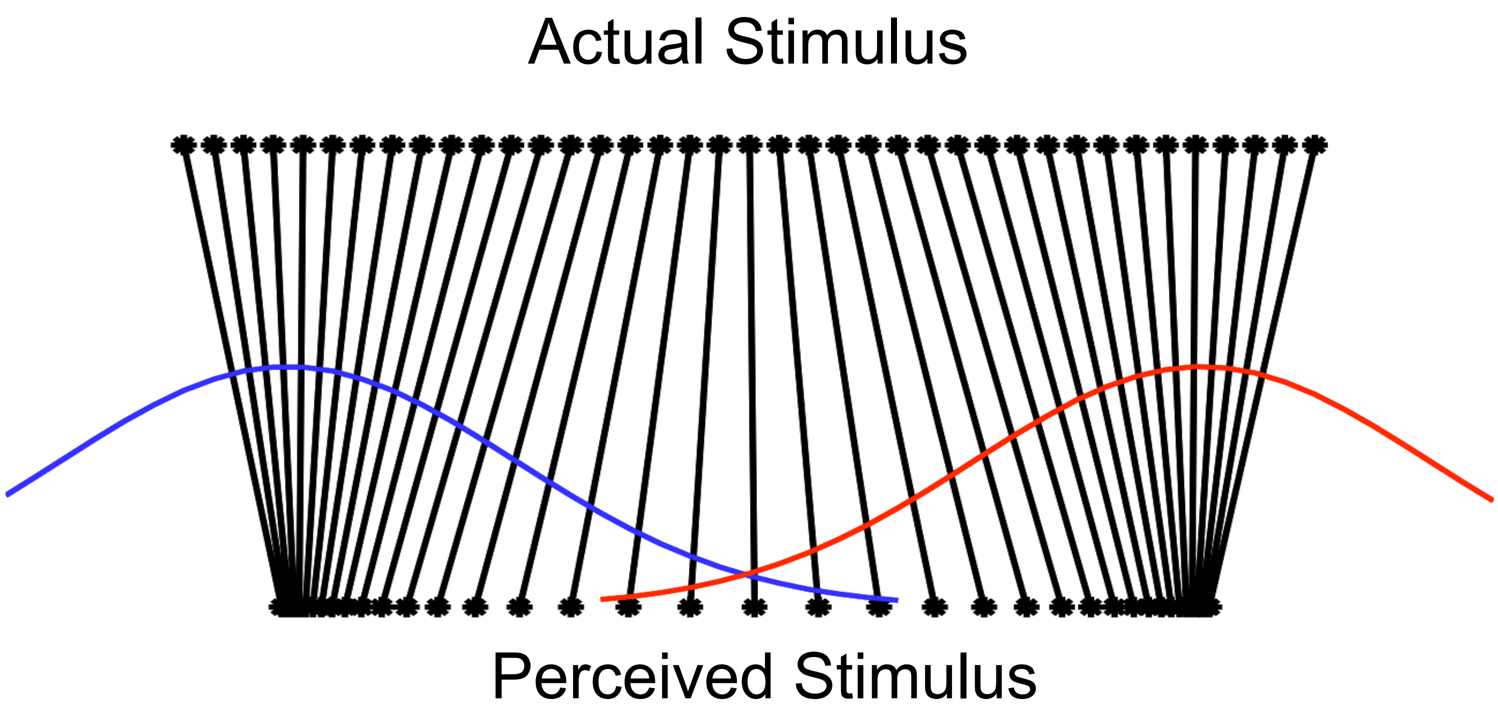
$$E\left[T \mid S, c\right] = \frac{\boxed{\sigma_c^{\,2}}\, S + \boxed{\sigma_S^{\,2}}\, \mu_c}{\sigma_c^{\,2} + \sigma_S^{\,2}}$$



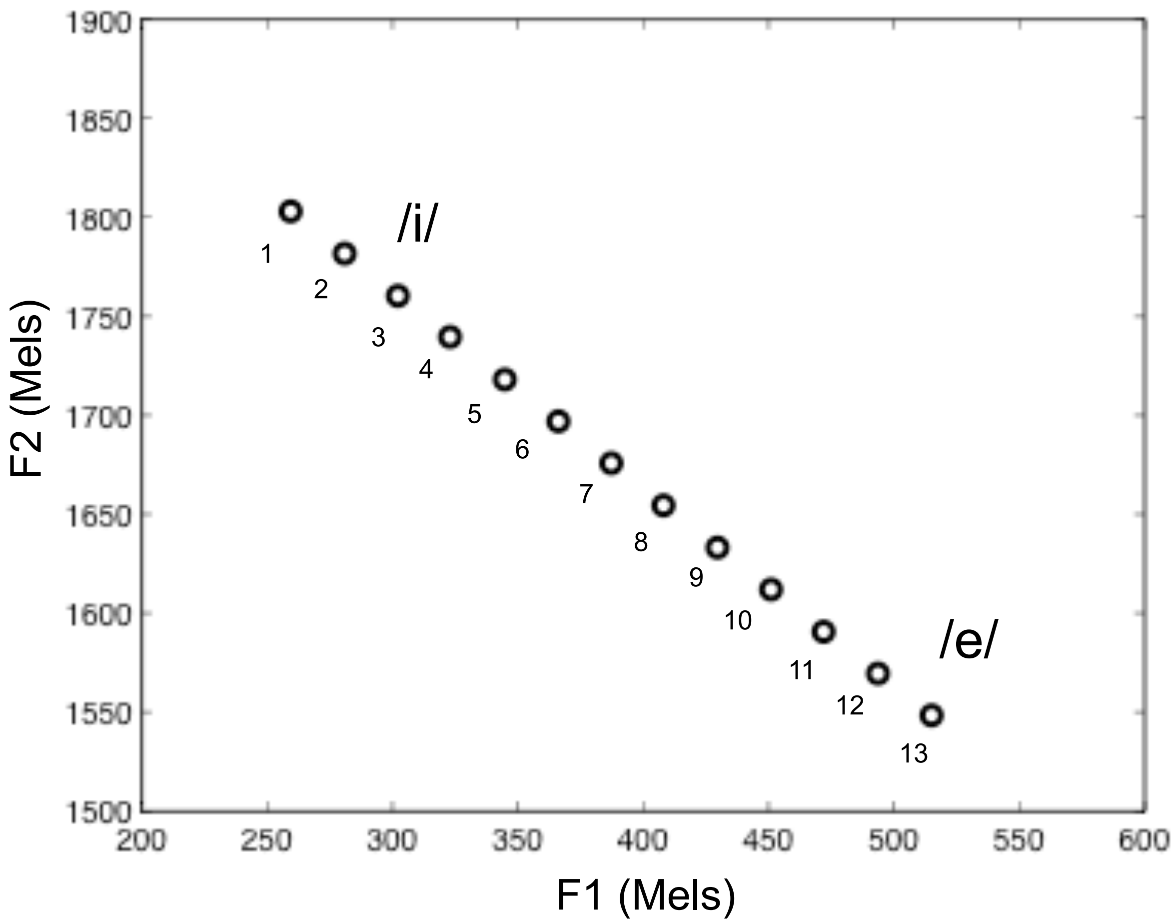
Low Noise Conditions



High Noise Conditions



Noise Experiment



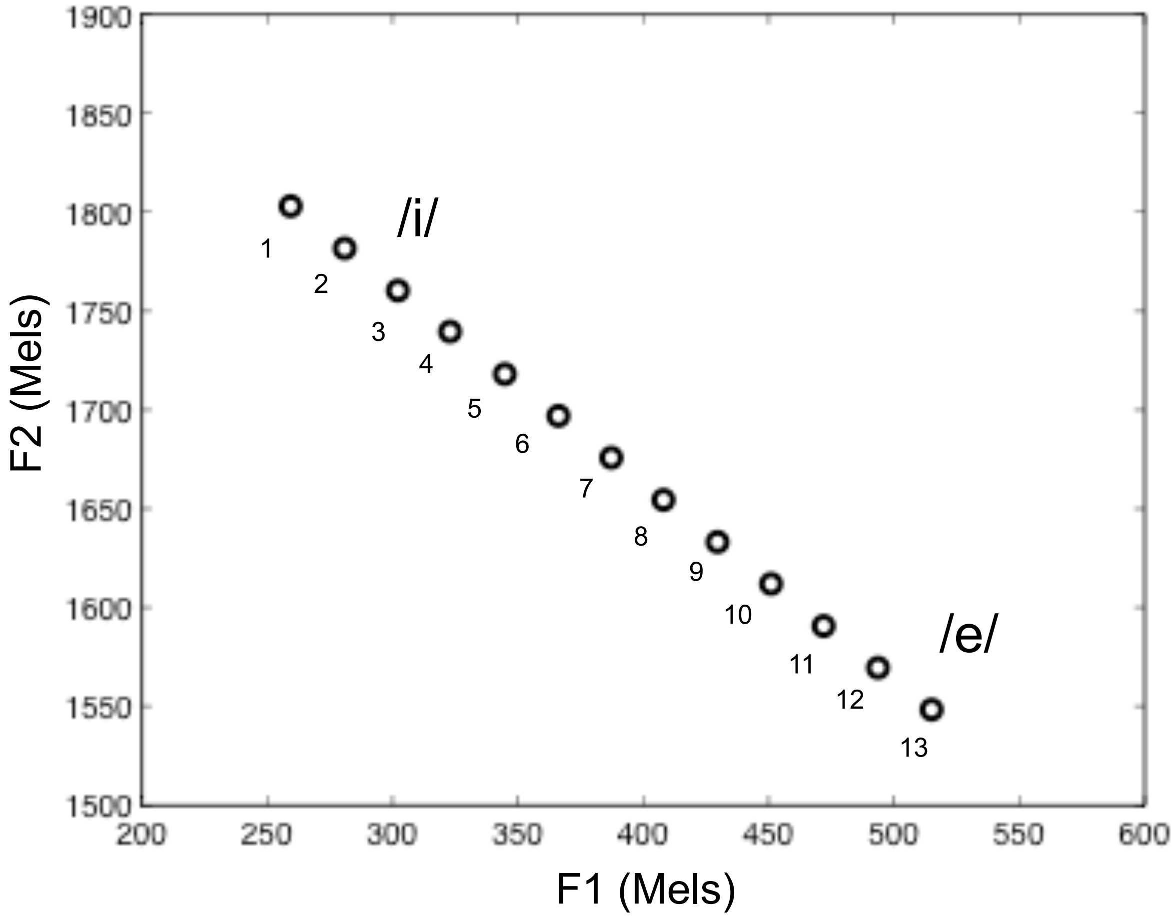
(Iverson & Kuhl, 1995; Feldman, Griffiths, & Morgan, 2009)

Noise Experiment

AX Discrimination Task:

Listeners hear all ordered pairs of stimuli

Determine whether pairs of sounds are identical



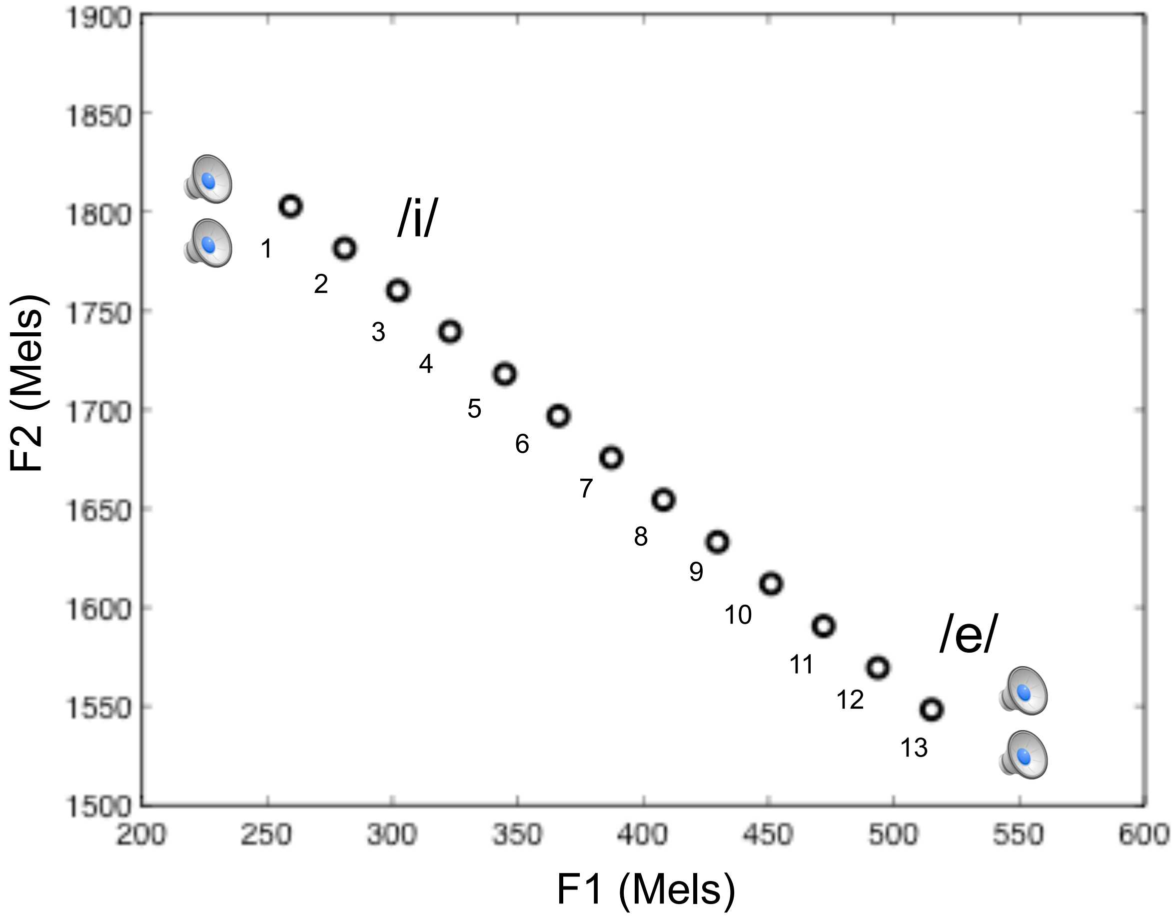
(Iverson & Kuhl, 1995; Feldman, Griffiths, & Morgan, 2009)

Noise Experiment

AX Discrimination Task:

Listeners hear all ordered pairs of stimuli

Determine whether pairs of sounds are identical



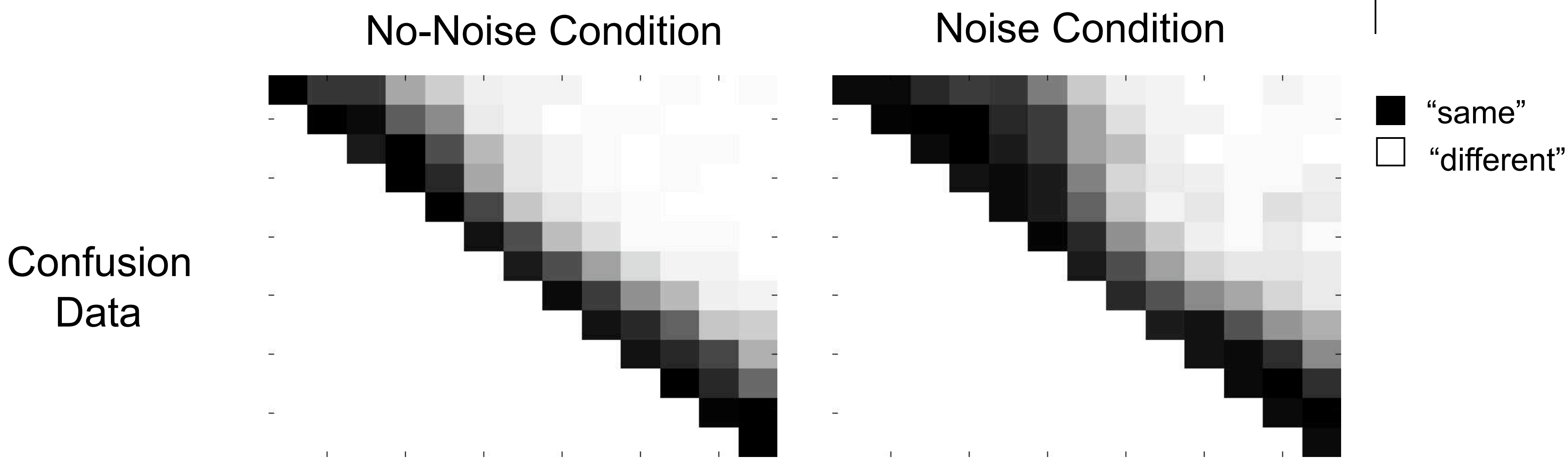
(Iverson & Kuhl, 1995; Feldman, Griffiths, & Morgan, 2009)

Confusion Matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	98.8	82.5	82.5	40.0	22.5	7.5	5.0	5.0	0.0	0.0	2.5	0.0	2.5
2		97.5	95.0	70.0	52.5	10.0	5.0	0.0	2.5	2.5	0.0	0.0	0.0
3			91.3	97.5	75.0	32.5	12.5	5.0	2.5	0.0	2.5	2.5	0.0
4				97.5	87.5	40.0	12.5	5.0	2.5	0.0	2.5	0.0	0.0
5					97.5	77.5	27.5	12.5	5.0	2.5	0.0	0.0	0.0
6						92.5	75.0	30.0	15.0	2.5	2.5	2.6	0.0
7							91.3	75.0	42.5	17.5	5.0	5.0	0.0
8								95.0	80.0	50.0	32.5	7.5	5.0
9									93.8	87.5	67.5	27.5	22.5
10										92.5	87.5	76.9	37.5
11											97.5	87.5	65.0
12												96.3	97.5
13													100

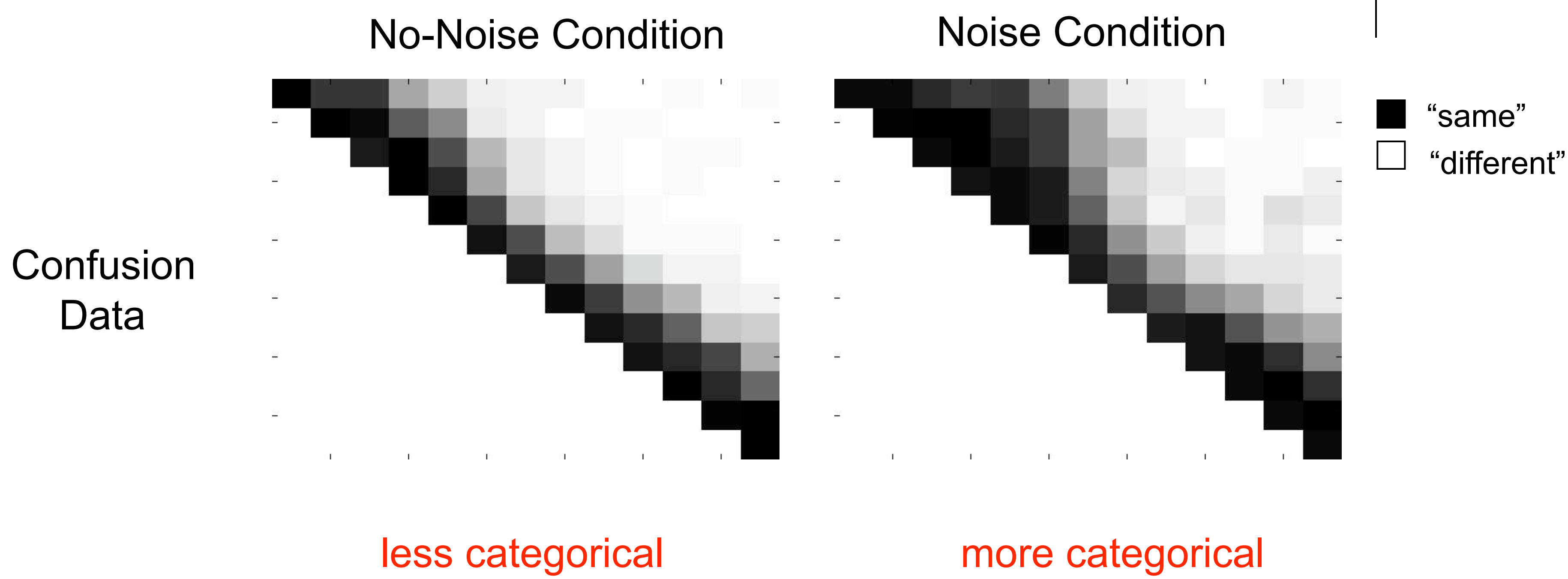
(Feldman, Griffiths, & Morgan, 2009)

Noise Experiment: Results



(Feldman, Griffiths, & Morgan, 2009)

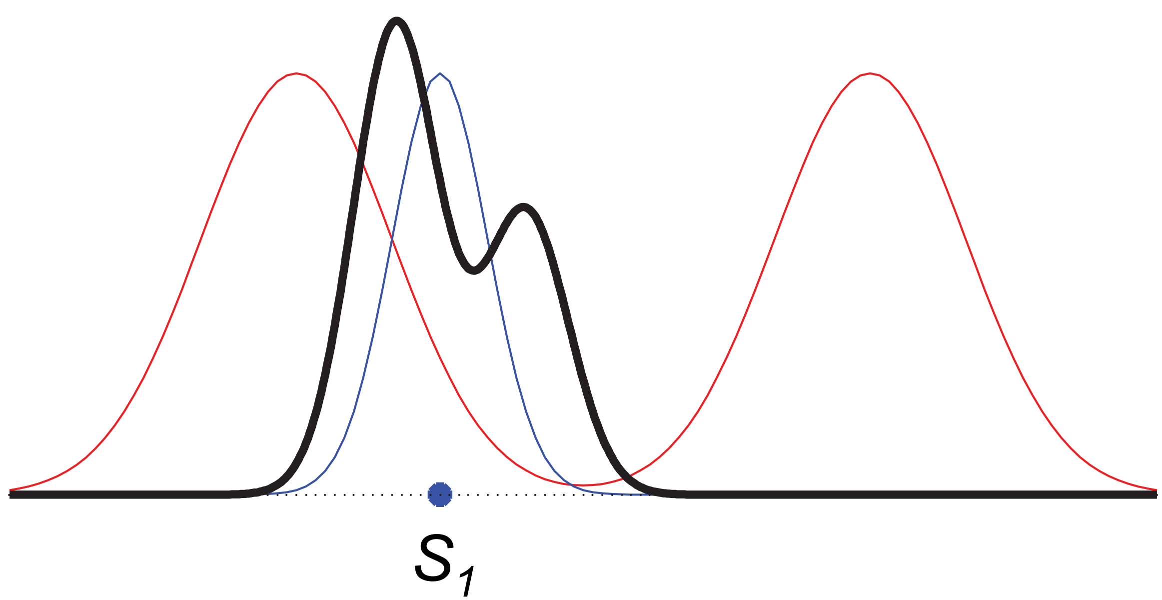
Noise Experiment: Results



(Feldman, Griffiths, & Morgan, 2009)

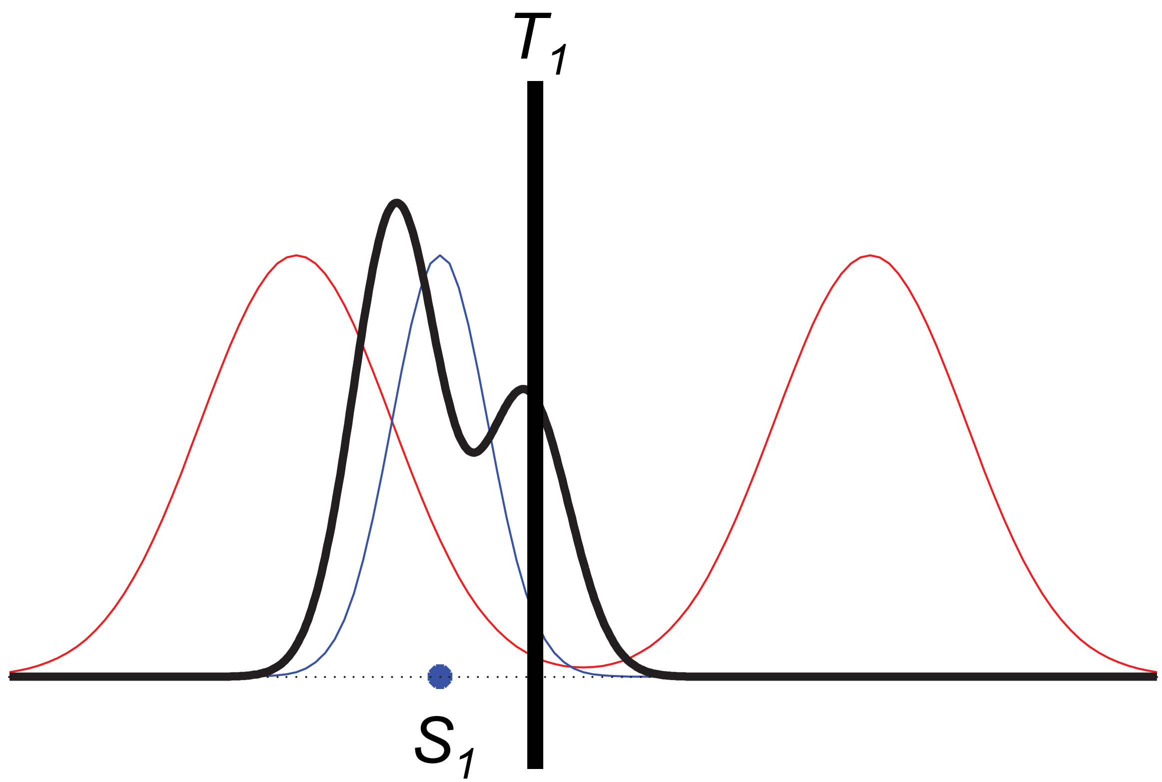
AX Discrimination Model

- Sample a “target production” from the posterior for each stimulus



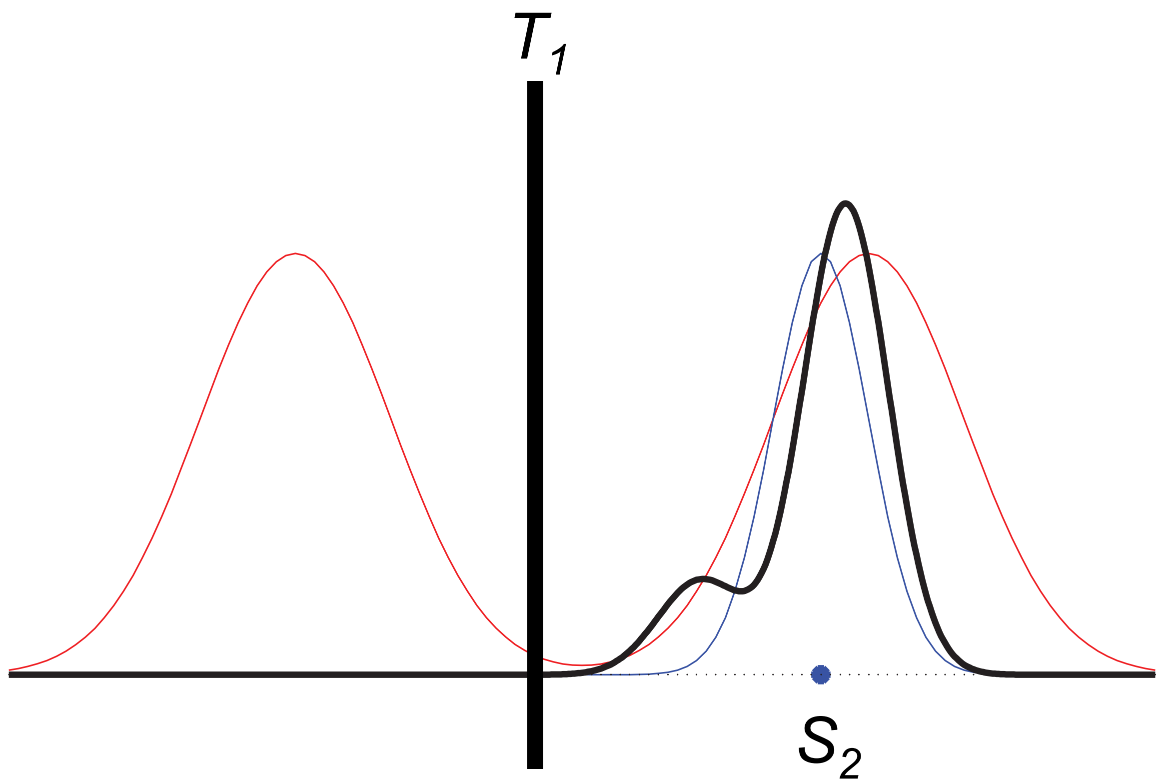
AX Discrimination Model

- Sample a “target production” from the posterior for each stimulus



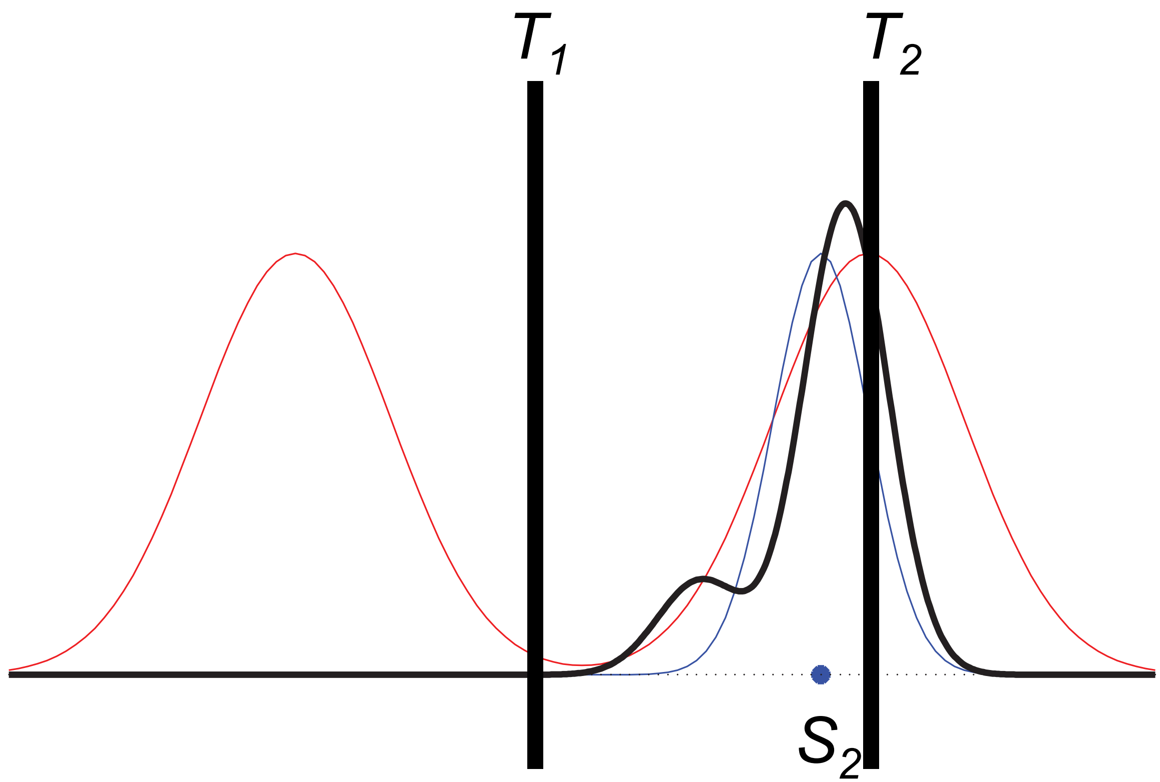
AX Discrimination Model

- Sample a “target production” from the posterior for each stimulus



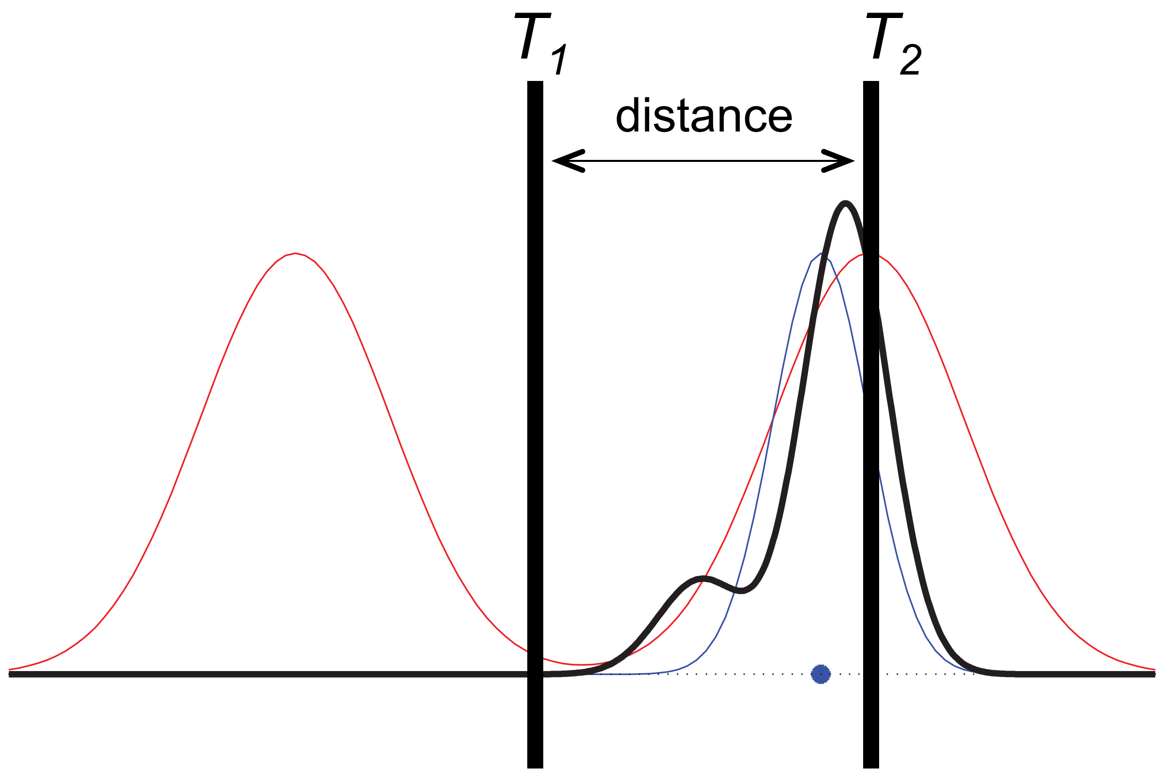
AX Discrimination Model

- Sample a “target production” from the posterior for each stimulus



AX Discrimination Model

- Sample a “target production” from the posterior for each stimulus
- Compare distance between target productions to threshold ε



Estimating Model Parameters

Identification data:

Listeners inferring category c from speech sound S

$$p(c \mid S) \propto p(S \mid c)\, p(c)$$

$$p(c \mid S) \propto N\left(\mu_c, \sigma_c^{\ 2} + \sigma_S^{\ 2}\right)(0.5)$$

Estimating Model Parameters

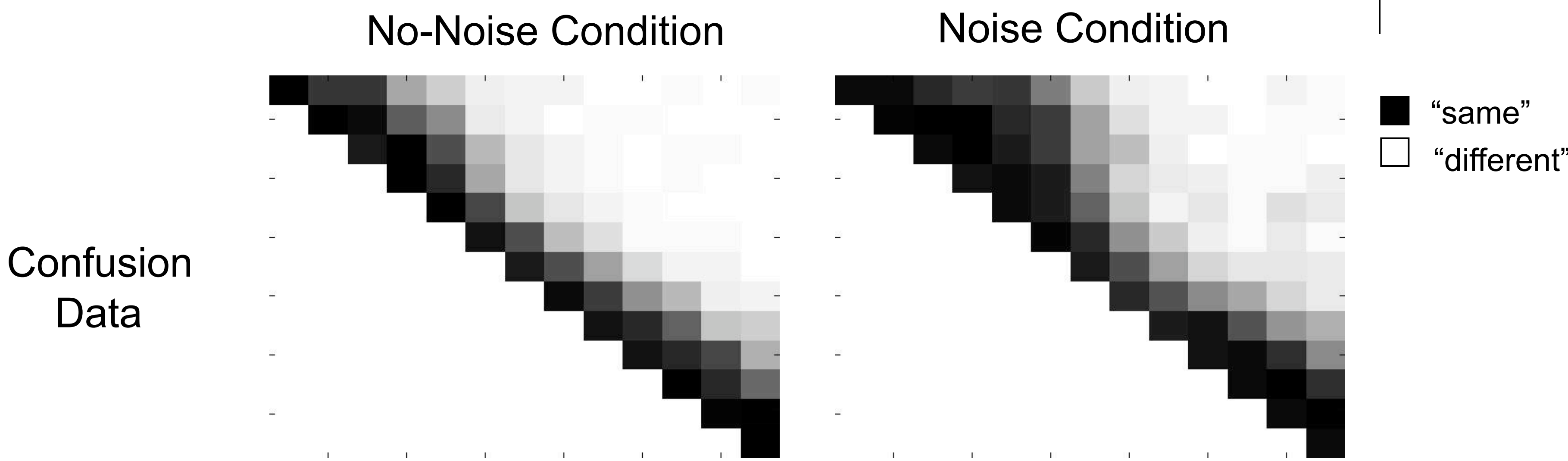
Discrimination data:

Listeners inferring target production T from speech sound S

$$p(T \mid S,c) \propto p(S \mid T)\, p(T \mid c)$$

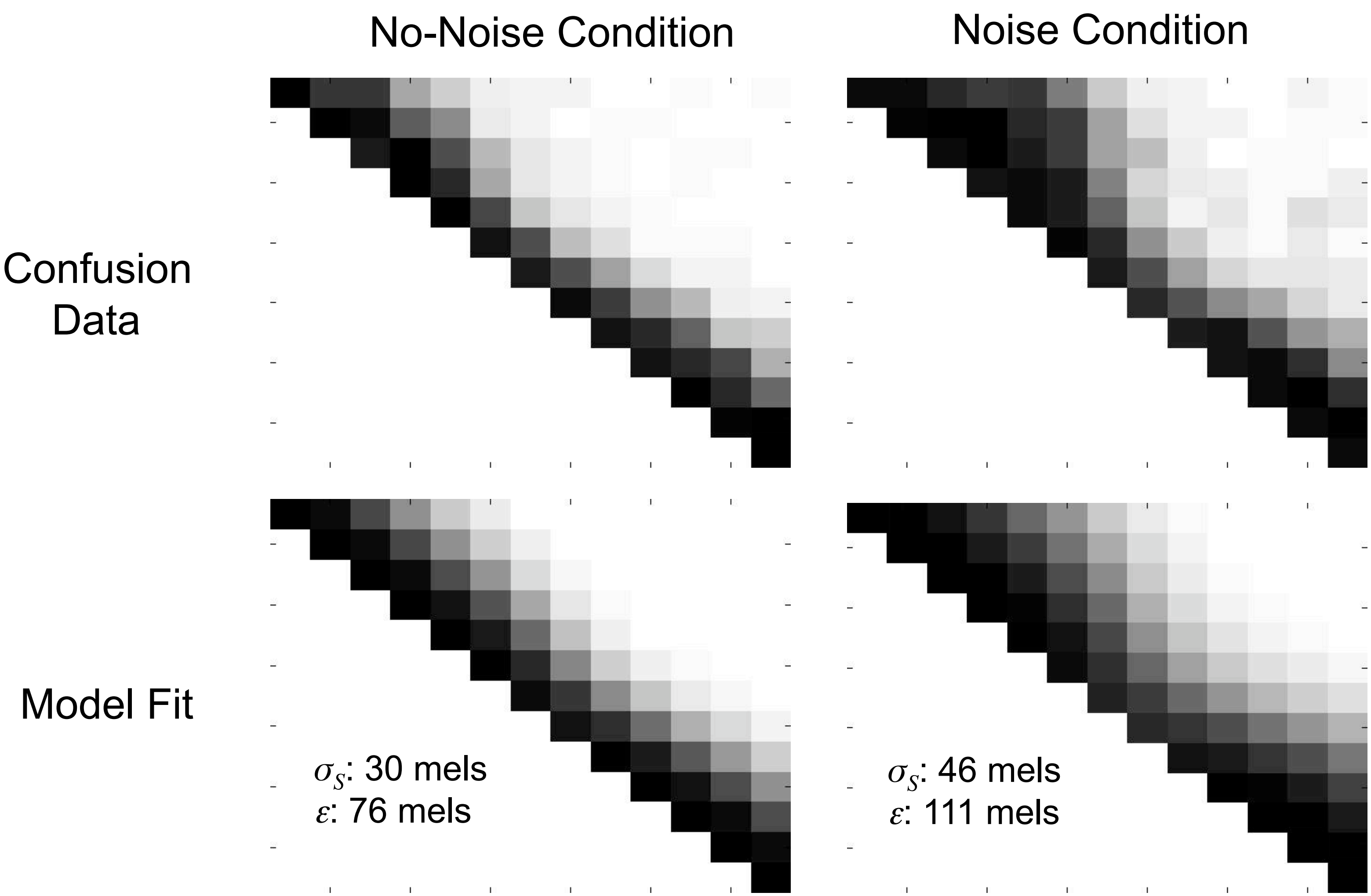
$$p(T \mid S,c) \propto N\left(T,\boxed{\sigma_S}^2\right) N\left(\mu_c,\boxed{\sigma_c}^2\right)$$

Noise Experiment: Results



(Feldman, Griffiths, & Morgan, 2009)

Noise Experiment: Results



"same"
"different"

Fits data significantly better than a nested model that doesn't vary the noise parameter, $p<0.001$

(Feldman, Griffiths, & Morgan, 2009)

Noise Experiment: Discussion

- Model accounts significantly for differences in noise by varying the noise parameter
- Vowels look more like consonants in noisy conditions (see also Pisoni, 1975; Repp, Healy, & Crowder, 1979)
- Can the same explanation account for differences between consonants and vowels?

Consonants vs. Vowels

Strong Effect

of Categories

Low $\sigma_c^{\,2}$

High $\sigma_S^{\,2}$

Weak Effect

of Categories

High $\sigma_c^{\,2}$

Low $\sigma_S^{\,2}$

Ratio of $\sigma_c^{\,2}$ to $\sigma_S^{\,2}$

Consonants vs. Vowels

Strong Effect

of Categories

Low σ_c^2

High σ_S^2

Weak Effect

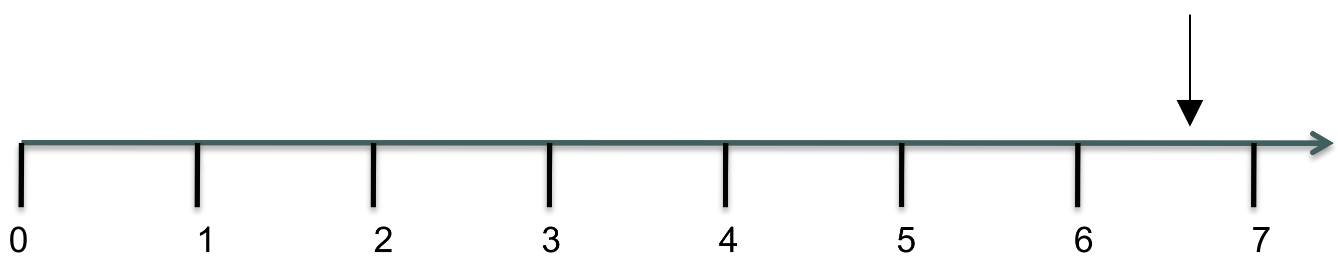
of Categories

High σ_c^2

Low σ_S^2

/e/,/i/

6.69



Ratio of σ_c^2 to σ_S^2

(Kronrod, Coppess, & Feldman, 2012)

Consonants vs. Vowels

Strong Effect

of Categories

Low σ_c^2

High σ_S^2

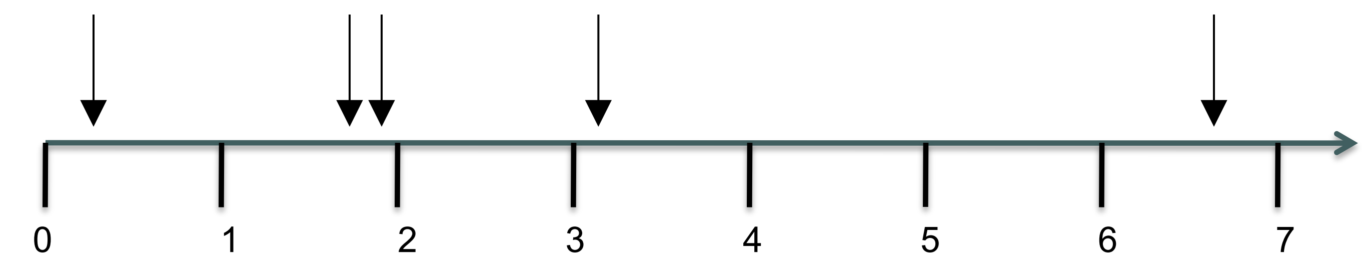
Weak Effect

of Categories

High σ_c^2

Low σ_S^2

/b/	/ʃ/	/s/	/p/	/e/,/i/
0.17	1.86	1.93	3.09	6.69



Ratio of σ_c^2 to σ_S^2

(Kronrod, Coppess, & Feldman, 2012)

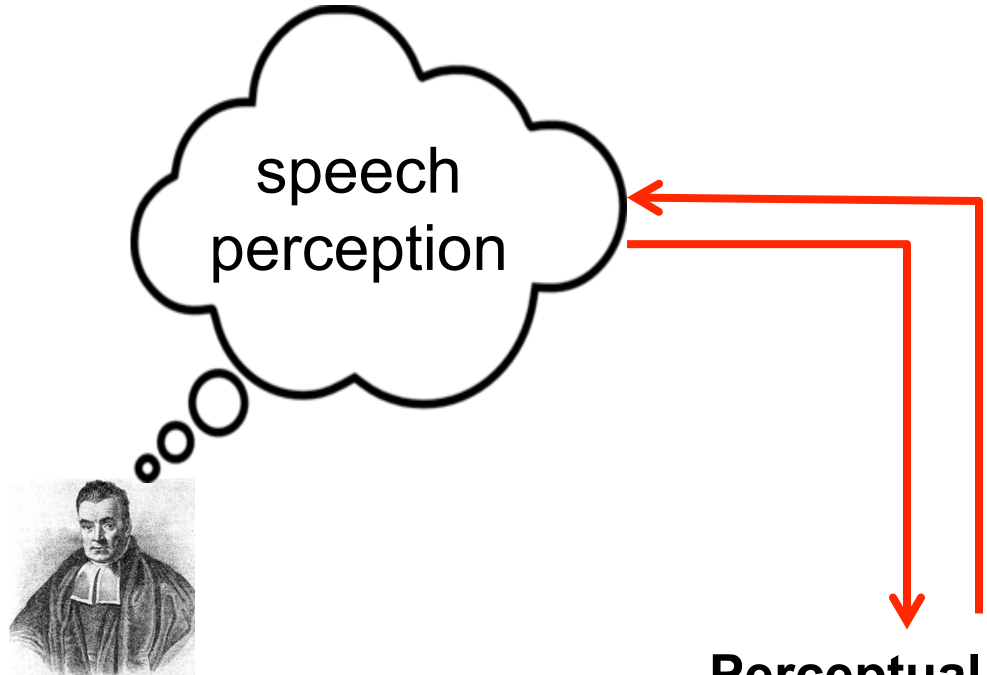
Modeling and Empirical Results

- Reproduces discrimination data from vowels, stop consonants, and fricatives (Feldman, Griffiths, & Morgan, 2009; Kronrod, Coppess, & Feldman, 2012)

- Correctly predicts stronger perceptual bias in noisy conditions than quiet conditions (Feldman et al., 2009)

- Captures differences in the strength of categorical effects with a single parameter (Kronrod, Coppess, & Feldman, 2012)

A Model of Speech Perception



Speech corpora
(prior distribution)

Perceptual data
(posterior distribution)

Outline

- Behavioral data in speech perception
- Cognitive model of speech perception
- Adapting the model to speech corpora

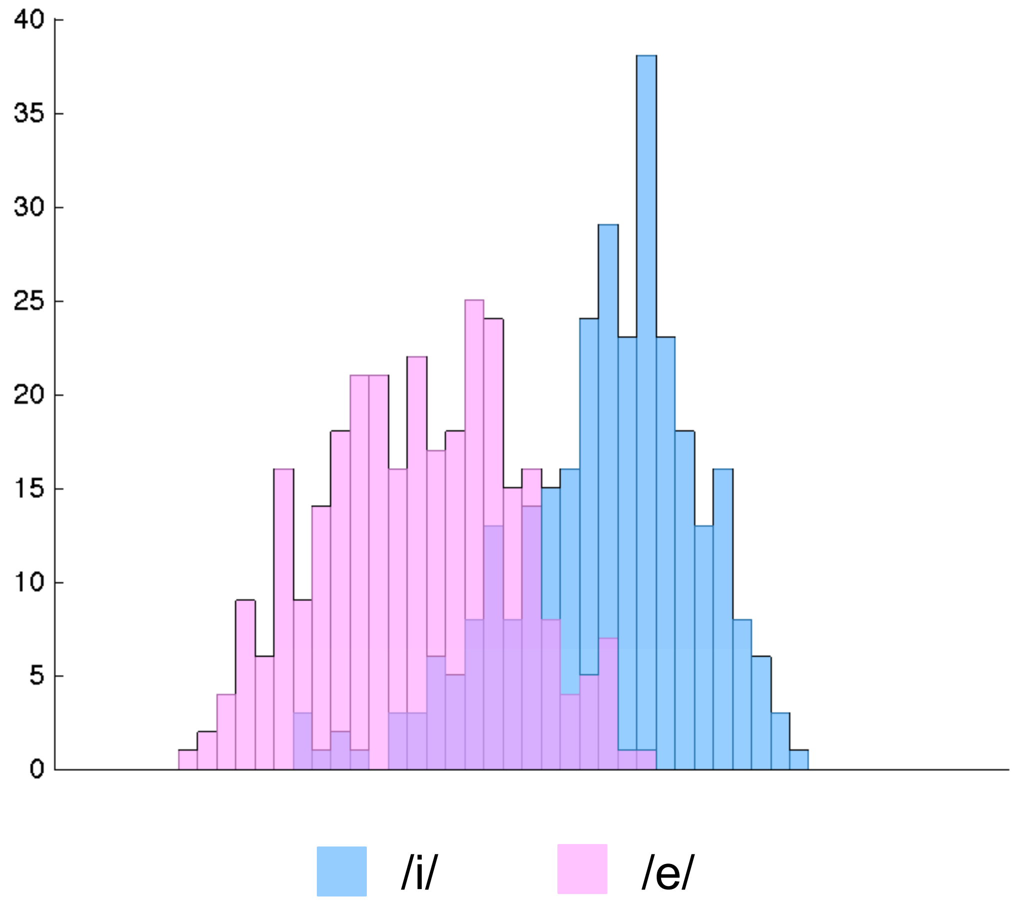
Lei Shi

Tom Griffiths

Adam Sanborn

- A case study: Speaker normalization

Speech Corpora as a Prior



Speech Corpora as a Prior

Challenge:

Sounds in speech corpora don’t fall neatly into Gaussian distributions

Speech Corpora as a Prior

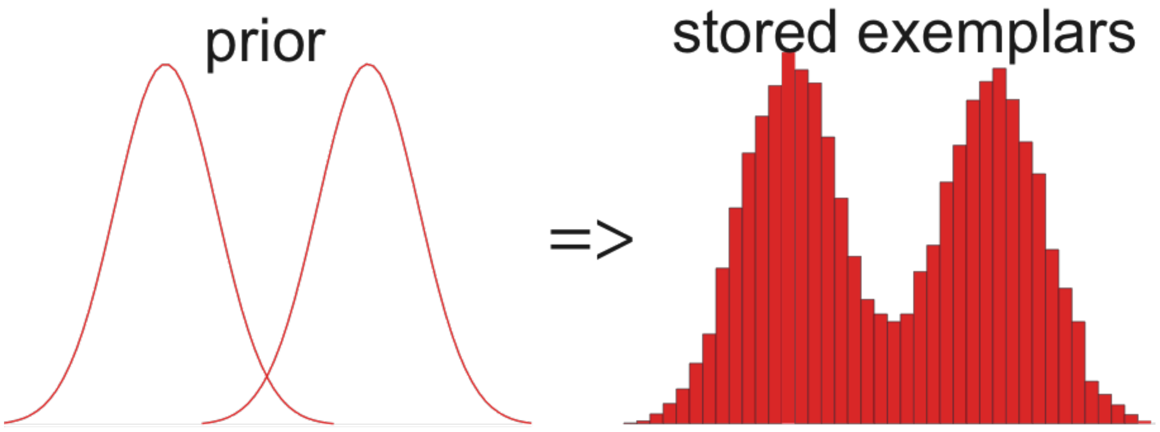
Challenge:

Sounds in speech corpora don’t fall neatly into Gaussian distributions

Solution:

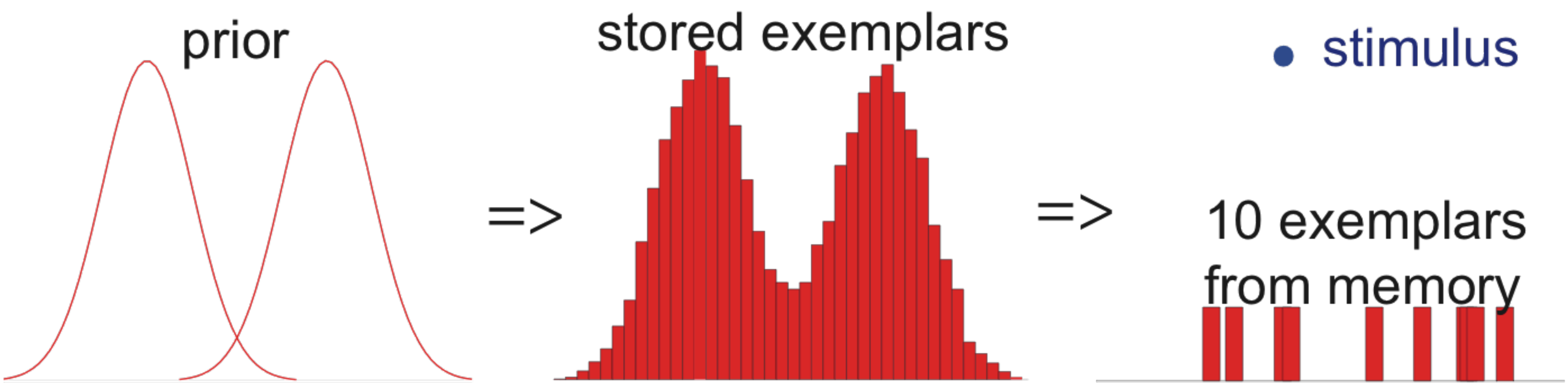
Use samples from the prior distribution to obtain a sample from the posterior distribution

Importance Sampling



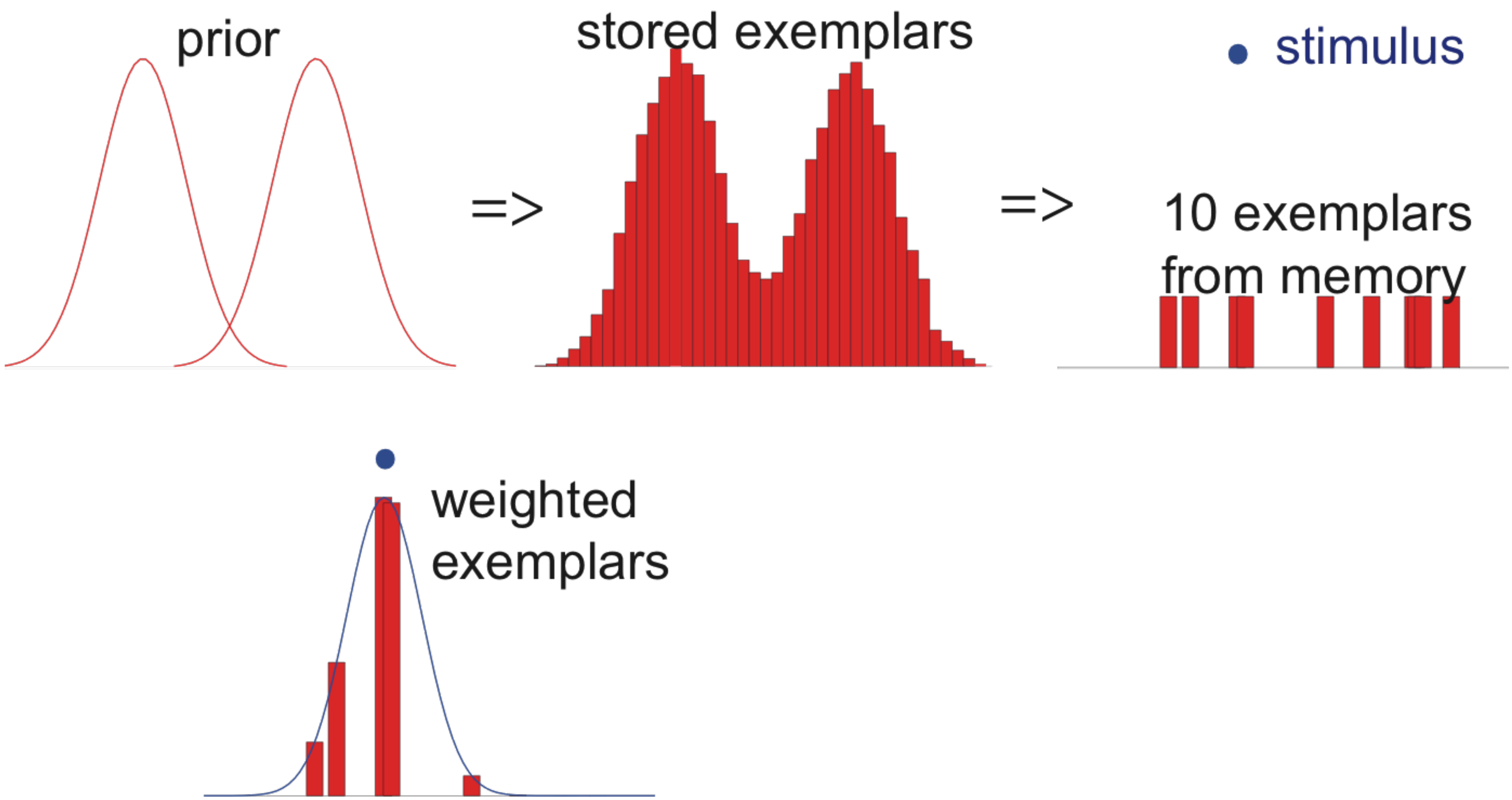
(Shi, Griffiths, Feldman, & Sanborn, 2010)

Importance Sampling



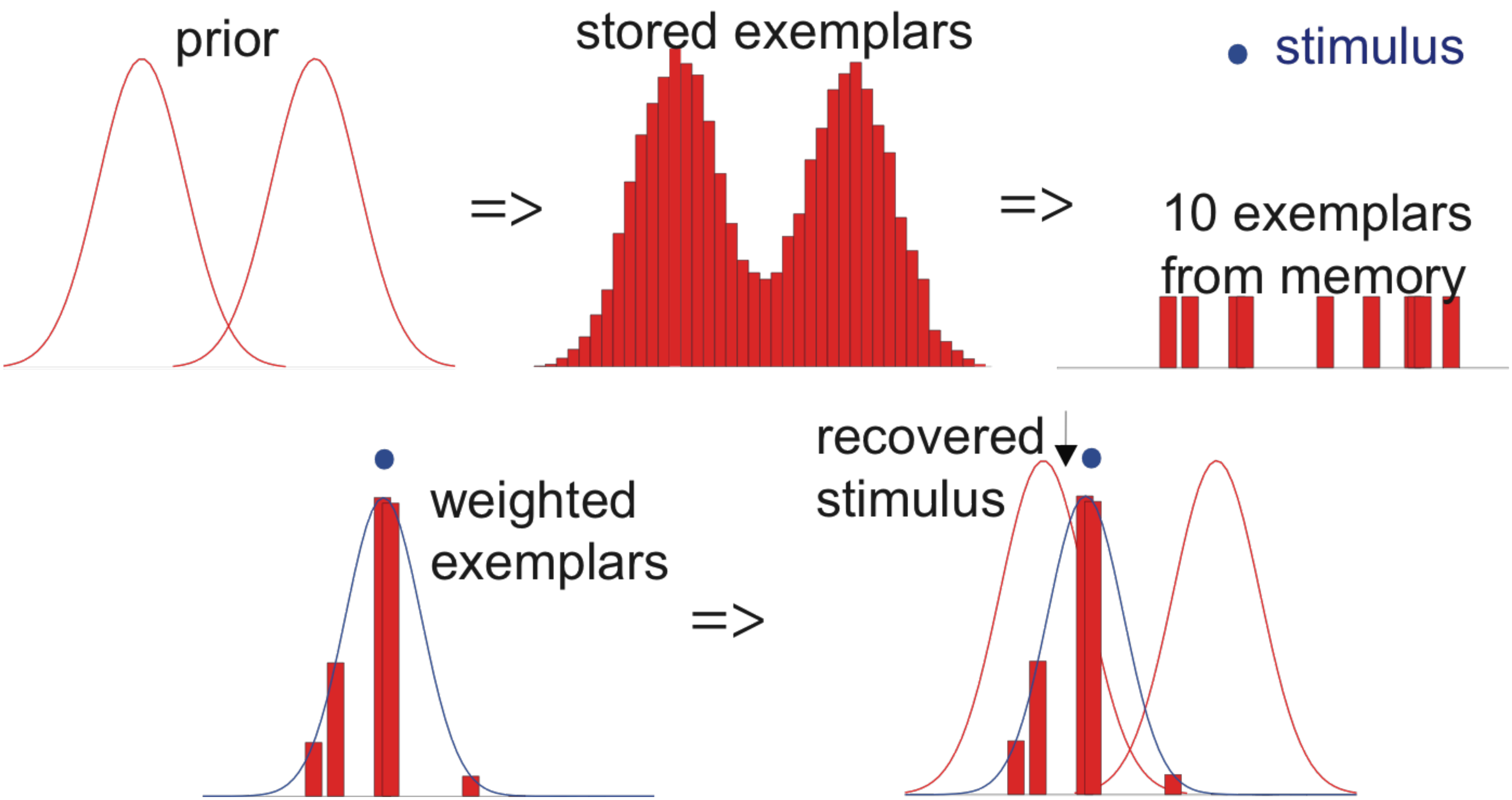
(Shi, Griffiths, Feldman, & Sanborn, 2010)

Importance Sampling



(Shi, Griffiths, Feldman, & Sanborn, 2010)

Importance Sampling



(Shi, Griffiths, Feldman, & Sanborn, 2010)

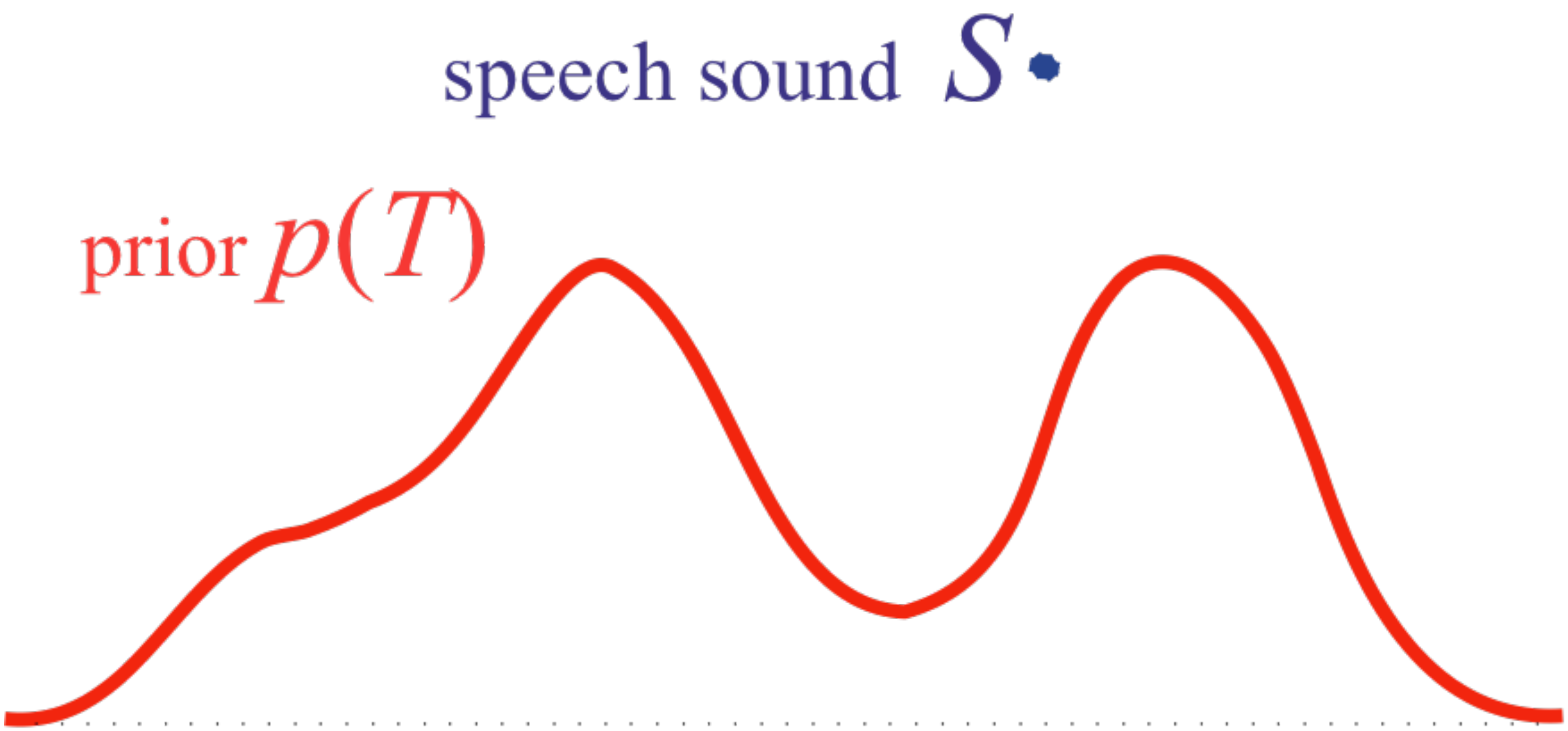
Importance Sampling

Exemplar models provide a general way of approximating Bayesian inference

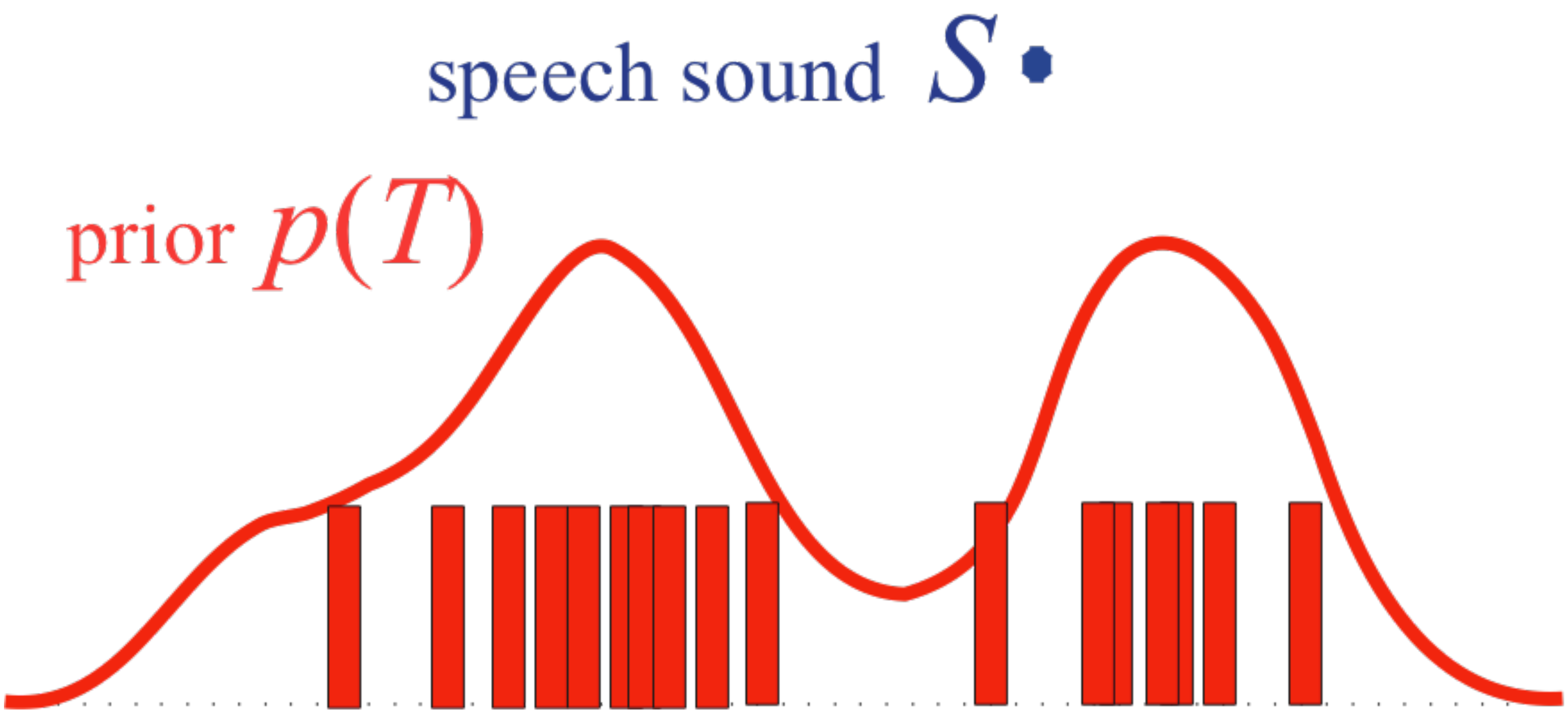
- Sample exemplars from the prior distribution
- Weight each exemplar by its likelihood
- These weighted samples behave like samples from the posterior distribution

(Shi, Griffiths, Feldman, & Sanborn, 2010)

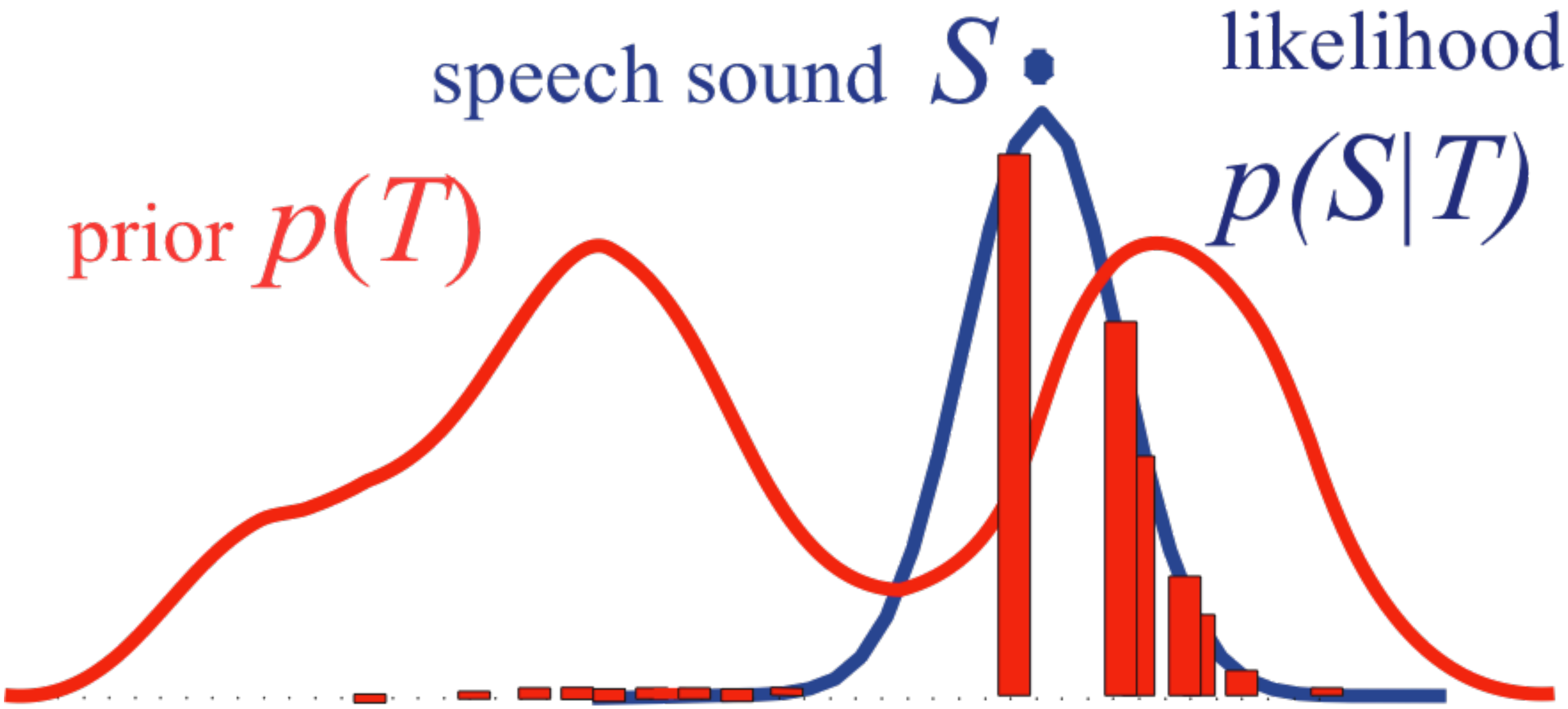
Importance Sampling



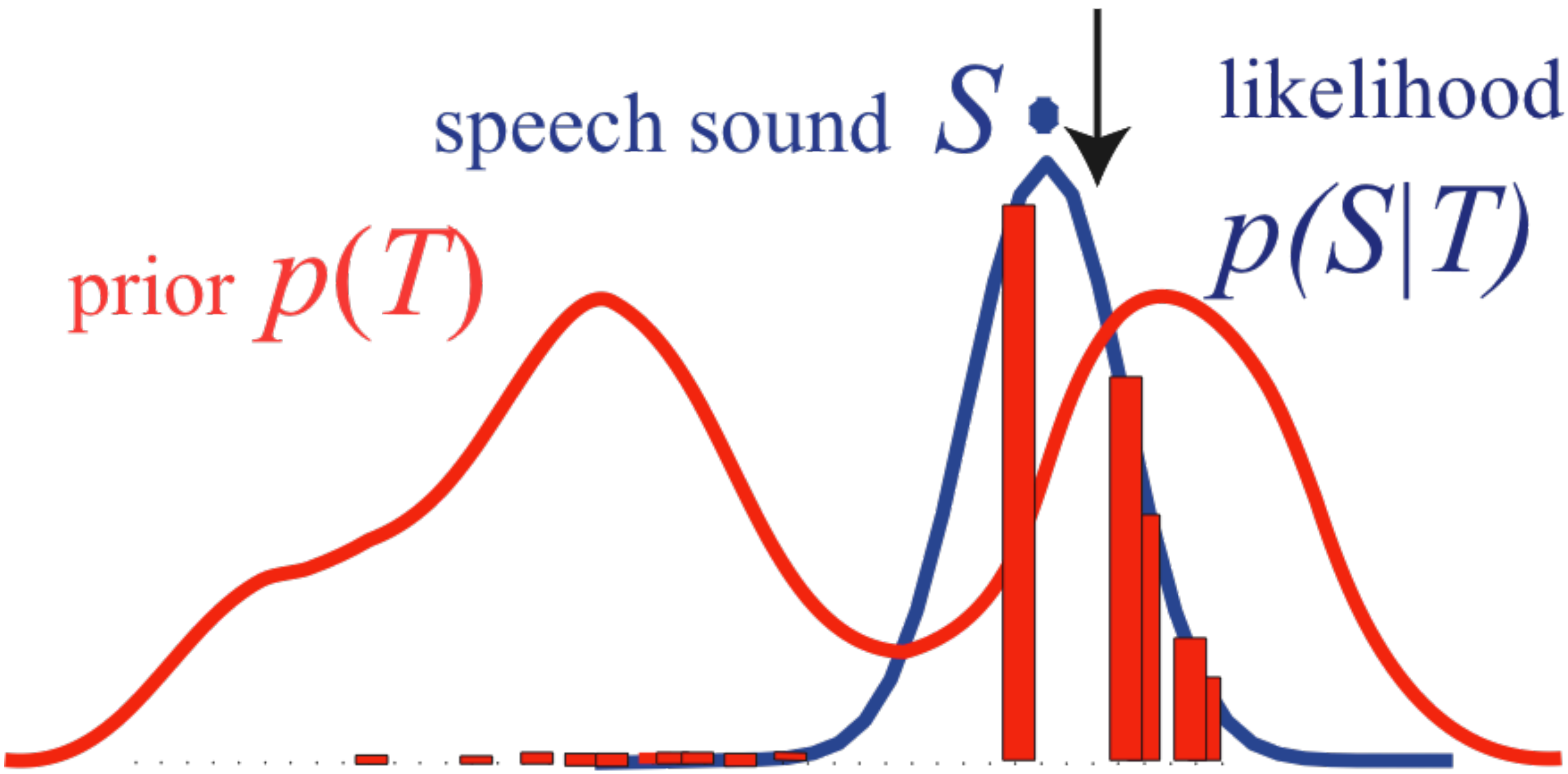
Importance Sampling



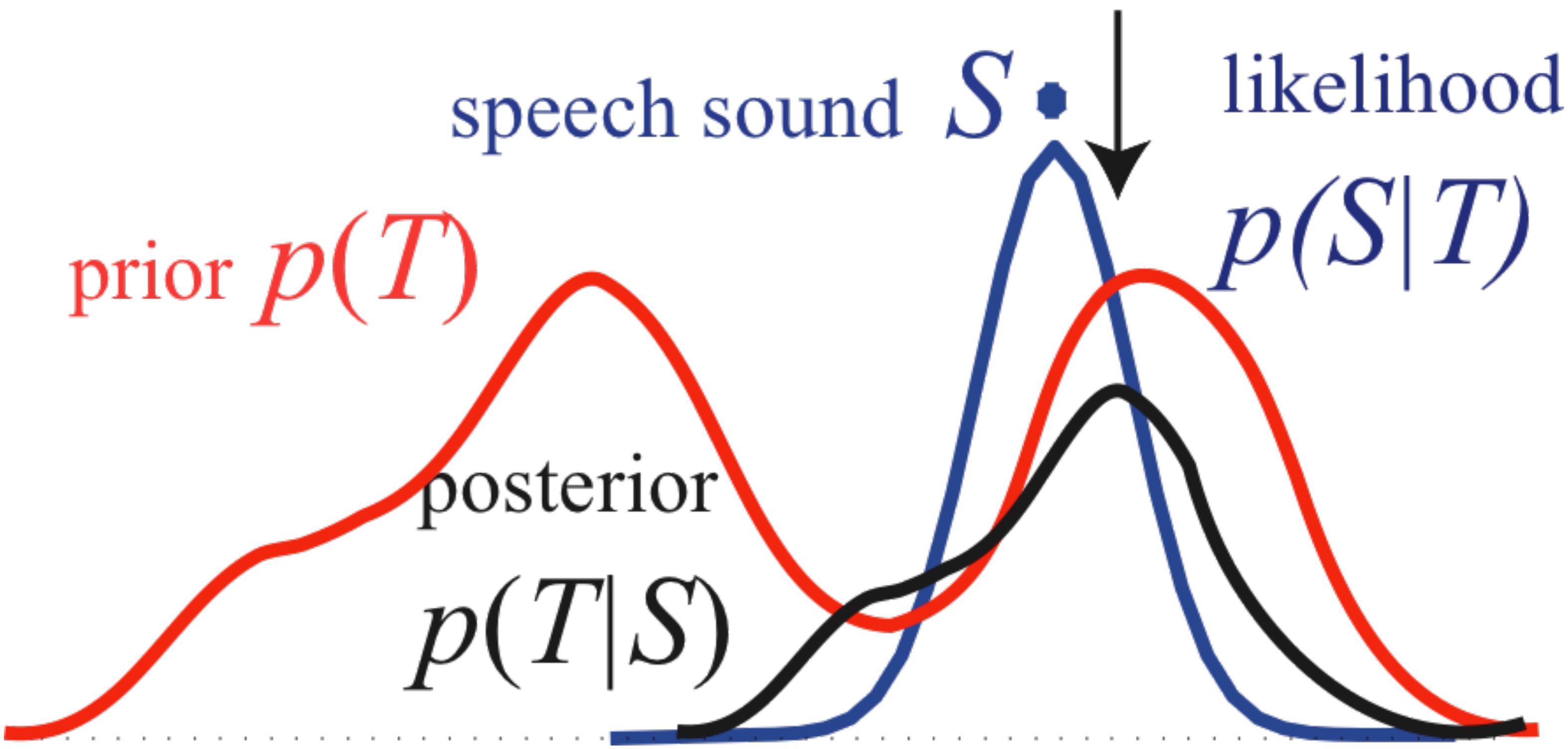
Importance Sampling



Importance Sampling



Importance Sampling

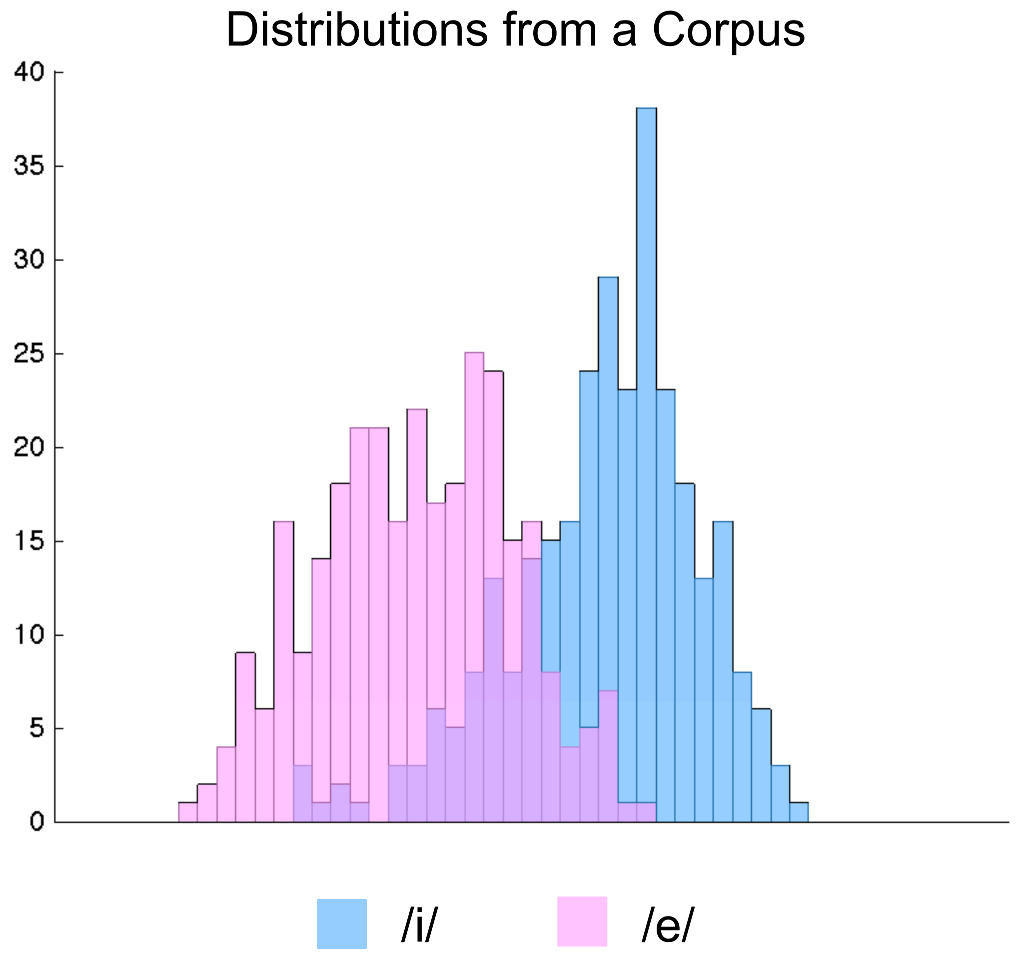


Given samples from the prior and a likelihood (noise) function, we can predict how people will perceive experimental stimuli

Given samples from the prior and a likelihood (noise) function, we can predict how people will perceive experimental stimuli

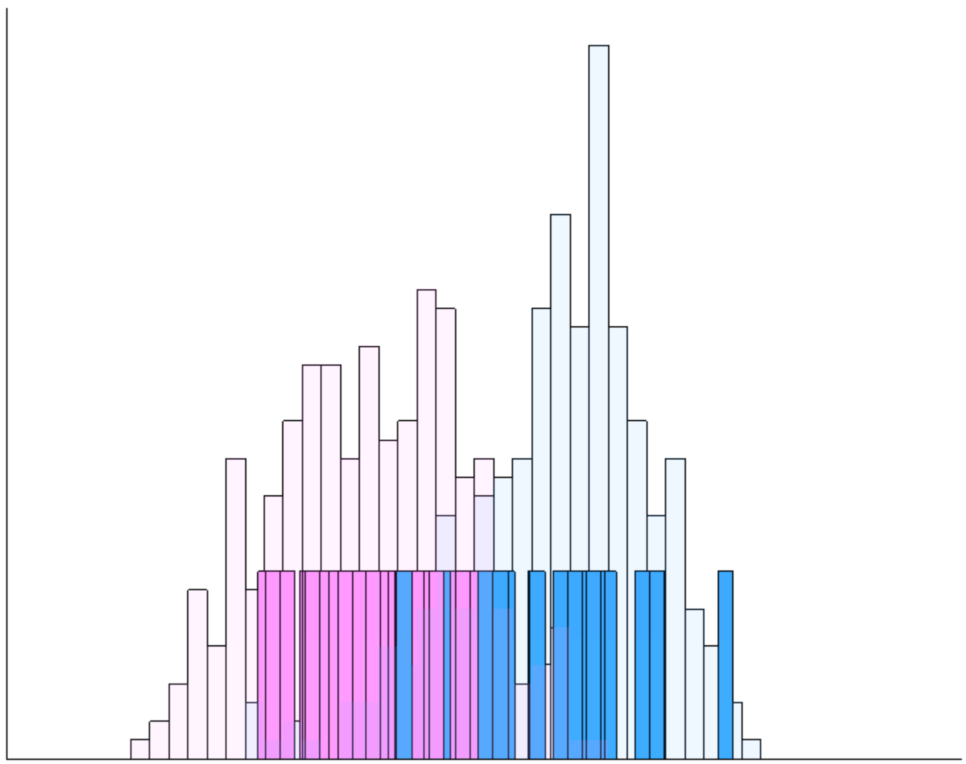
Speech corpora consist of samples from the prior distribution!

Simulation of an AX Trial

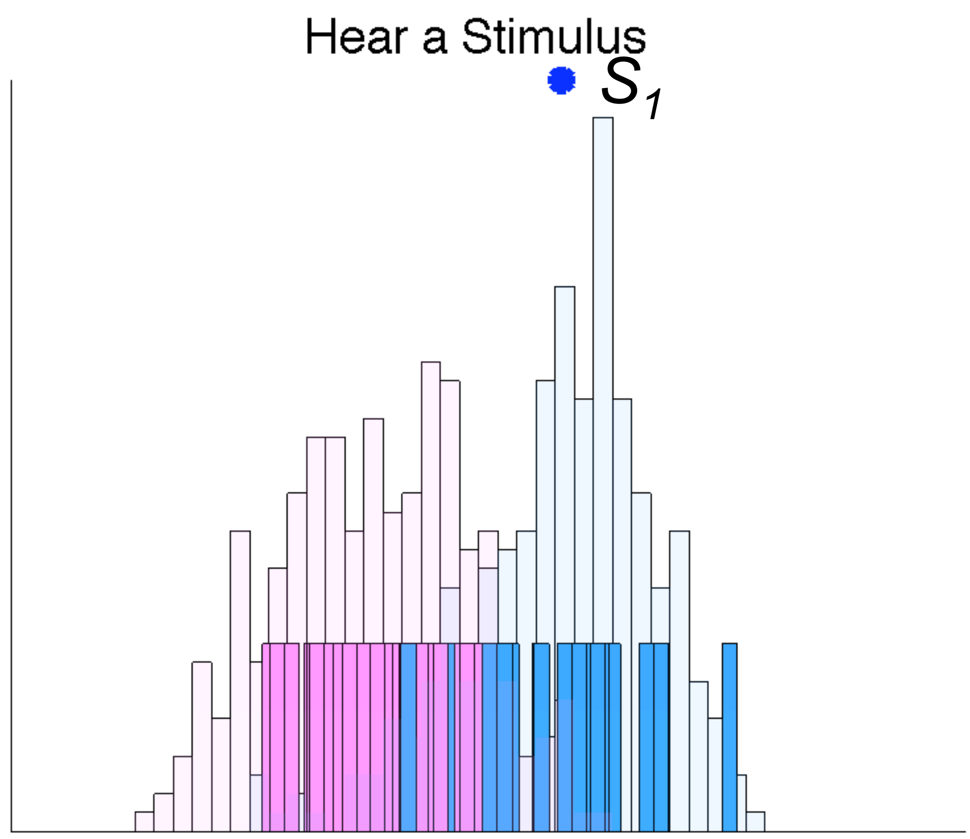


Perceiving First Stimulus

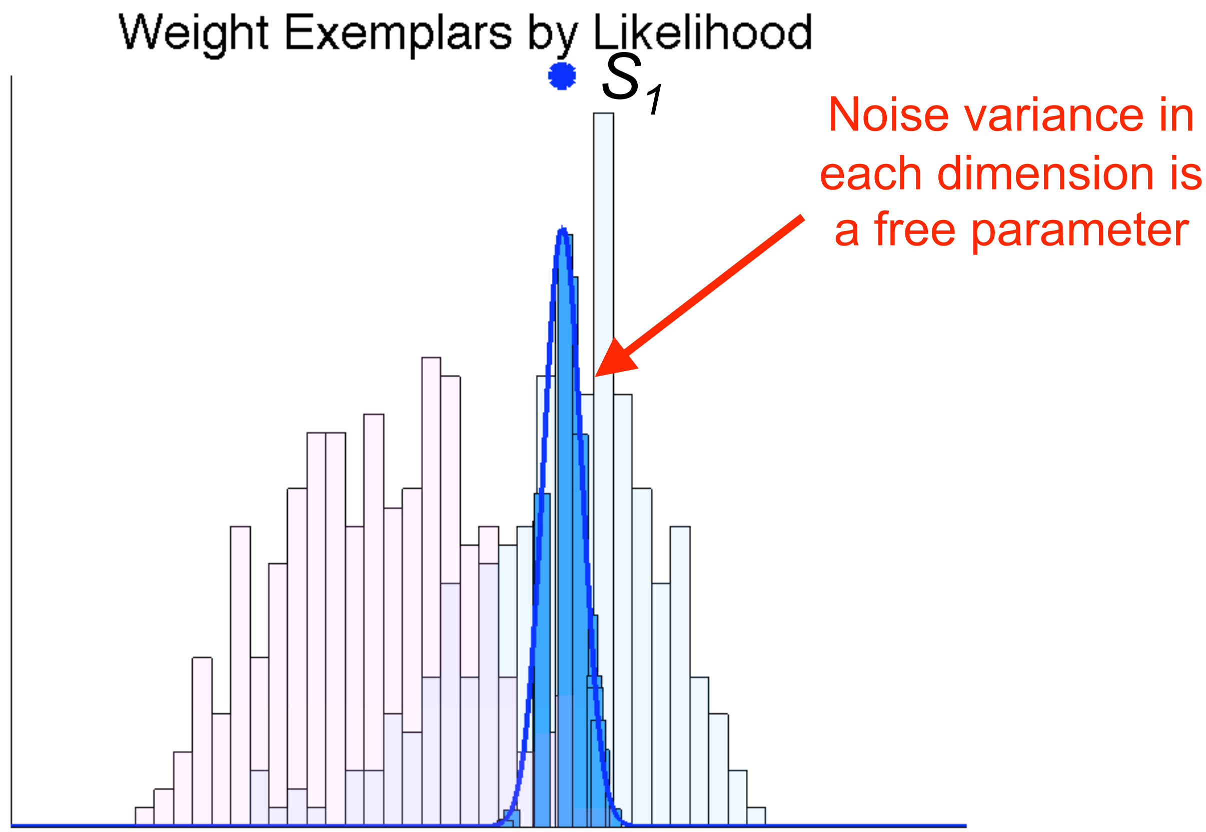
Sample Exemplars



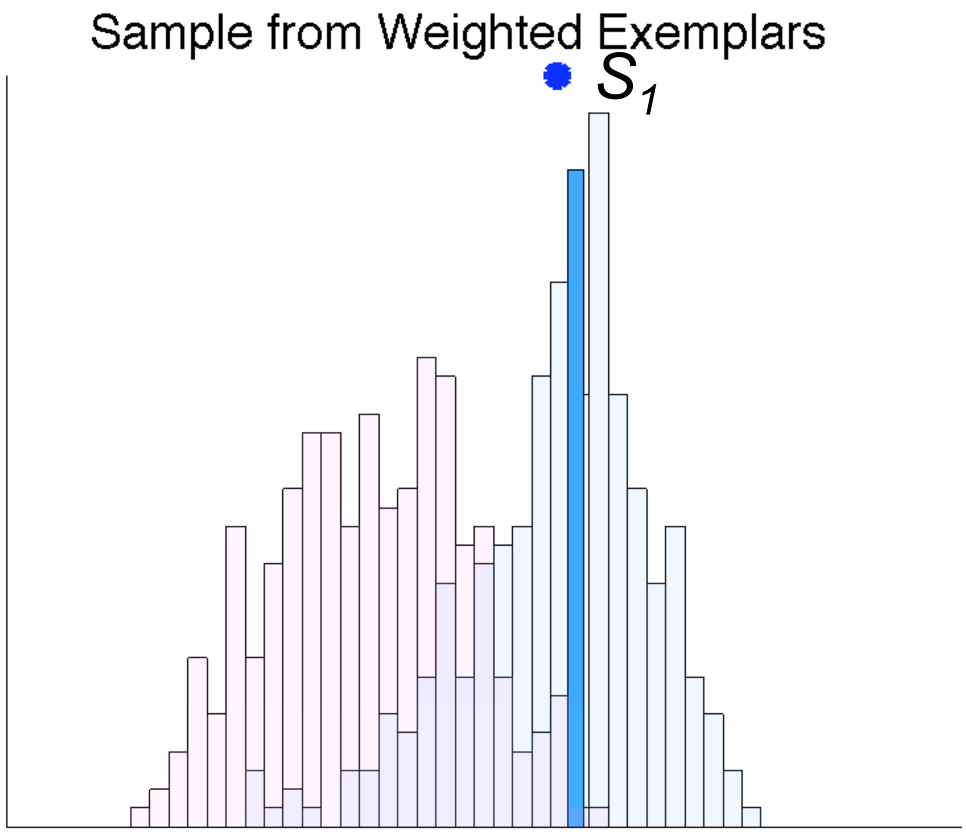
Perceiving First Stimulus



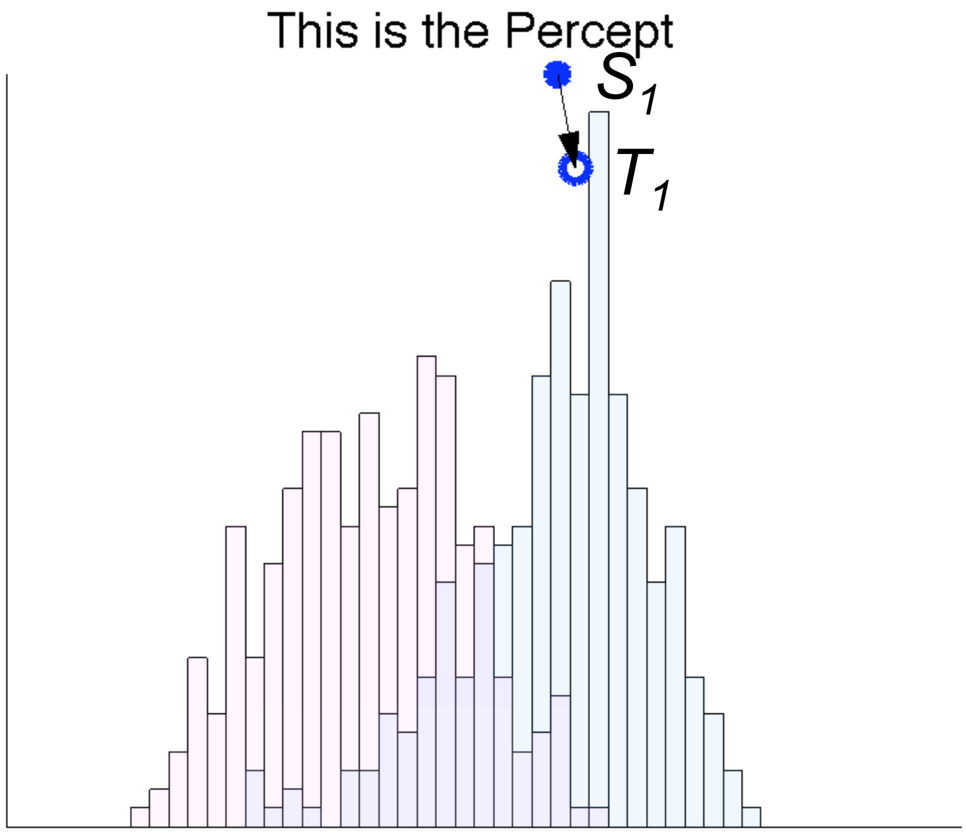
Perceiving First Stimulus



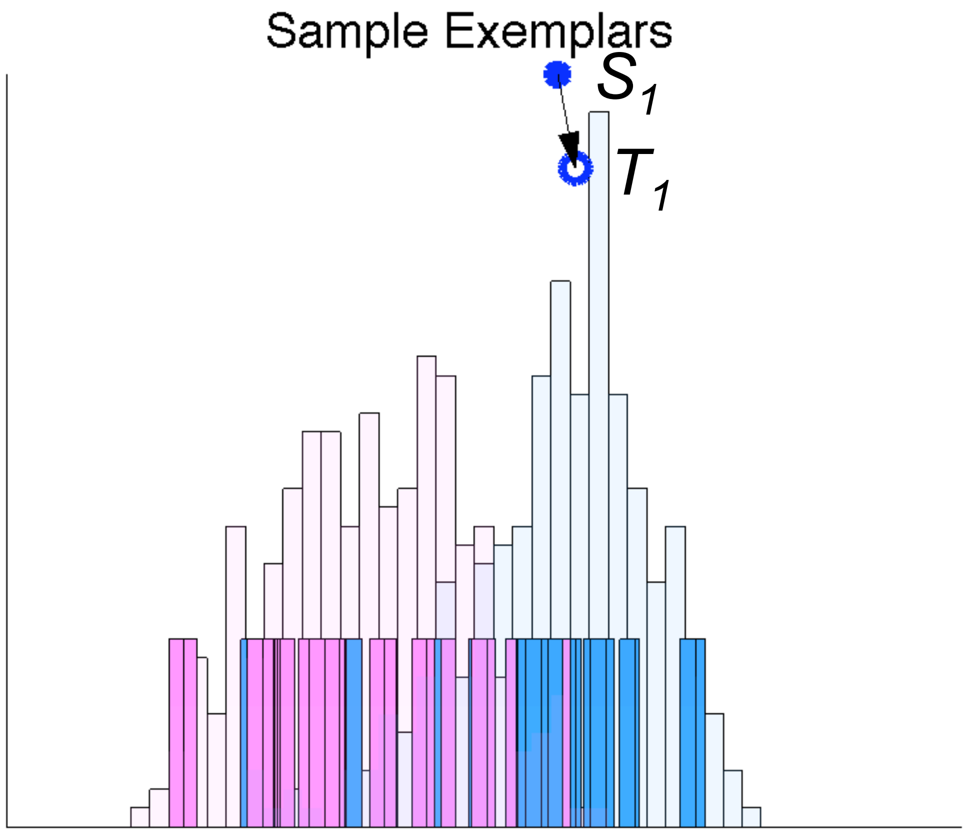
Perceiving First Stimulus



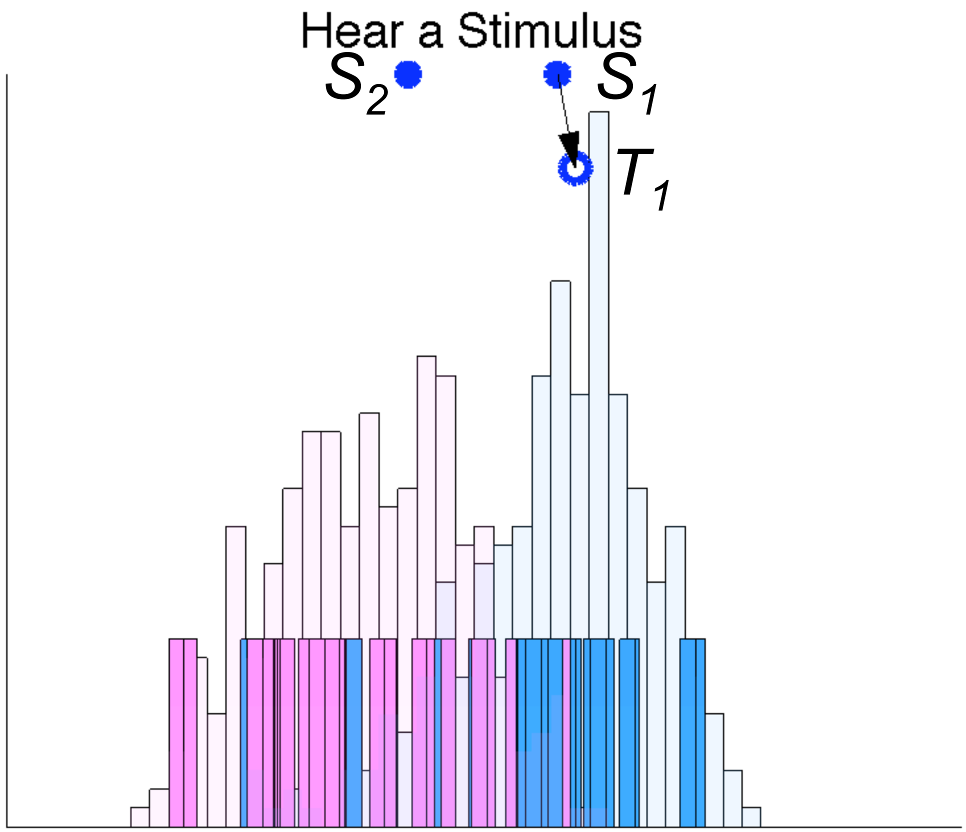
Perceiving First Stimulus



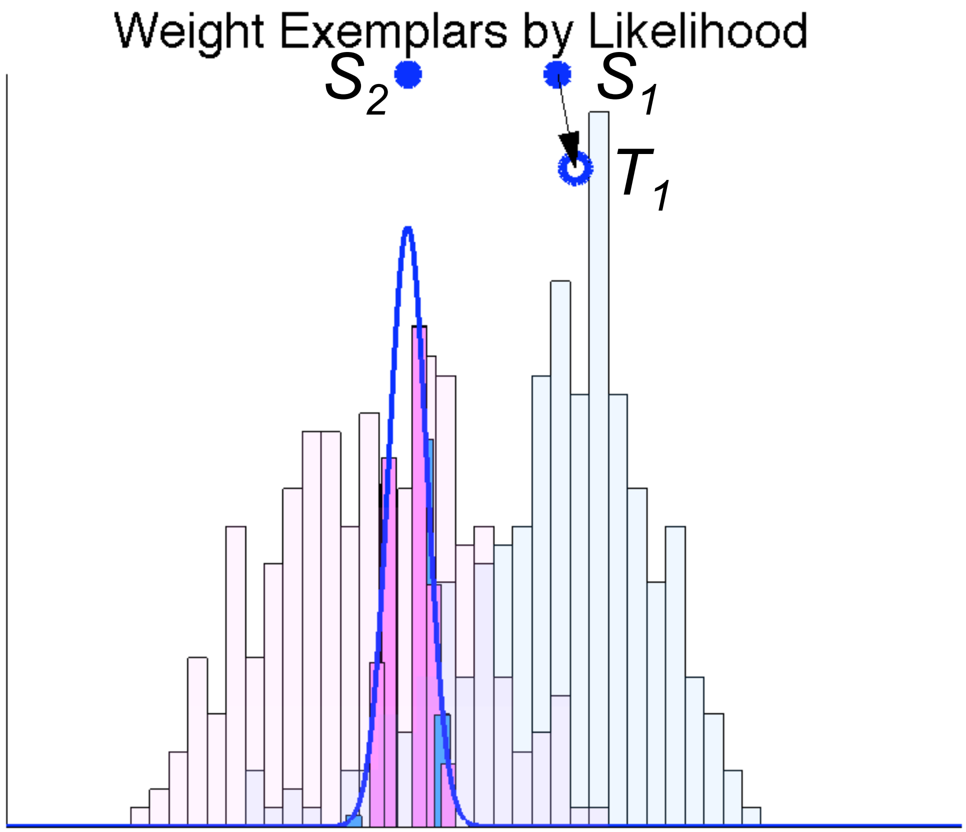
Perceiving Second Stimulus



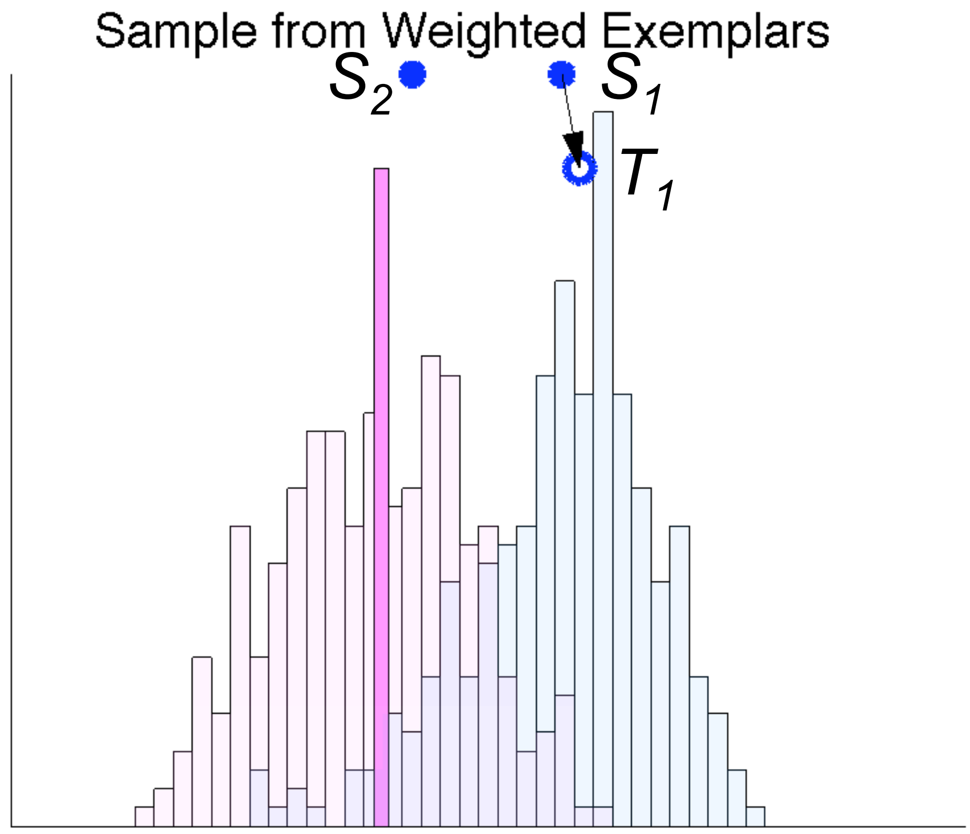
Perceiving Second Stimulus



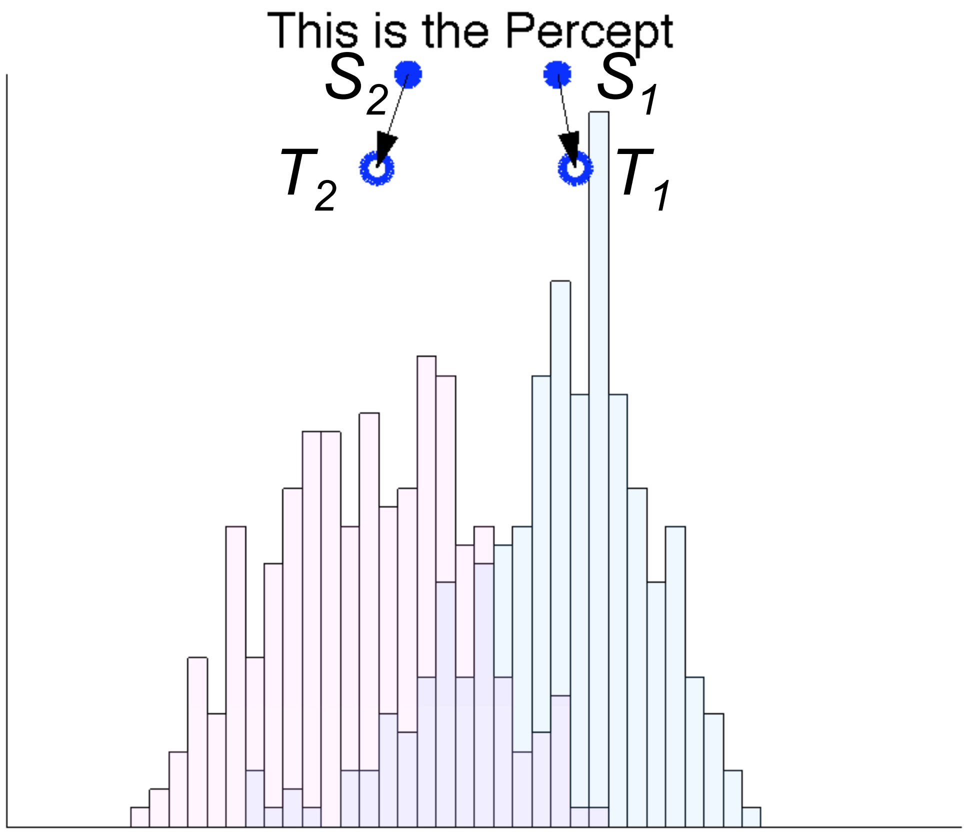
Perceiving Second Stimulus



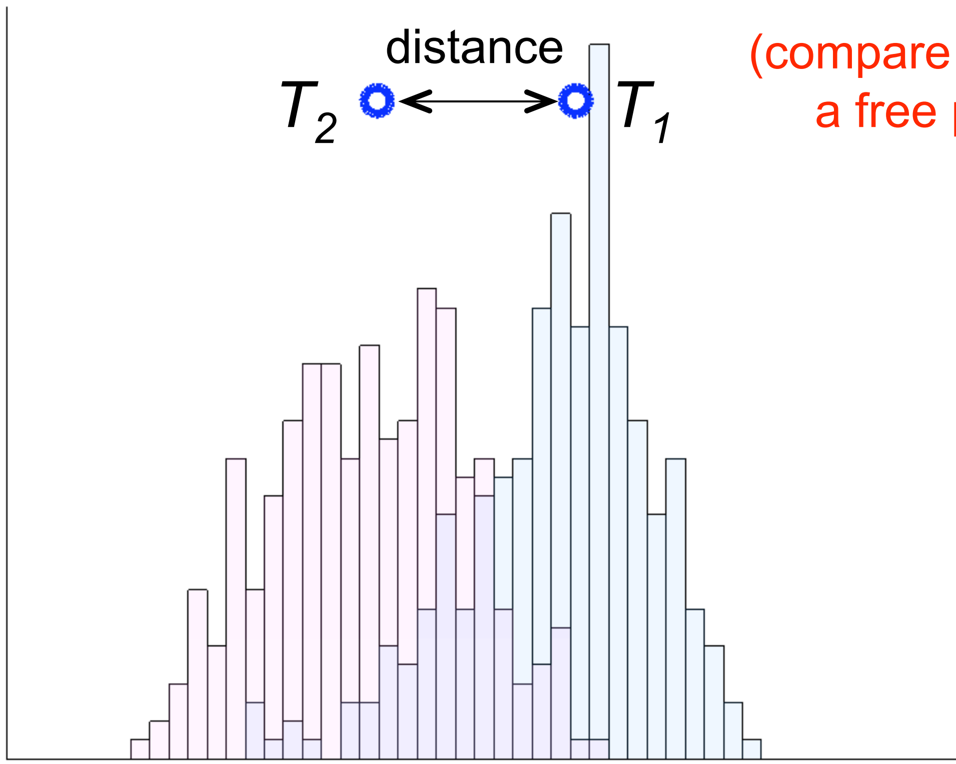
Perceiving Second Stimulus



Perceiving Second Stimulus

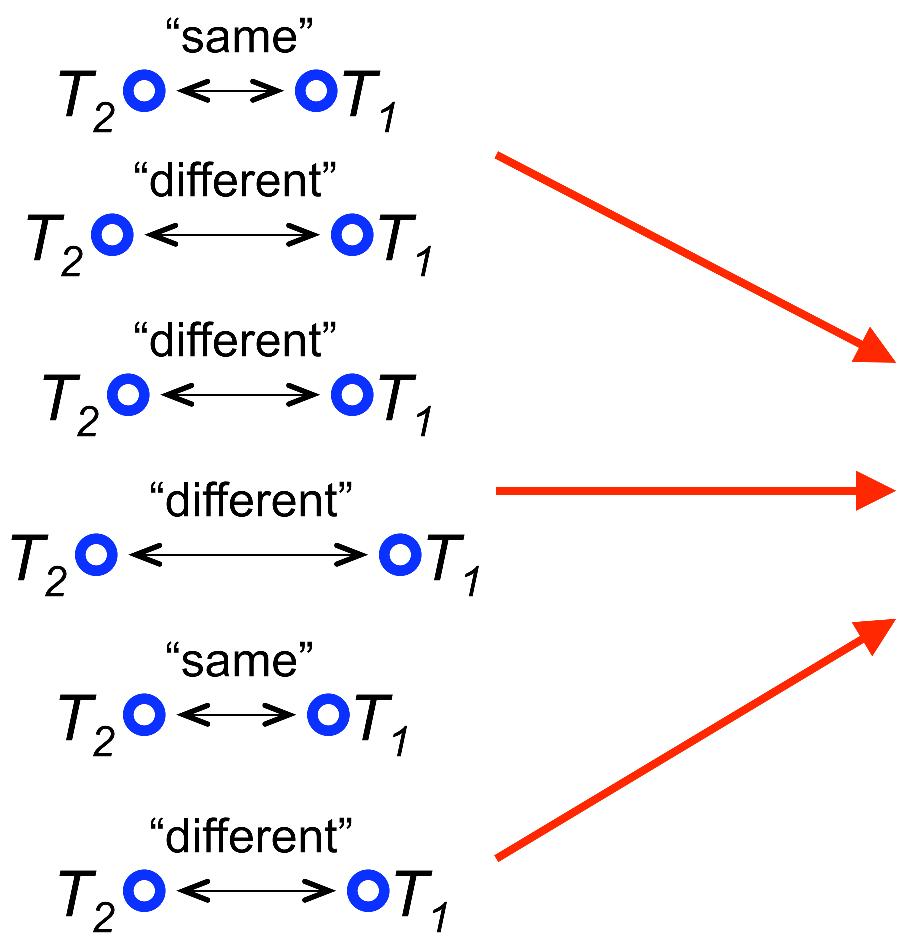


Same or Different?



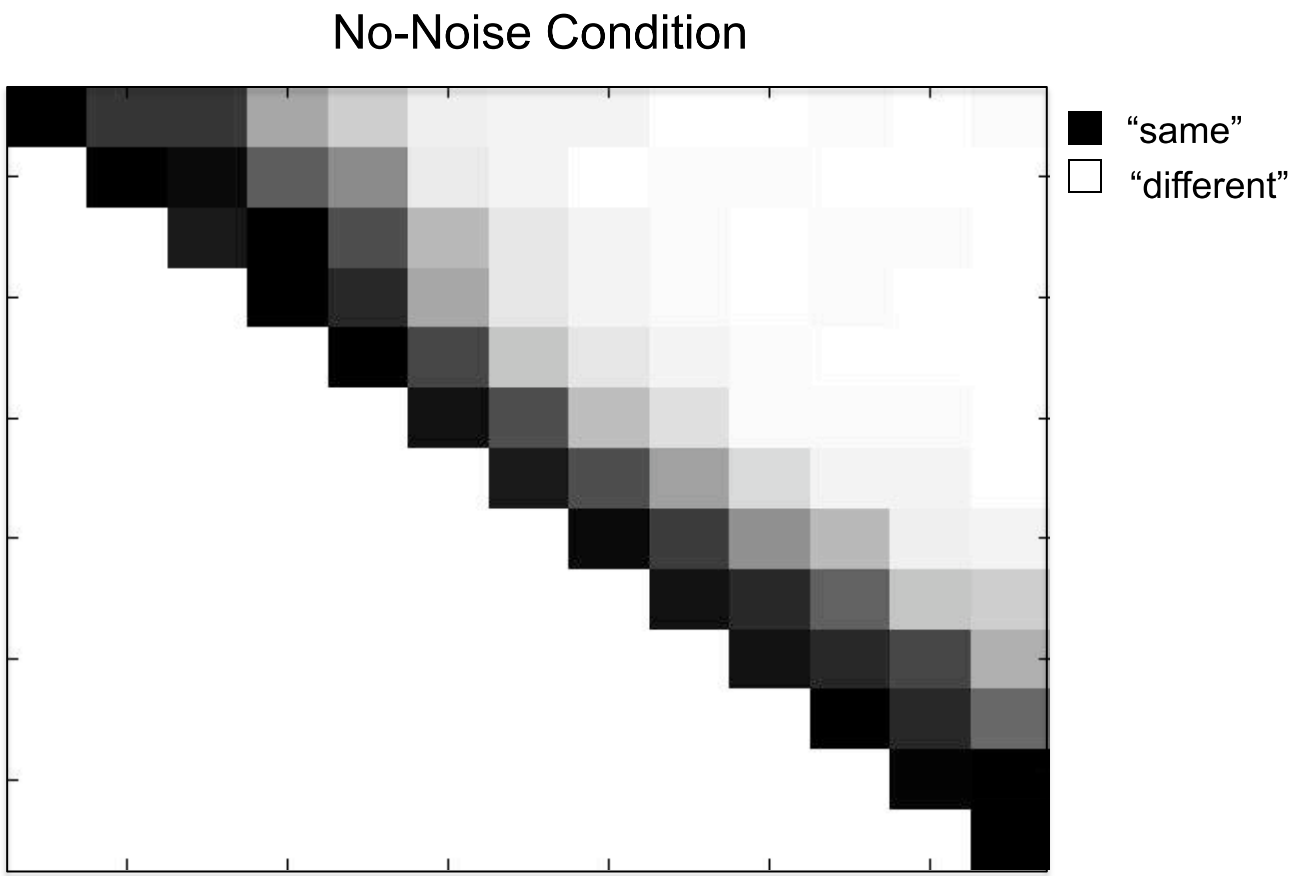
(compare to threshold ε,
a free parameter)

Probability of *Same* Response



Monte Carlo estimate of a binomial parameter for proportion of time listeners respond “same” to this trial

Compare with Human Data



Working from Corpus Data

1. Assume sounds in corpus are a sample from listeners' prior distribution
2. Sample from listeners' posterior distribution for each experimental stimulus
3. Use those samples to estimate listeners' probability of responding "same" on each trial
4. Compare model predictions to discrimination data

Outline

- Behavioral data in speech perception
- Cognitive model of speech perception
- Adapting the model to speech corpora
- A case study: Speaker normalization

Caitlin Richter

Aren Jansen

Representing Speech

Linguists

- Formant frequencies
- Formant transitions
- Voice onset time
- Pitch
- Duration

Engineers

- Mel frequency cepstral coefficients (MFCC)
- Perceptual linear prediction (PLP)
- Relative spectral encoding (RASTA)
- Posteriorgrams

Representing Speech

- Distributions of sounds may look different depending on how you represent the speech signal
- Which features predict human data best?

Speaker Normalization

- Human listeners generalize across talkers
  - Infants generalize across talkers at 6 months (Kuhl, 1979)
  - Adults normalize for a range of vocal tract lengths in recognizing vowels (Smith, Patterson, Turner, Kawahara, & Irino, 2005)

- Vocal tract length normalization improves performance in ASR systems (Wegmann, McAllaster, Orloff, & Peskin, 1996)

- Removing predictable variability improves a cognitive model of fricative identification (McMurray & Jongman, 2011)

Speaker Normalization

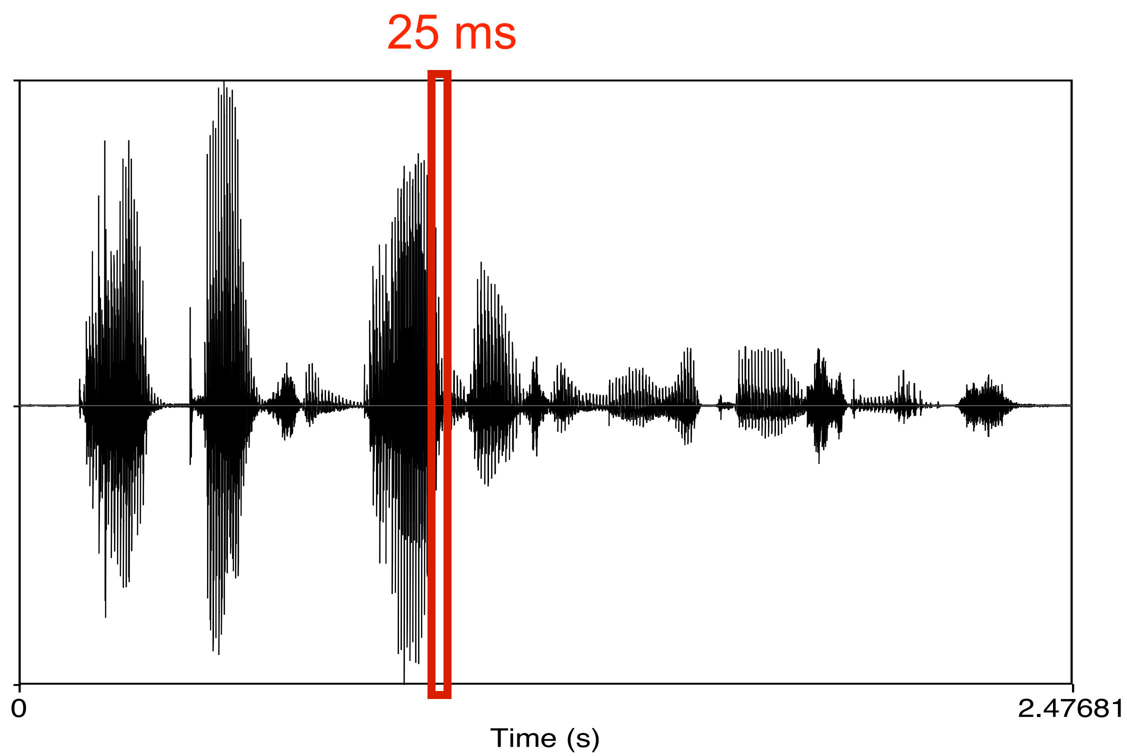
Is there a benefit of vocal tract length normalization in predicting human discrimination data?

Mel frequency Cepstral coefficients (MFCCs)

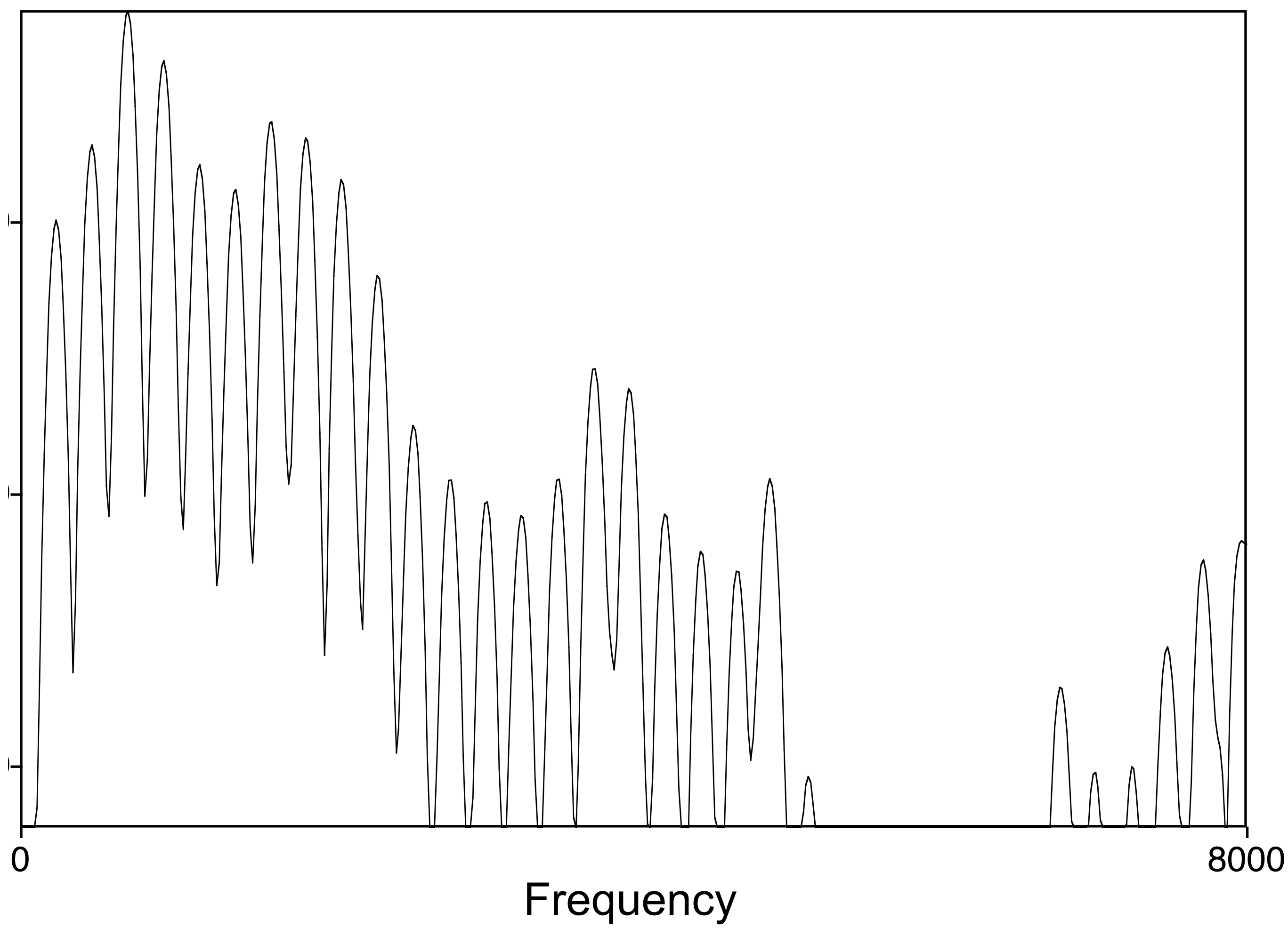
vs.

Mel frequency Cepstral coefficients (MFCCs)
with vocal tract length normalization (VTLN)

Waveform



Spectrum



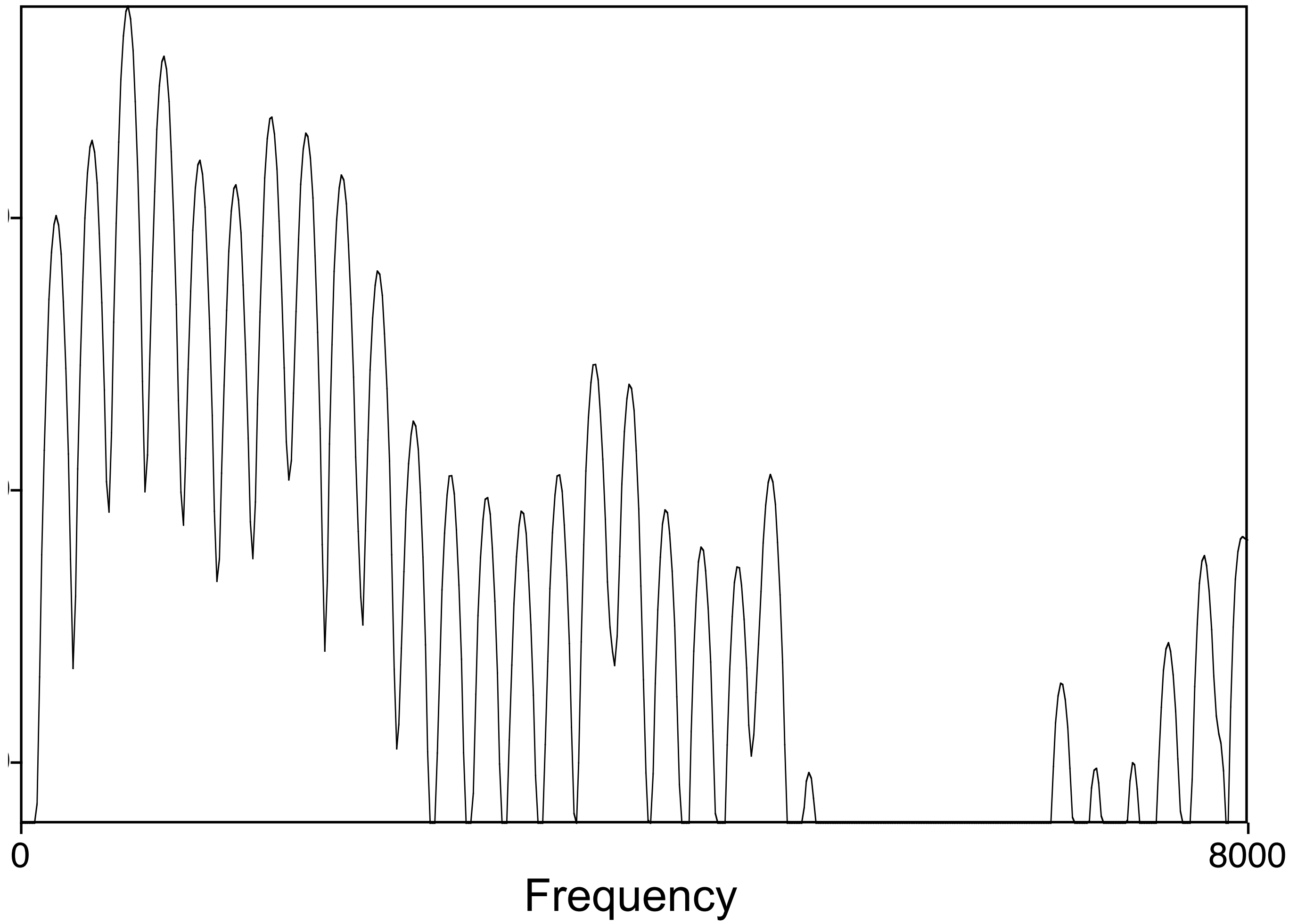
Mel Frequency Cepstral Coefficients

- Compute the log mel power spectrum for each frame

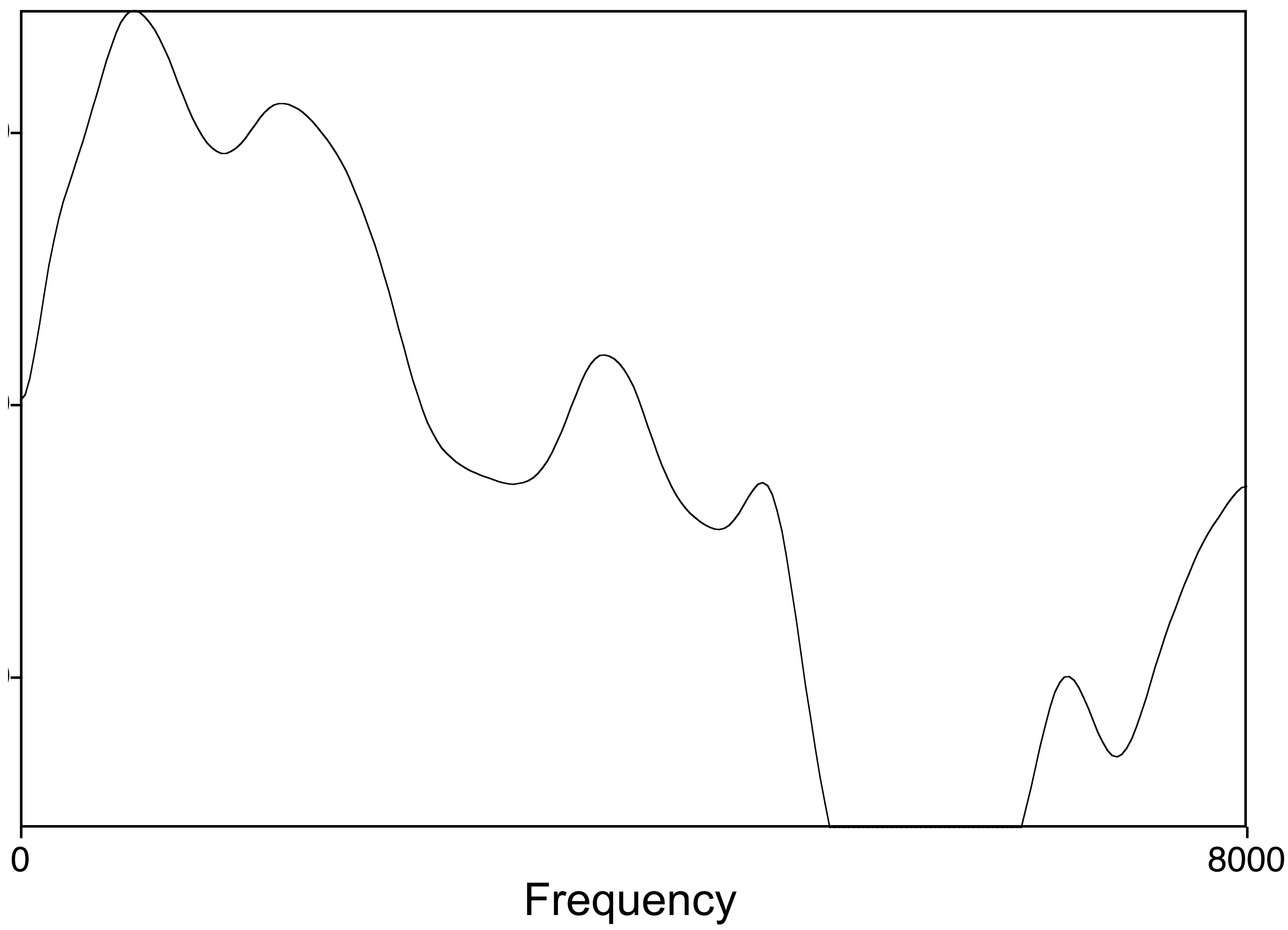
- Take first several low-frequency coefficients of the discrete cosine transform
  - Captures broad peaks in the spectrum
  - Ignores narrower (high-frequency) peaks

(Davis & Mermelstein, 1980)

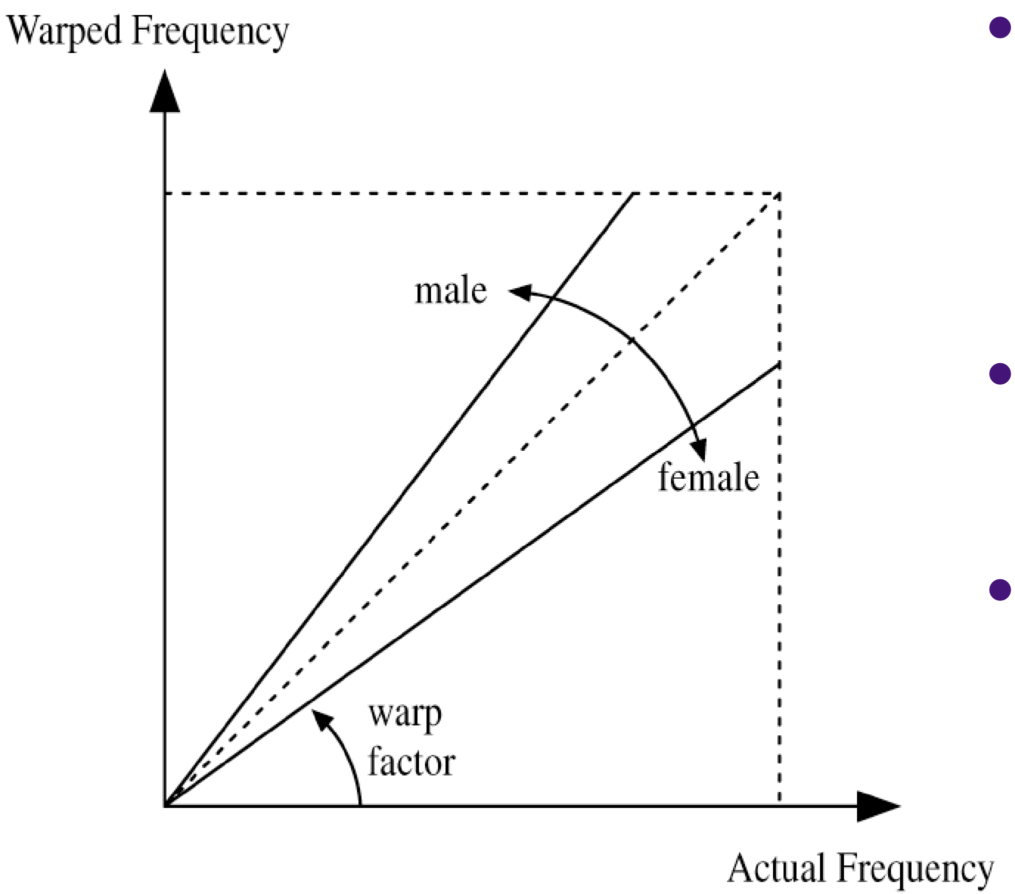
Spectrum



Low-Frequency Components



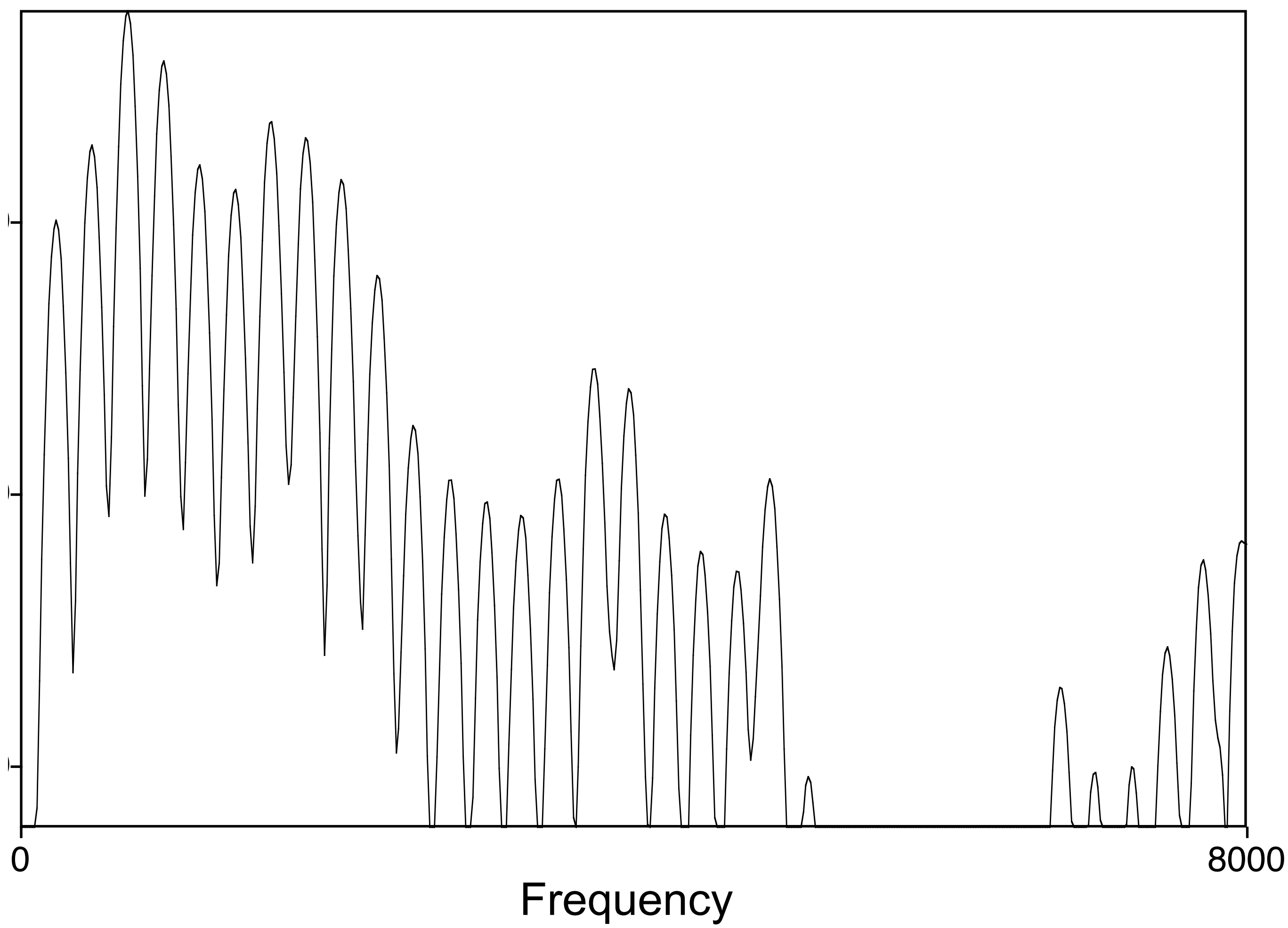
Vocal Tract Length Normalization



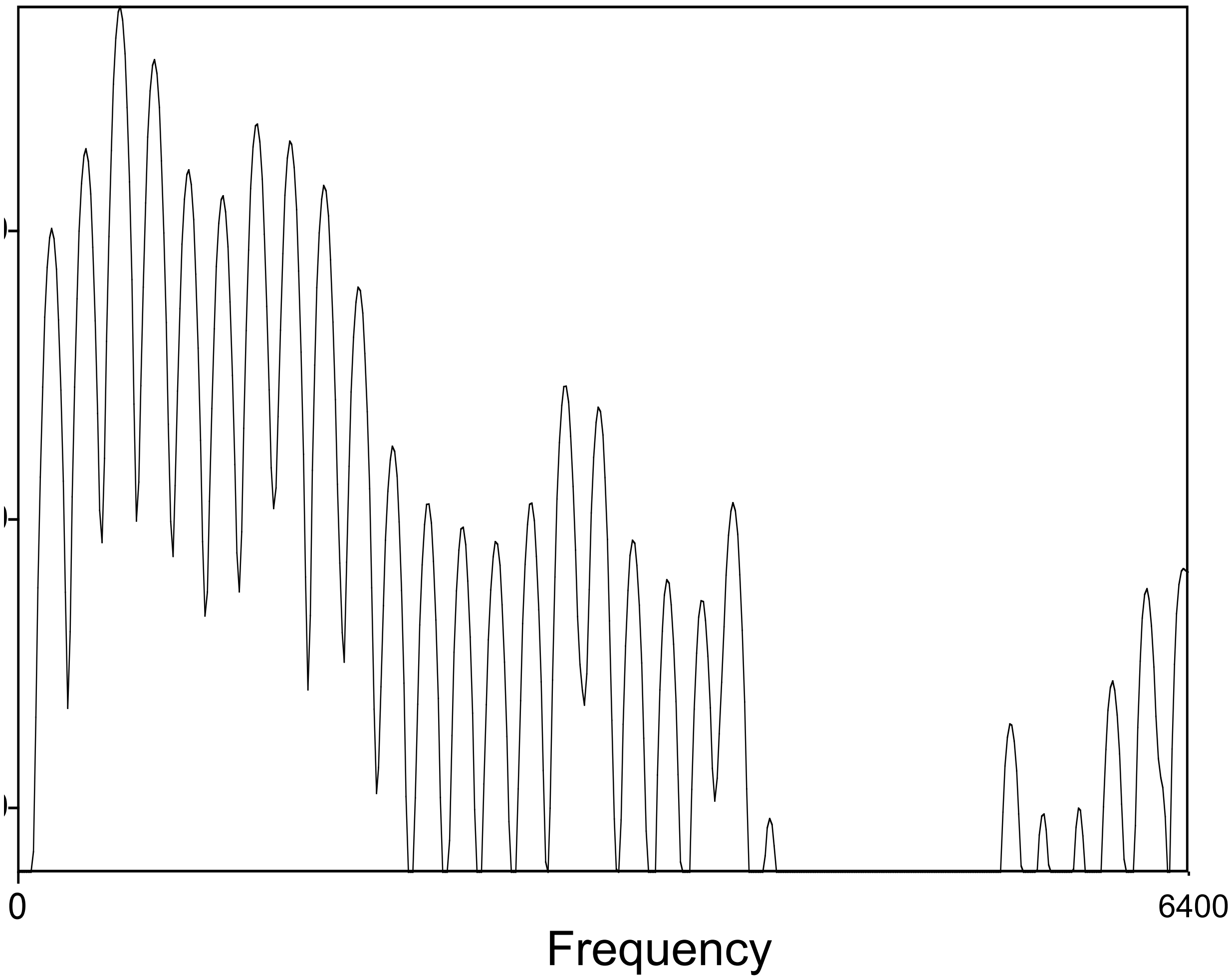
- Frequencies in filter bank scaled linearly by a ***warp factor*** before features are computed
- Warp factors ranged from 0.8 to 1.2, in increments of 0.05
- Warp factor for each speaker was chosen to maximize the likelihood of [i] frames in a Gaussian mixture model

(Wegmann, McAllaster, Orloff, & Peskin, 1996)

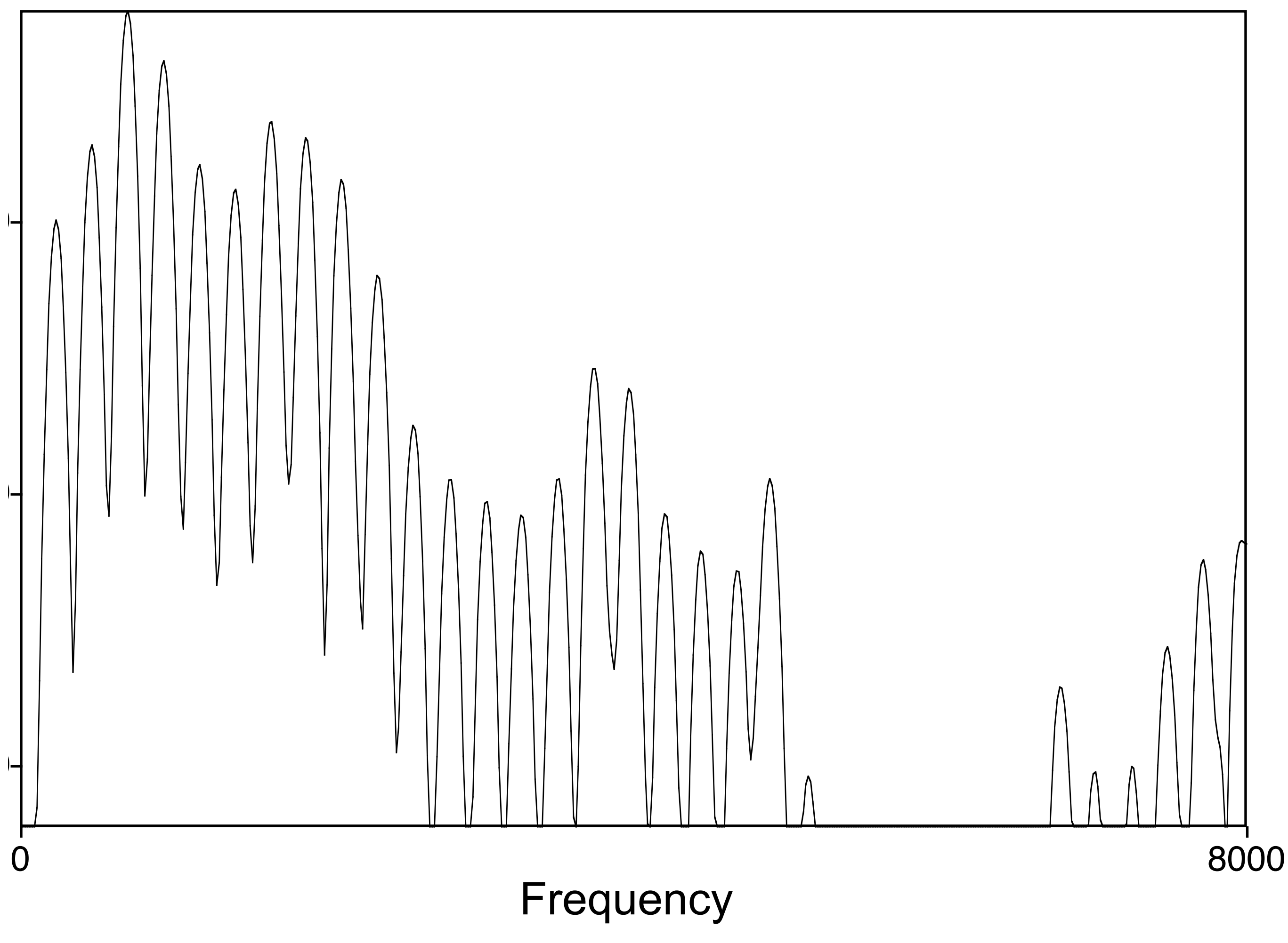
Spectrum



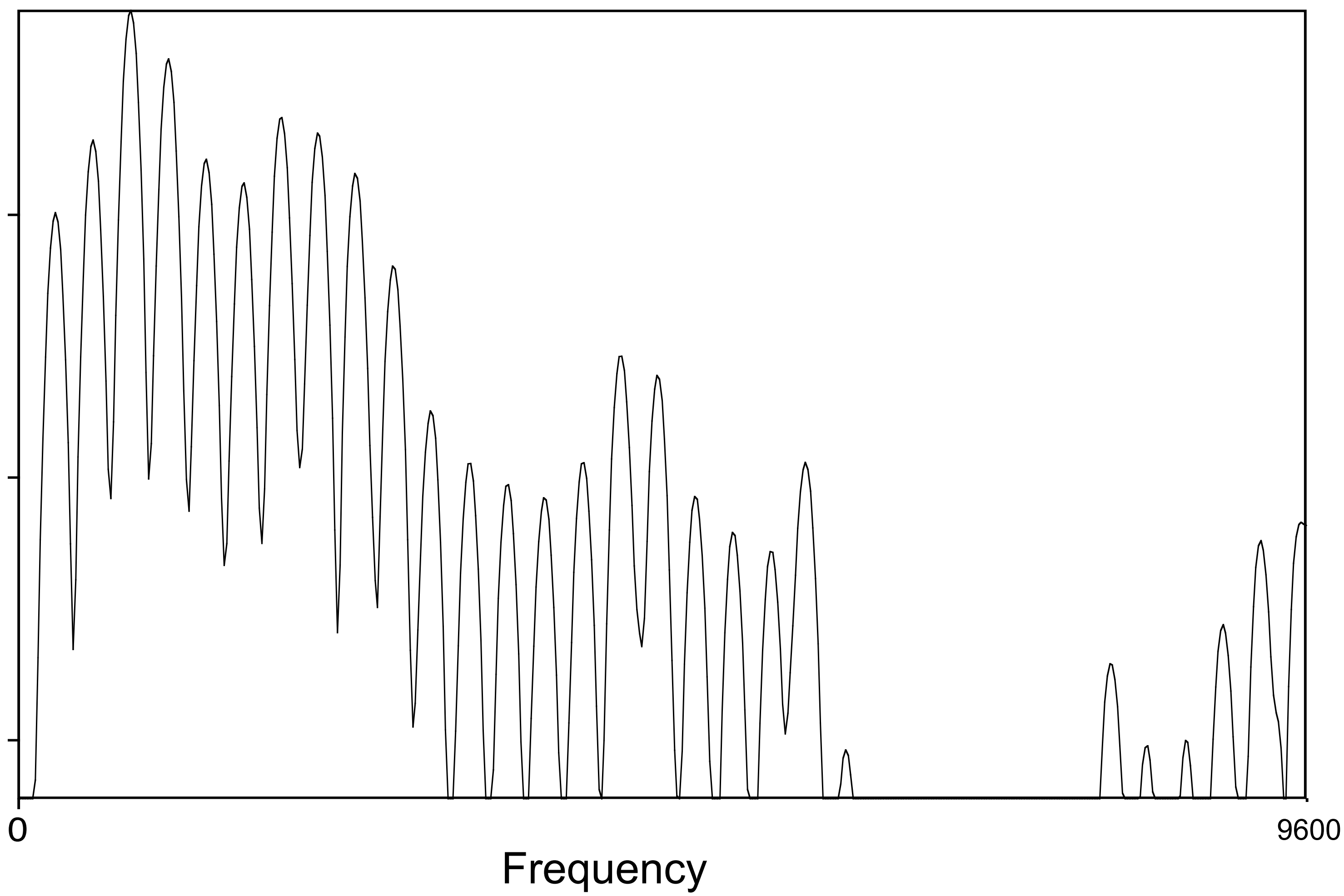
Spectrum



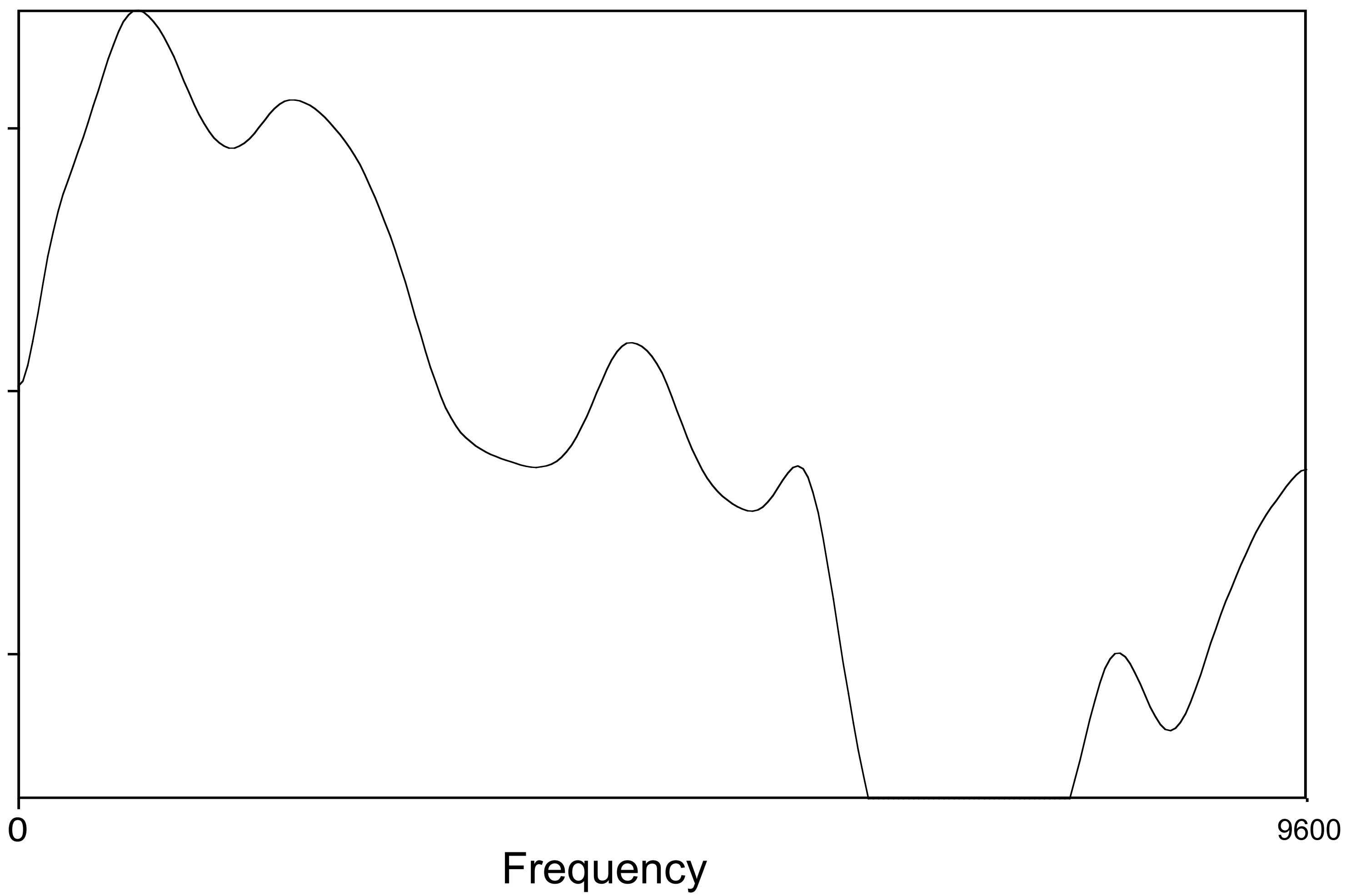
Spectrum



Spectrum



Low-Frequency Components



Feature Representation

	Frame 1	Frame 2	Frame 3	Frame 4	Frame 5
0	-1.9797	-0.5123	-1.4678	-1.9308	-3.2854
1	0.9850	0.9496	0.7751	1.0833	0.3002
2	0.1707	-0.7913	-0.0947	-0.8188	-0.1646
3	0.4207	1.0198	1.3087	1.1888	0.9085
4	0.6101	-0.1021	-0.1033	-0.3476	1.8199
5	0.2639	0.0865	0.2027	-0.4556	-0.2619
6	0.2932	-0.4730	-0.2413	1.1715	2.2628
7	0.6992	0.3412	0.1339	-0.5760	0.1275
8	-0.1473	-0.1811	0.0271	2.0721	0.8592
9	-0.1296	-0.8160	-0.7020	-0.8623	-0.5312
10	0.3189	0.2433	0.4987	-0.1196	0.0098
11	0.5889	0.9983	0.7926	1.1798	0.7214
12	-1.4080	-1.7503	-1.6107	-1.5995	-0.9872

Feature Representation

frames in 10-ms steps

	Frame 1	Frame 2	Frame 3	Frame 4	Frame 5
0	-1.9797	-0.5123	-1.4678	-1.9308	-3.2854
1	0.9850	0.9496	0.7751	1.0833	0.3002
2	0.1707	-0.7913	-0.0947	-0.8188	-0.1646
3	0.4207	1.0198	1.3087	1.1888	0.9085
4	0.6101	-0.1021	-0.1033	-0.3476	1.8199
5	0.2639	0.0865	0.2027	-0.4556	-0.2619
6	0.2932	-0.4730	-0.2413	1.1715	2.2628
7	0.6992	0.3412	0.1339	-0.5760	0.1275
8	-0.1473	-0.1811	0.0271	2.0721	0.8592
9	-0.1296	-0.8160	-0.7020	-0.8623	-0.5312
10	0.3189	0.2433	0.4987	-0.1196	0.0098
11	0.5889	0.9983	0.7926	1.1798	0.7214
12	-1.4080	-1.7503	-1.6107	-1.5995	-0.9872

Feature Representation

frames in 10-ms steps

dimensions

	Frame 1	Frame 2	Frame 3	Frame 4	Frame 5
0	-1.9797	-0.5123	-1.4678	-1.9308	-3.2854
1	0.9850	0.9496	0.7751	1.0833	0.3002
2	0.1707	-0.7913	-0.0947	-0.8188	-0.1646
3	0.4207	1.0198	1.3087	1.1888	0.9085
4	0.6101	-0.1021	-0.1033	-0.3476	1.8199
5	0.2639	0.0865	0.2027	-0.4556	-0.2619
6	0.2932	-0.4730	-0.2413	1.1715	2.2628
7	0.6992	0.3412	0.1339	-0.5760	0.1275
8	-0.1473	-0.1811	0.0271	2.0721	0.8592
9	-0.1296	-0.8160	-0.7020	-0.8623	-0.5312
10	0.3189	0.2433	0.4987	-0.1196	0.0098
11	0.5889	0.9983	0.7926	1.1798	0.7214
12	-1.4080	-1.7503	-1.6107	-1.5995	-0.9872

Feature Representation

frames in 10-ms steps

dimensions

	Frame 1	Frame 2	Frame 3	Frame 4	Frame 5
[redacted]					
1	0.9850	0.9496	0.7751	1.0833	0.3002
2	0.1707	-0.7913	-0.0947	-0.8188	-0.1646
3	0.4207	1.0198	1.3087	1.1888	0.9085
4	0.6101	-0.1021	-0.1033	-0.3476	1.8199
5	0.2639	0.0865	0.2027	-0.4556	-0.2619
6	0.2932	-0.4730	-0.2413	1.1715	2.2628
7	0.6992	0.3412	0.1339	-0.5760	0.1275
8	-0.1473	-0.1811	0.0271	2.0721	0.8592
9	-0.1296	-0.8160	-0.7020	-0.8623	-0.5312
10	0.3189	0.2433	0.4987	-0.1196	0.0098
11	0.5889	0.9983	0.7926	1.1798	0.7214
12	-1.4080	-1.7503	-1.6107	-1.5995	-0.9872

Feature Representation

frames in 10-ms steps

dimensions

	Frame 1	Frame 2	Frame 3	Frame 4	Frame 5
[redacted]					
1	0.9850	0.9496	0.7751	1.0833	0.3002
2	0.1707	-0.7913	-0.0947	-0.8188	-0.1646
3	0.4207	1.0198	1.3087	1.1888	0.9085
4	0.6101	-0.1021	-0.1033	-0.3476	1.8199
5	0.2639	0.0865	0.2027	-0.4556	-0.2619
6	0.2932	-0.4730	-0.2413	1.1715	2.2628
7	0.6992	0.3412	0.1339	-0.5760	0.1275
8	-0.1473	-0.1811	0.0271	2.0721	0.8592
9	-0.1296	-0.8160	-0.7020	-0.8623	-0.5312
10	0.3189	0.2433	0.4987	-0.1196	0.0098
11	0.5889	0.9983	0.7926	1.1798	0.7214
12	-1.4080	-1.7503	-1.6107	-1.5995	-0.9872

Feature Representation

frames in 10-ms steps

dimensions

		Center Frame	
	[redacted]		[redacted]
1	[redacted]	0.7751	[redacted]
2	[redacted]	-0.0947	[redacted]
3	[redacted]	1.3087	[redacted]
4	[redacted]	-0.1033	[redacted]
5	[redacted]	0.2027	[redacted]
6	[redacted]	-0.2413	[redacted]
7	[redacted]	0.1339	[redacted]
8	[redacted]	0.0271	[redacted]
9	[redacted]	-0.7020	[redacted]
10	[redacted]	0.4987	[redacted]
11	[redacted]	0.7926	[redacted]
12	[redacted]	-1.6107	[redacted]

Simulations

- Compute mel frequency cepstral coefficients (MFCCs), with and without vocal tract length normalization (VTLN), for the midpoint of each vowel in the corpus and each stimulus
- Simulate listeners using each type of feature
- Compare model predictions with human data

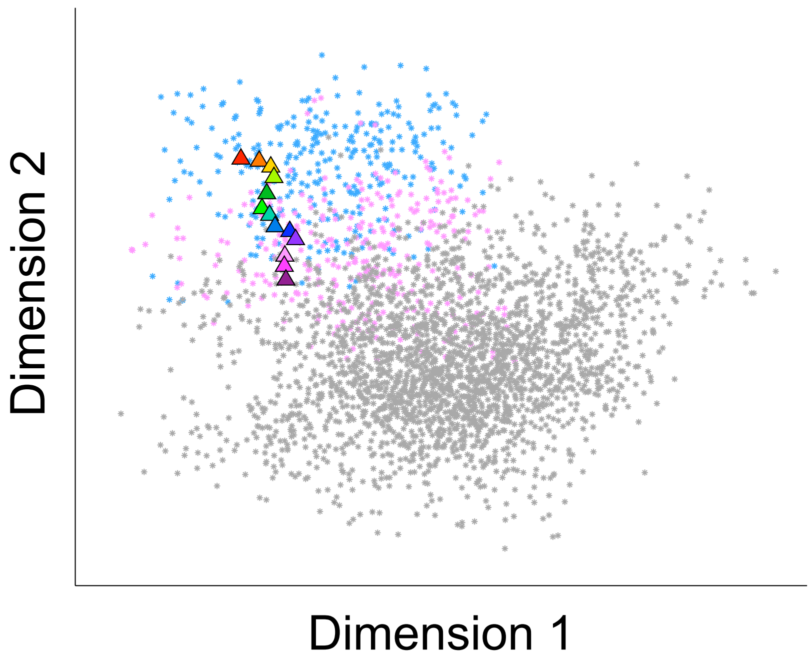
Nationwide Speech Project

- Recordings of 5 male, 5 female speakers from each of 6 dialect regions of the United States
- Each speaker produced 5 repetitions of 10 vowels in /hVd/ contexts: heed, hid, hayed, head, had, hod, hud, hoed, hood, who'd

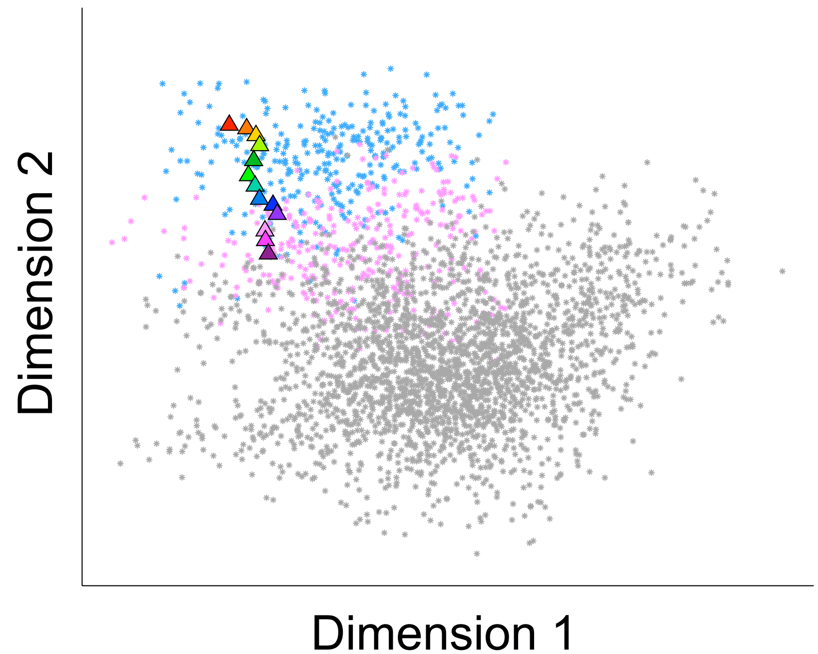
(Clopper & Pisoni, 2006)

Distributions of Exemplars

Raw MFCCs



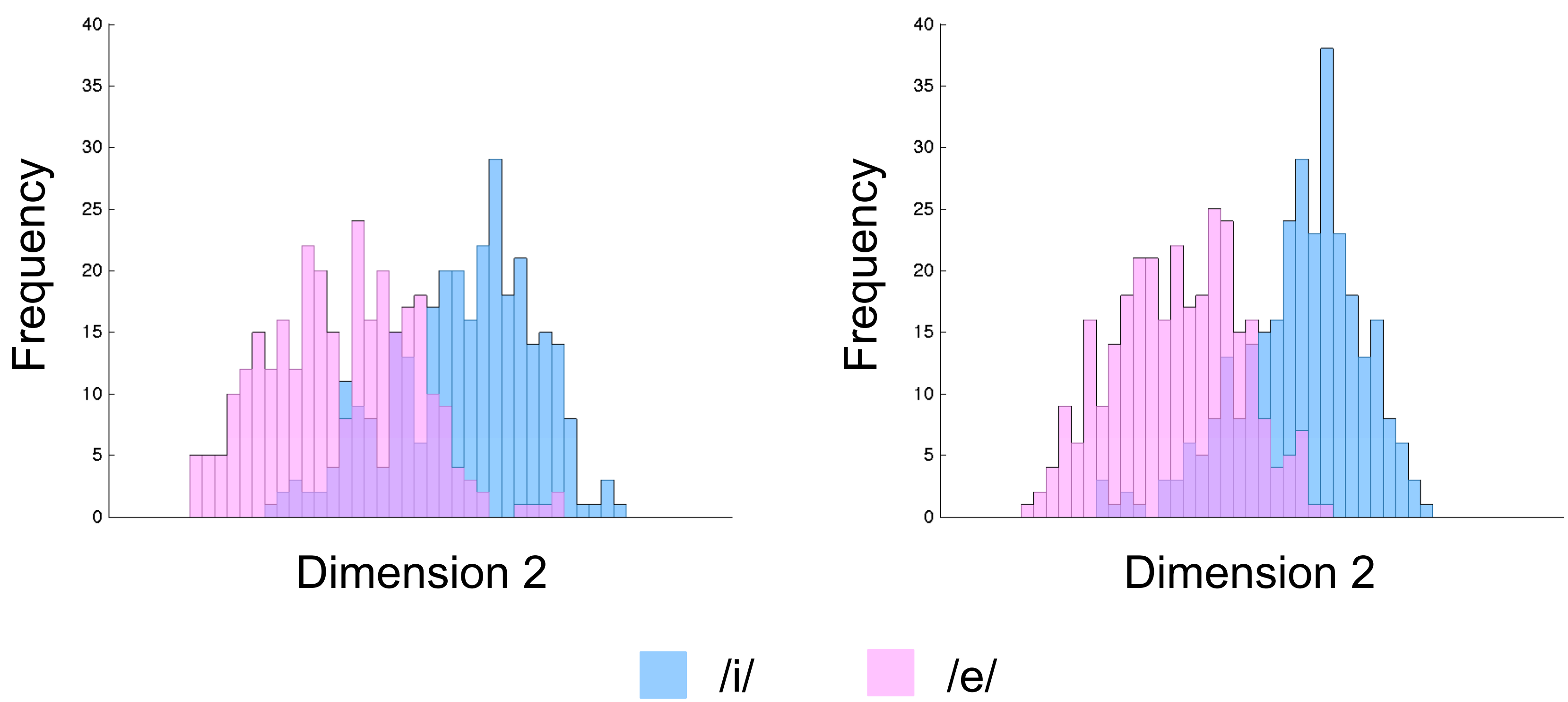
MFCCs with VTLN



/i/ /e/ all other vowels

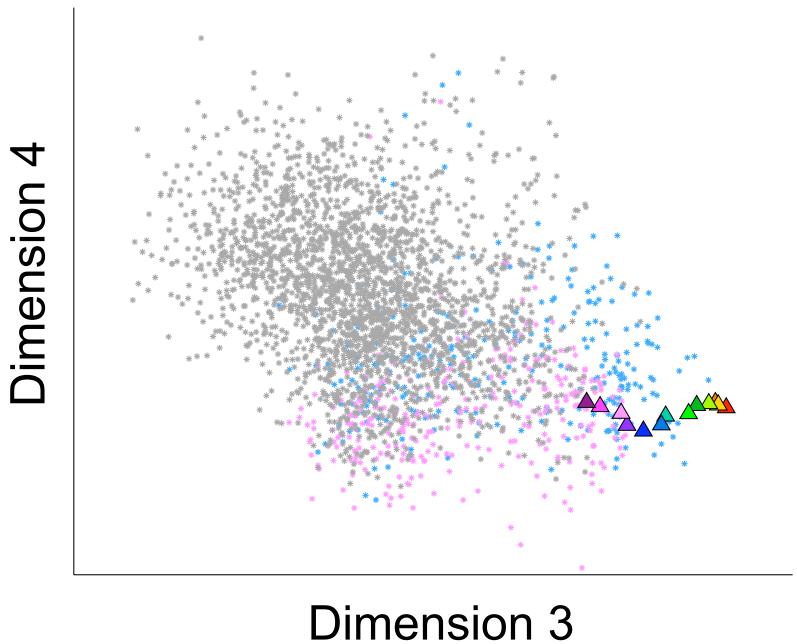
Distributions of Exemplars

Raw MFCCs **MFCCs with VTLN**

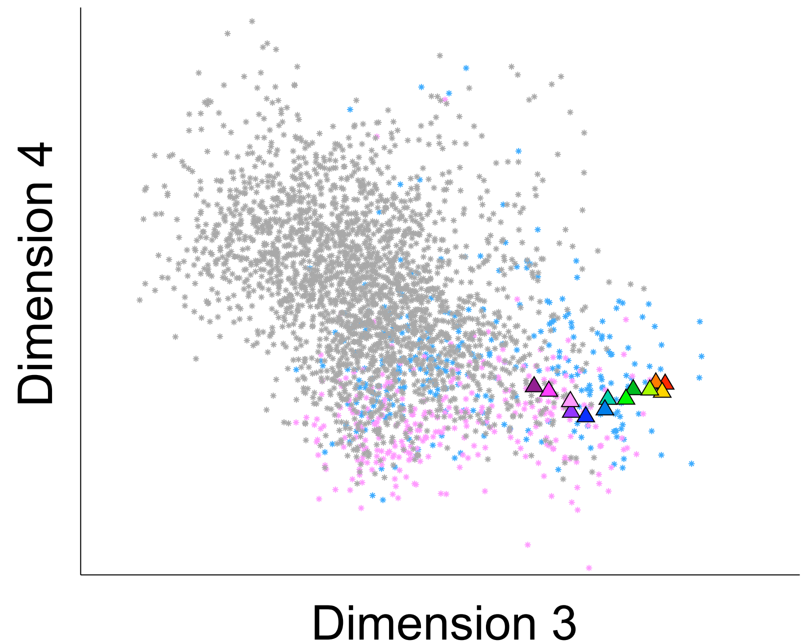


Scatterplots with Stimuli

Raw MFCCs

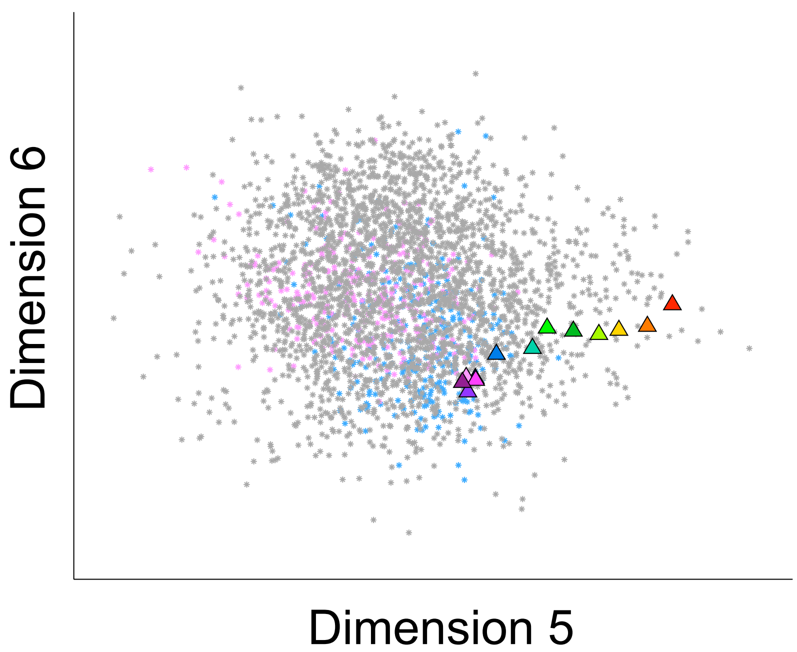


MFCCs with VTLN

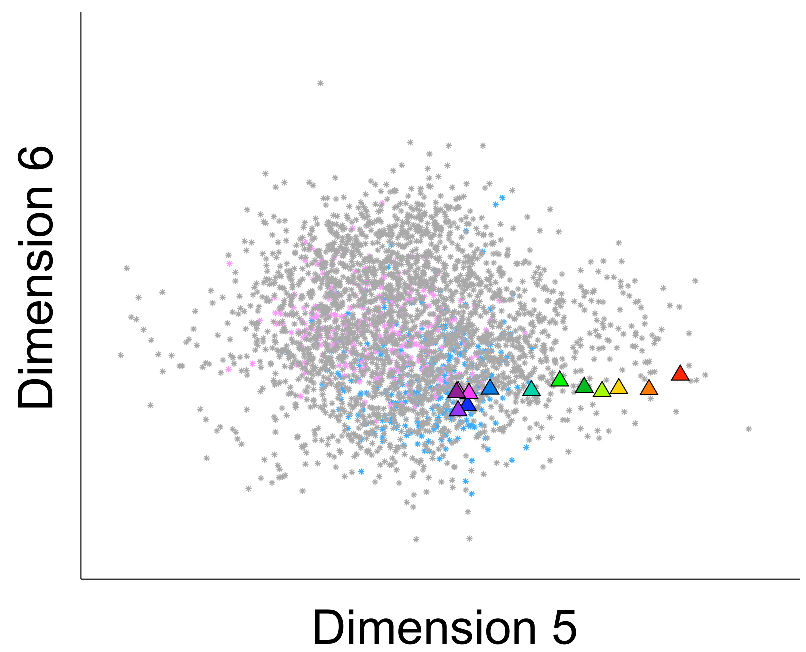


Scatterplots with Stimuli

Raw MFCCs



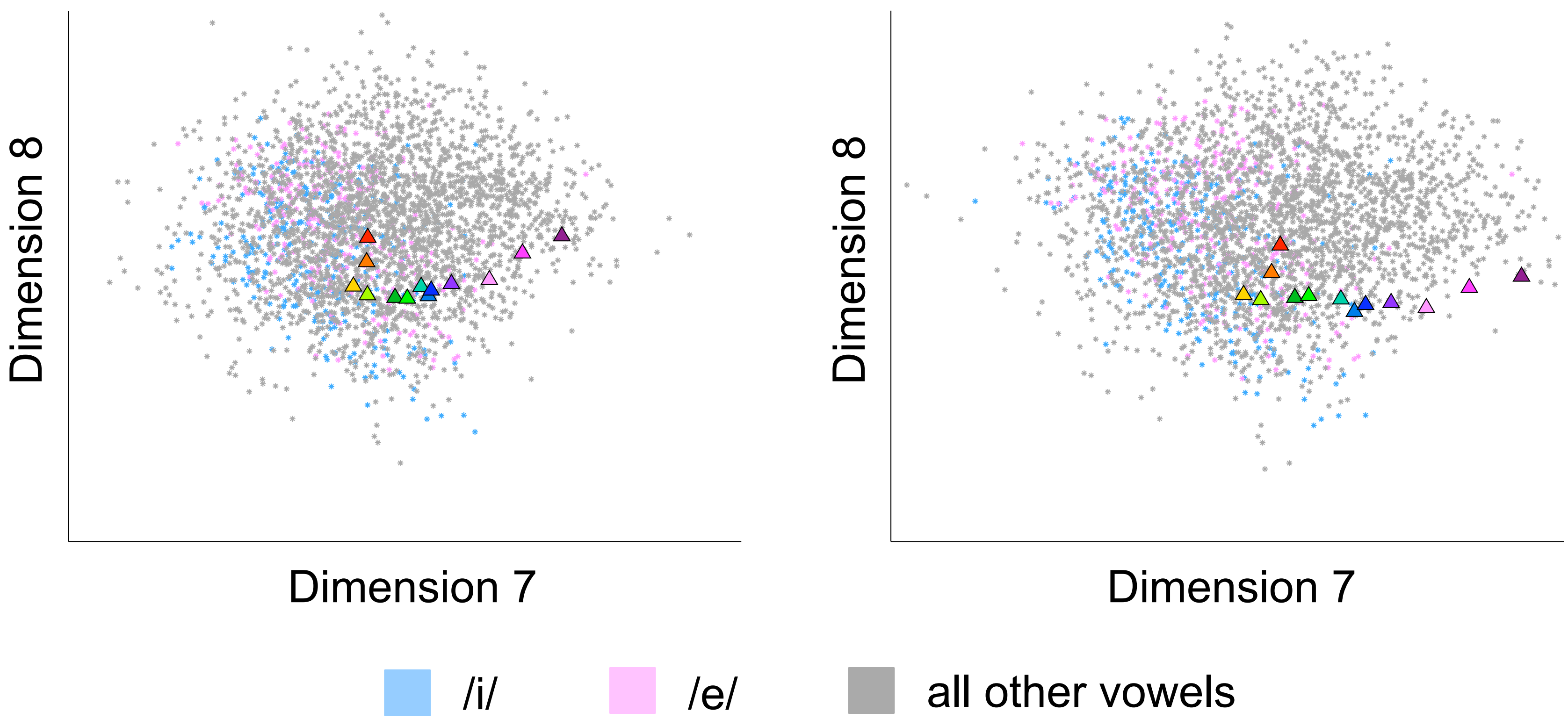
MFCCs with VTLN



/i/ /e/ all other vowels

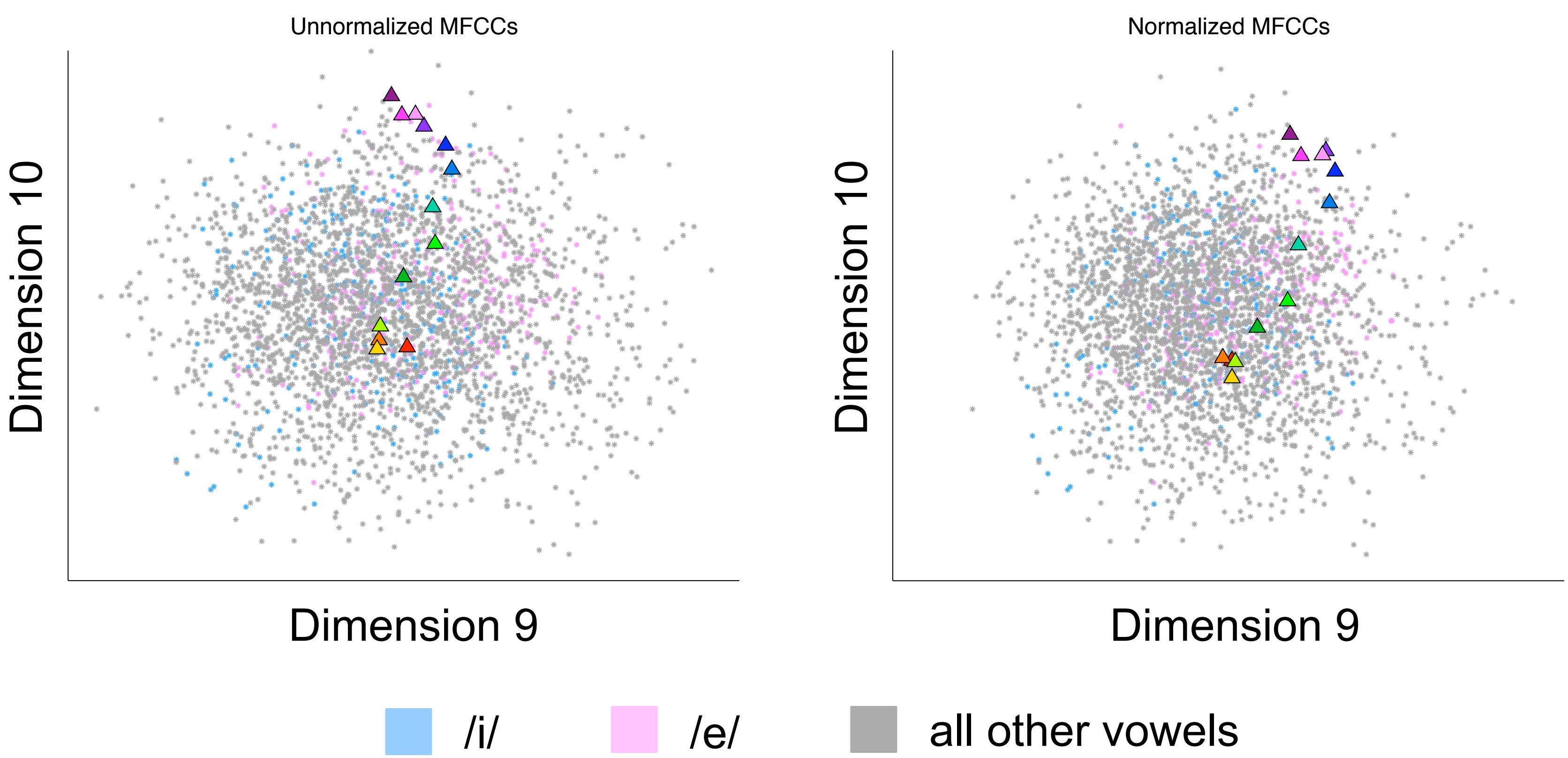
Scatterplots with Stimuli

Raw MFCCs | **MFCCs with VTLN**



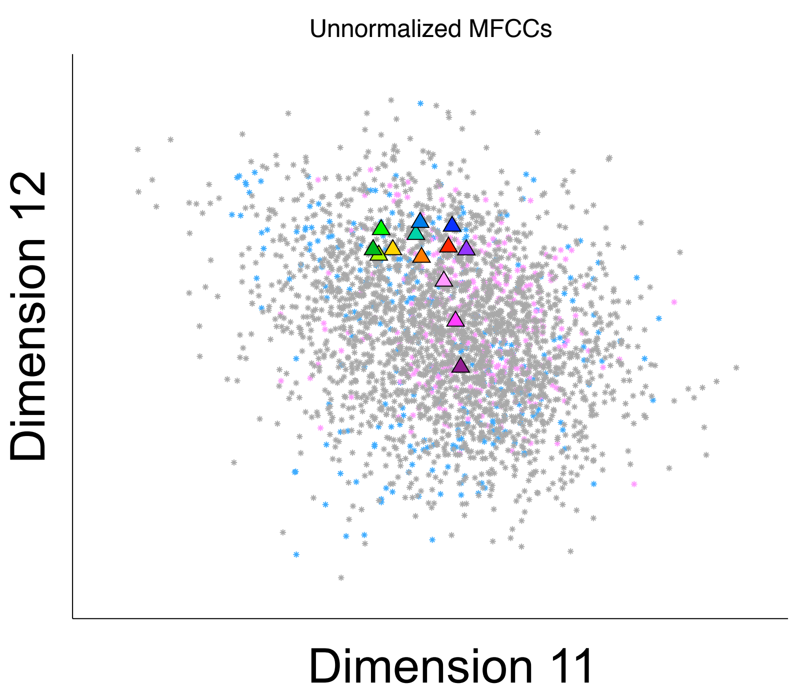
Scatterplots with Stimuli

Raw MFCCs | **MFCCs with VTLN**

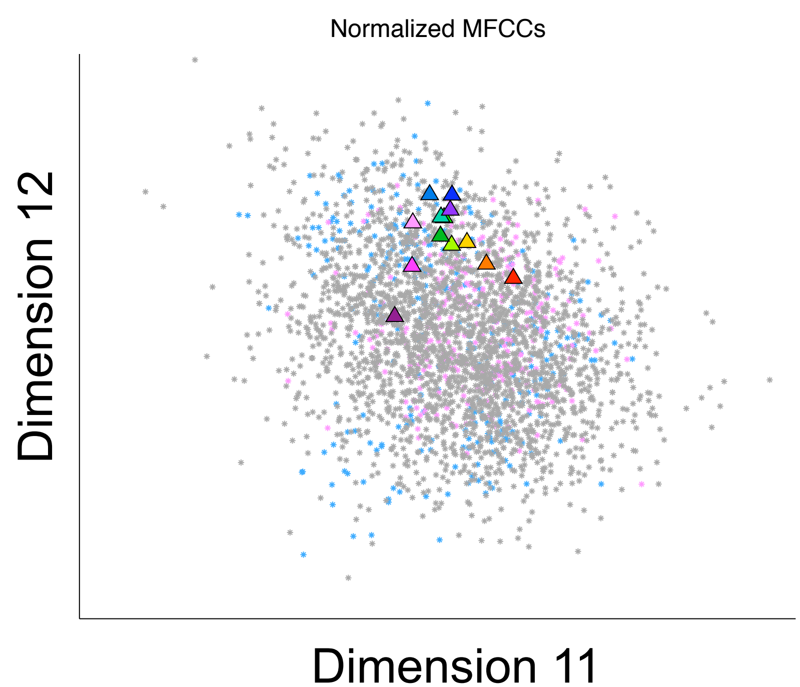


Scatterplots with Stimuli

Raw MFCCs



MFCCs with VTLN



/i/ /e/ all other vowels

Input to the Model

frames in 10-ms steps

dimensions

	[redacted]	**Center Frame**	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
1	[redacted]	0.7751	[redacted]
2	[redacted]	-0.0947	[redacted]
3	[redacted]	1.3087	[redacted]
4	[redacted]	-0.1033	[redacted]
5	[redacted]	0.2027	[redacted]
6	[redacted]	-0.2413	[redacted]
7	[redacted]	0.1339	[redacted]
8	[redacted]	0.0271	[redacted]
9	[redacted]	-0.7020	[redacted]
10	[redacted]	0.4987	[redacted]
11	[redacted]	0.7926	[redacted]
12	[redacted]	-1.6107	[redacted]

(Davis & Mermelstein, 1980)

Input to the Model

frames in 10-ms steps

dimensions

	Center Frame
1	0.7751
2	-0.0947
3	1.3087
4	-0.1033

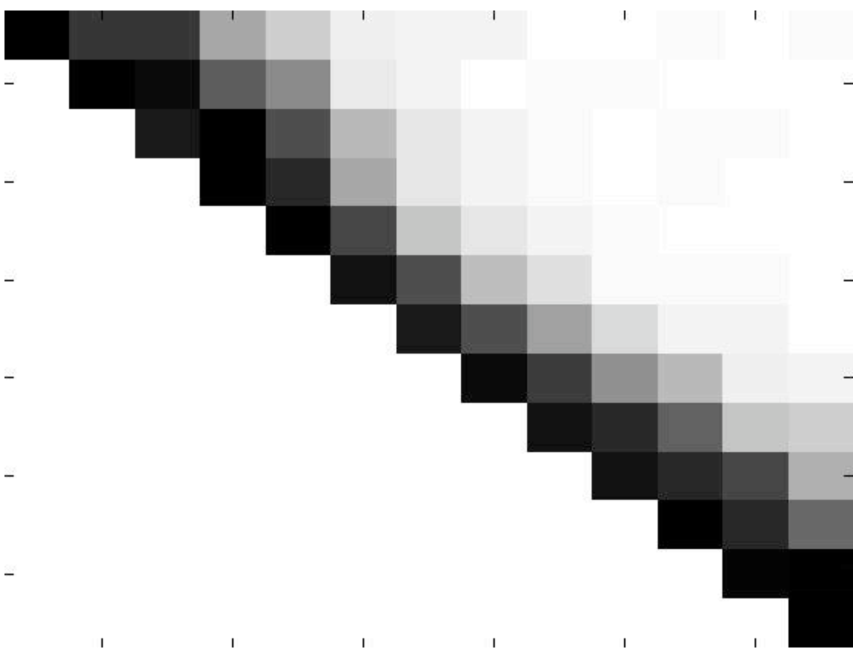
(Davis & Mermelstein, 1980)

Fitting the Model

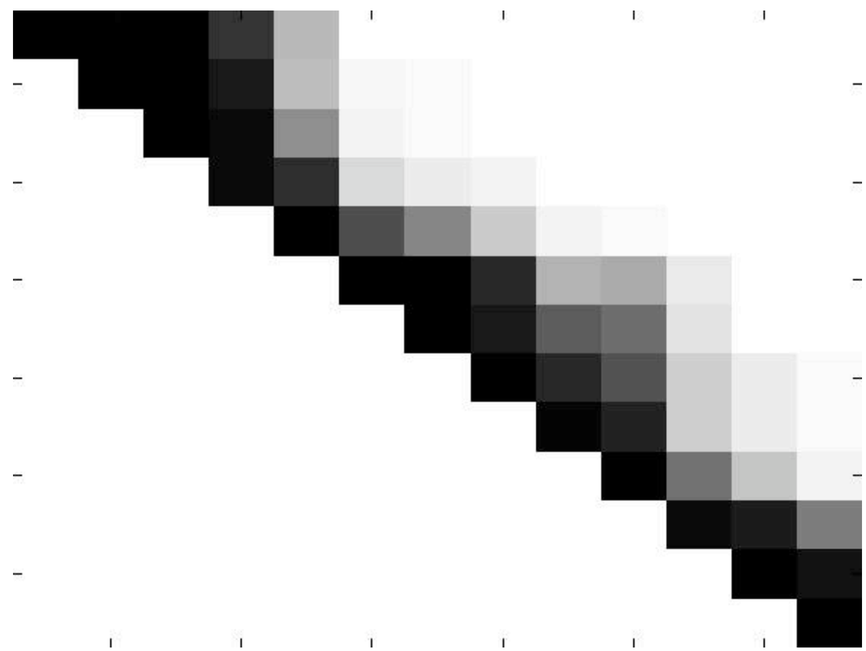
- Need to fit parameters for our simulation
  - Noise covariance matrix (constrained to be diagonal)
  - Response threshold
- MCMC to find parameter values with high likelihoods
- Half of exemplars in the corpus used for parameter fitting
- Model likelihoods computed on untrained exemplars using Monte Carlo simulation

Results: Raw MFCCs

Humans

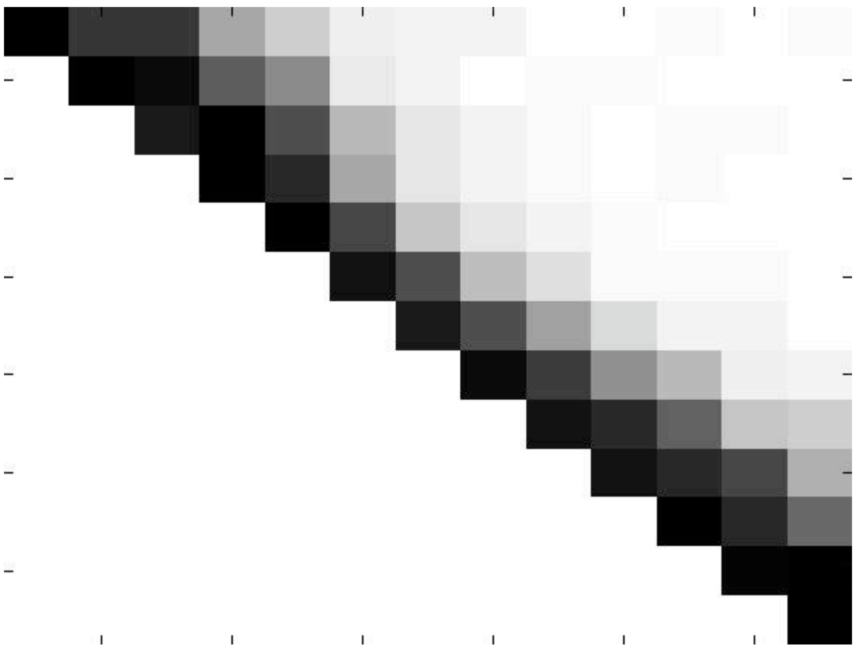


Model

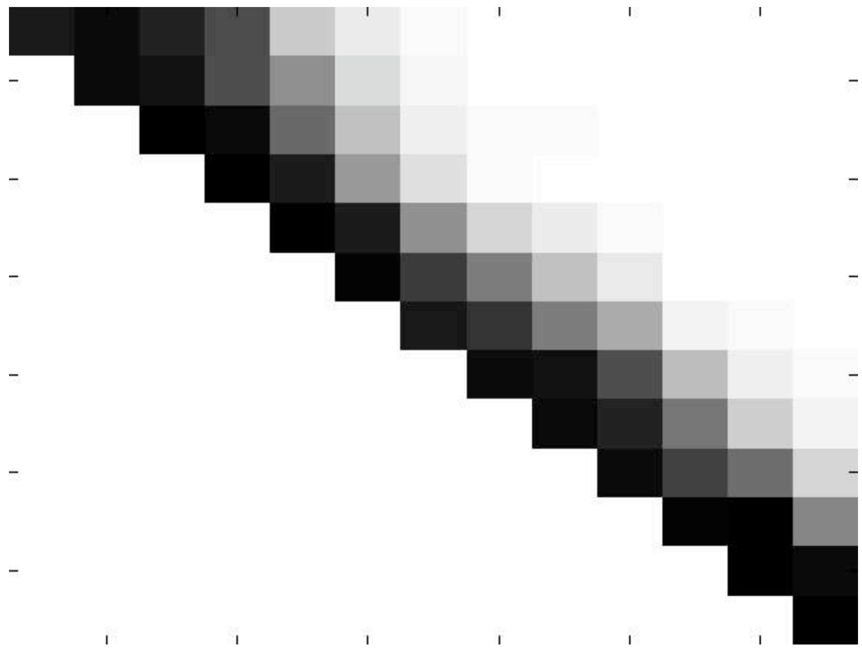


Results: MFCCs with VTLN

Humans



Model

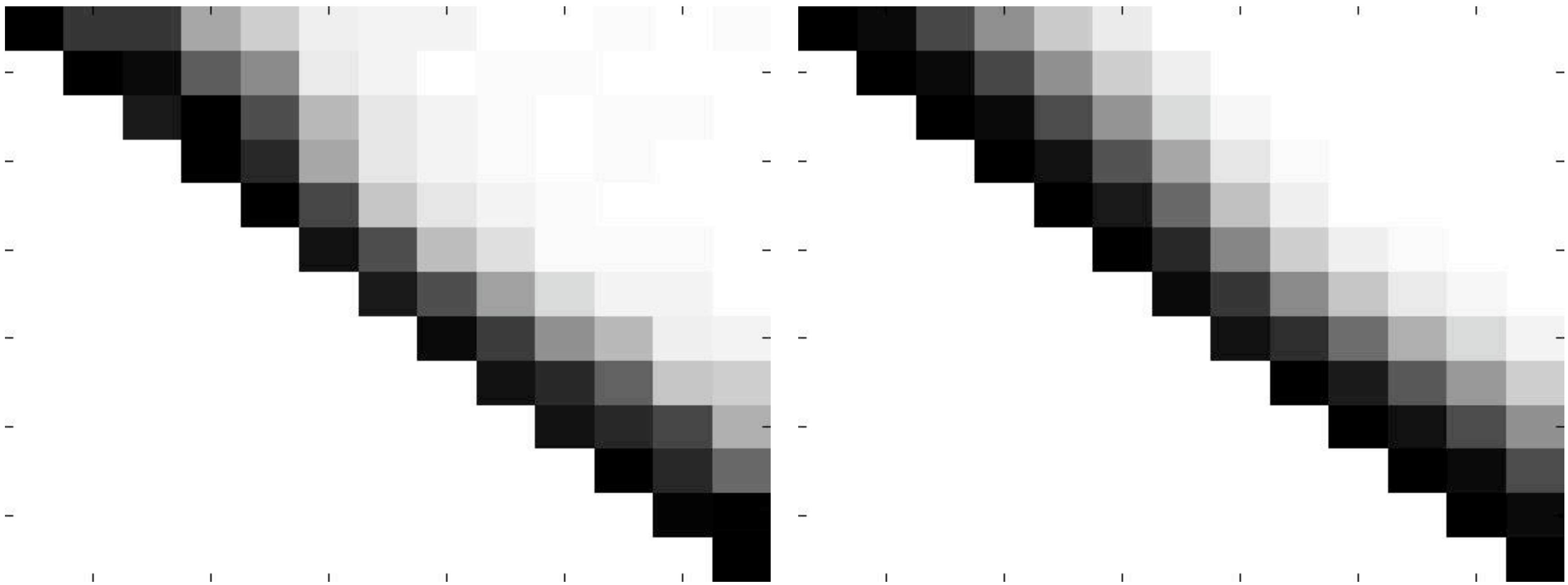


Results: Likelihoods

	Log Likelihood
Raw MFCCs	-490
MFCCs with VTLN	-255

Perceptual Baseline

Humans **Model**



Likelihoods

	Log Likelihood
Raw MFCCs	-490
MFCCs with VTLN	-255
Gaussians estimated from perceptual identification data	-223

Discussion

- Vocal tract length normalization improves prediction of human perceptual data from speech exemplars
- Underperforms prior distribution estimated from perceptual data
  → Neither of these sets of dimensions is exactly right

An Evaluation Metric

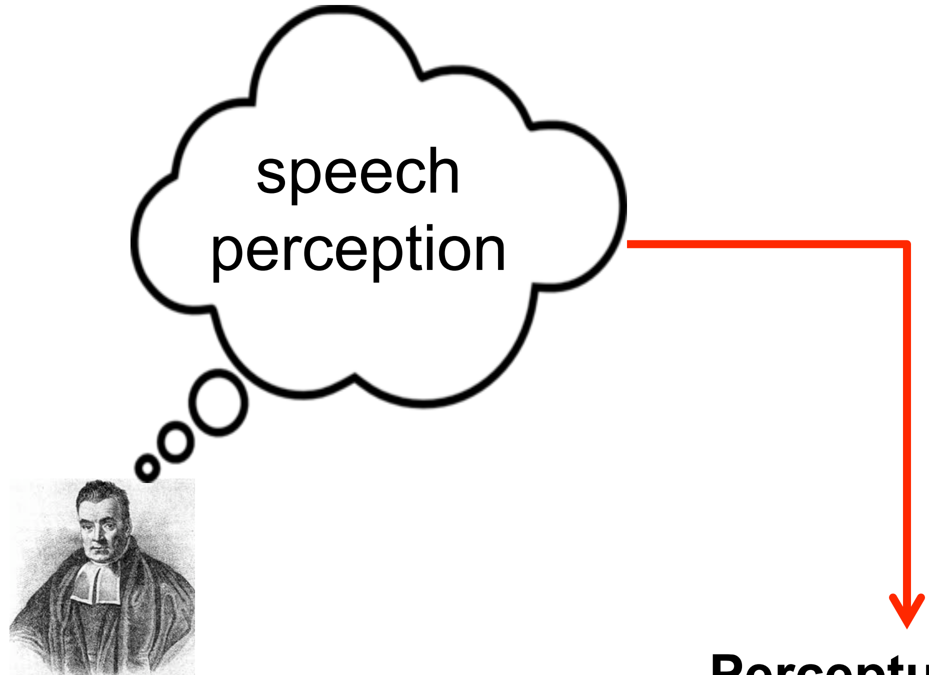
Linguists

- Formant frequencies
- Formant transitions
- Voice onset time
- Pitch
- Duration

Engineers

- Mel frequency cepstral coefficients (MFCC)
- Perceptual linear prediction (PLP)
- Relative spectral encoding (RASTA)
- Posteriorgrams

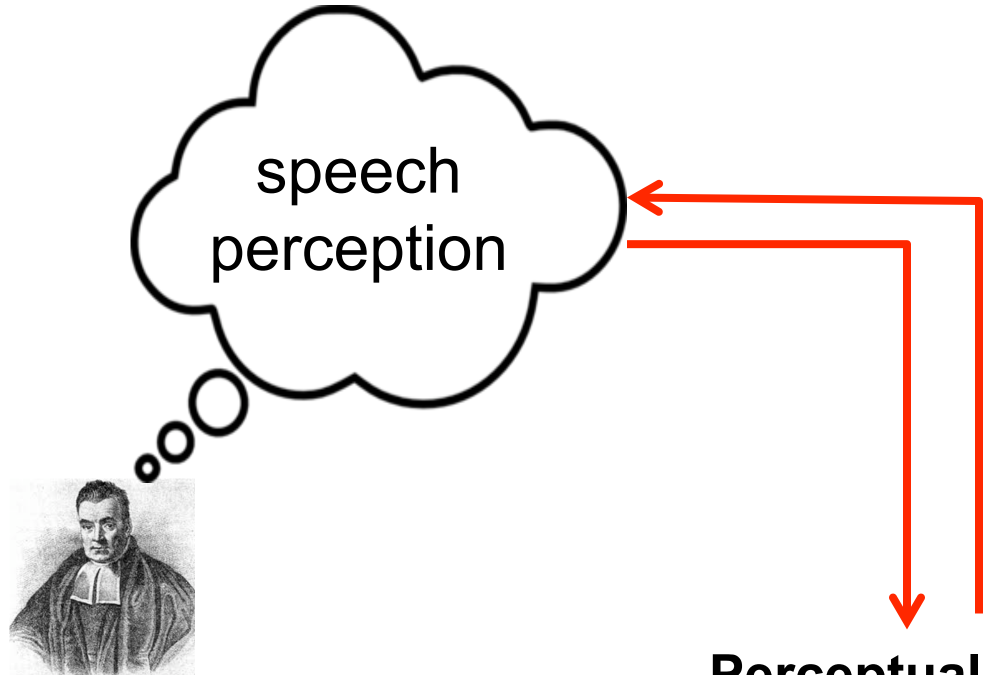
A Model of Speech Perception



Speech corpora
(prior distribution)

Perceptual data
(posterior distribution)

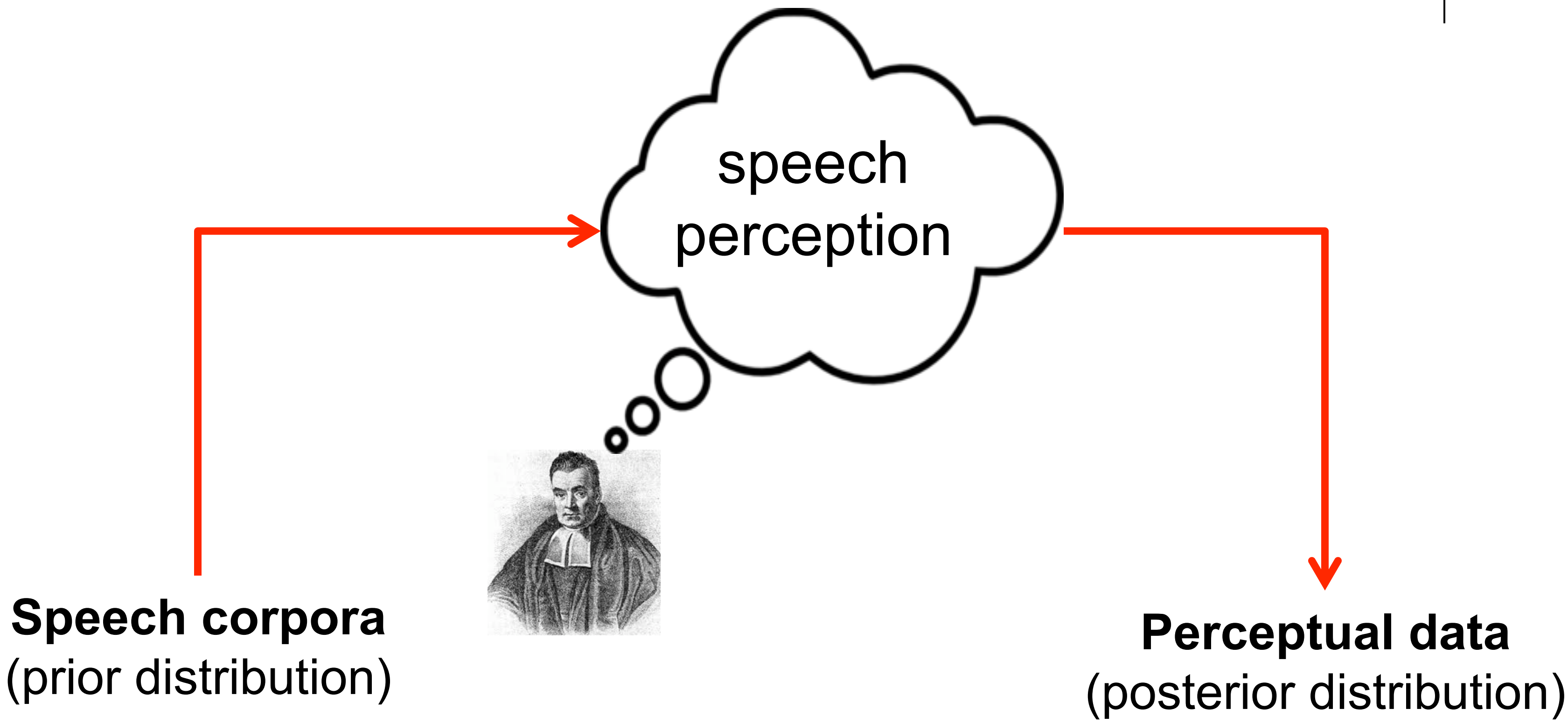
A Model of Speech Perception



Speech corpora
(prior distribution)

Perceptual data
(posterior distribution)

A Model of Speech Perception



Conclusions

- A model of speech perception that captures behavioral data in a more ecologically valid setting
  - Unifies perceptual data from consonants and vowels
  - Predicts perception in noise
  - Method for evaluating which speech features are most similar to the perceptual dimensions used by human listeners

- Cognitive models provide a way to link corpus data with behavioral psycholinguistic data in a principled way

Acknowledgments

Vowel perception model: Joint work with Tom Griffiths, James Morgan

Consonants vs. vowels: Joint work with Yakov Kronrod, Emily Coppess

Importance sampling: Joint work with Lei Shi, Tom Griffiths, Adam Sanborn

Speech corpora: Joint work with Caitlin Richter, Aren Jansen

Thanks to Sheila Blumstein, Adam Darlow, Bill Idsardi, Josh Falk, Sharon Goldwater, Hynek Hermansky, Sol Lago, Feipeng Li, Vijay Peddinti, Lori Rolfe, Phani Sankar, Mathias Scharinger, and Amy Weinberg for helpful comments, sharing data, and advice on speech features

This work was supported by:
NSF BCS-0631518, BCS-1320410
NSF IGERT DGE-9870676, DGE-0801465
NIH HD032005
AFOSR FA9550-07-1-0351