Gaussian Mixture Model Phonetics

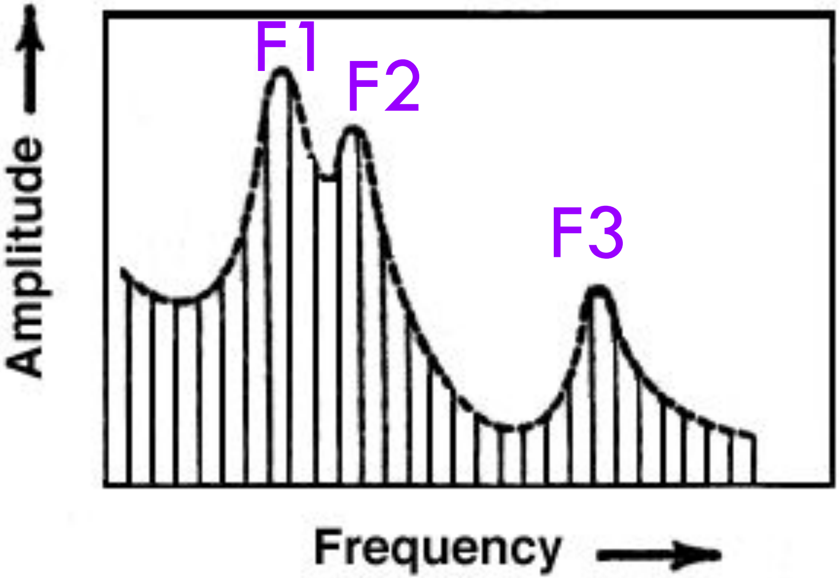
LING 492B

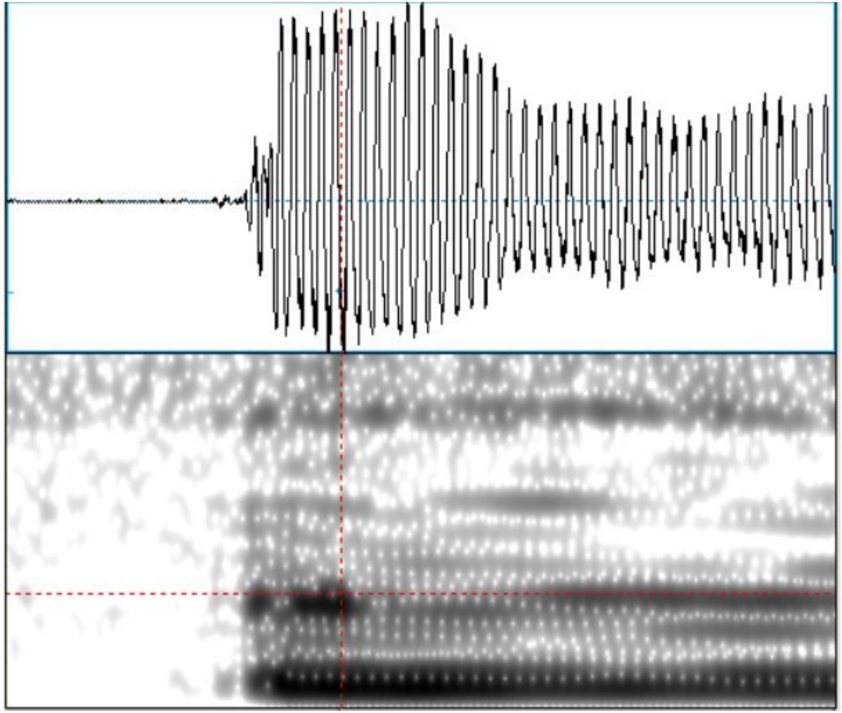
Motivation: Phonetic Distributions

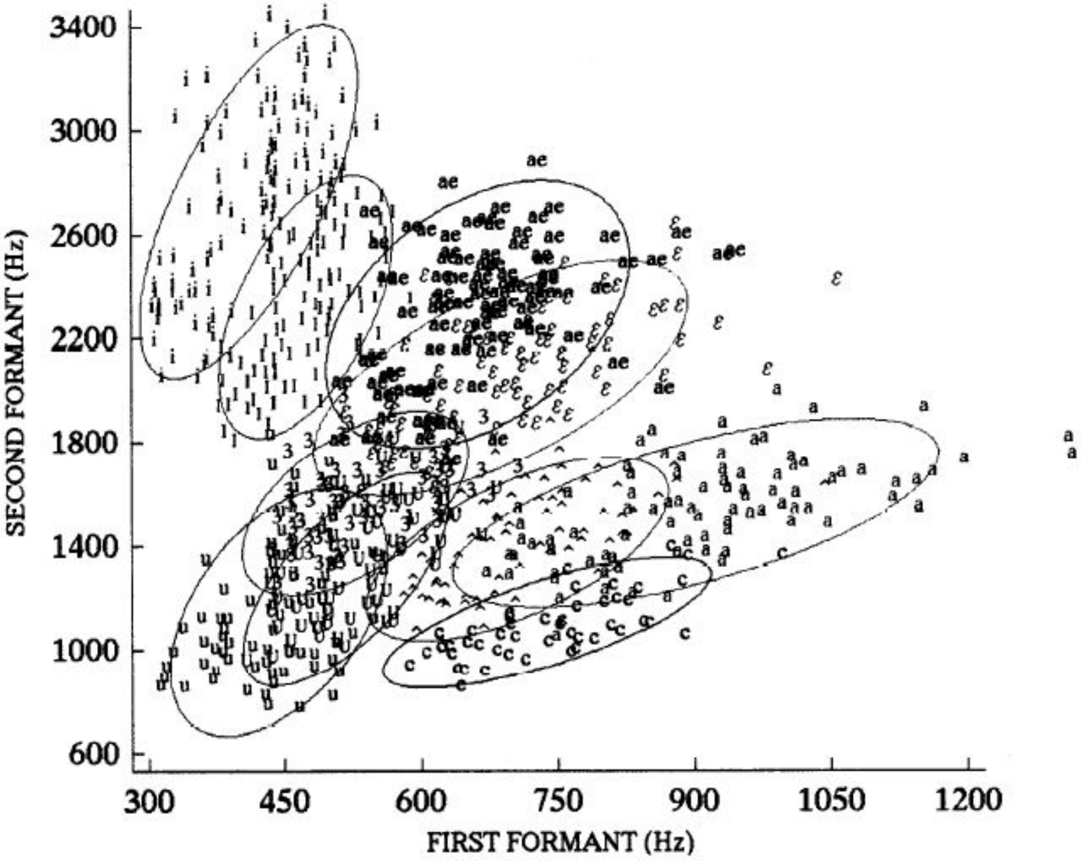
- Formants: resonant frequencies (Hz)
    - Vocal tract shape
    - e.g. F1, F2

- Acoustic cue to vowel category

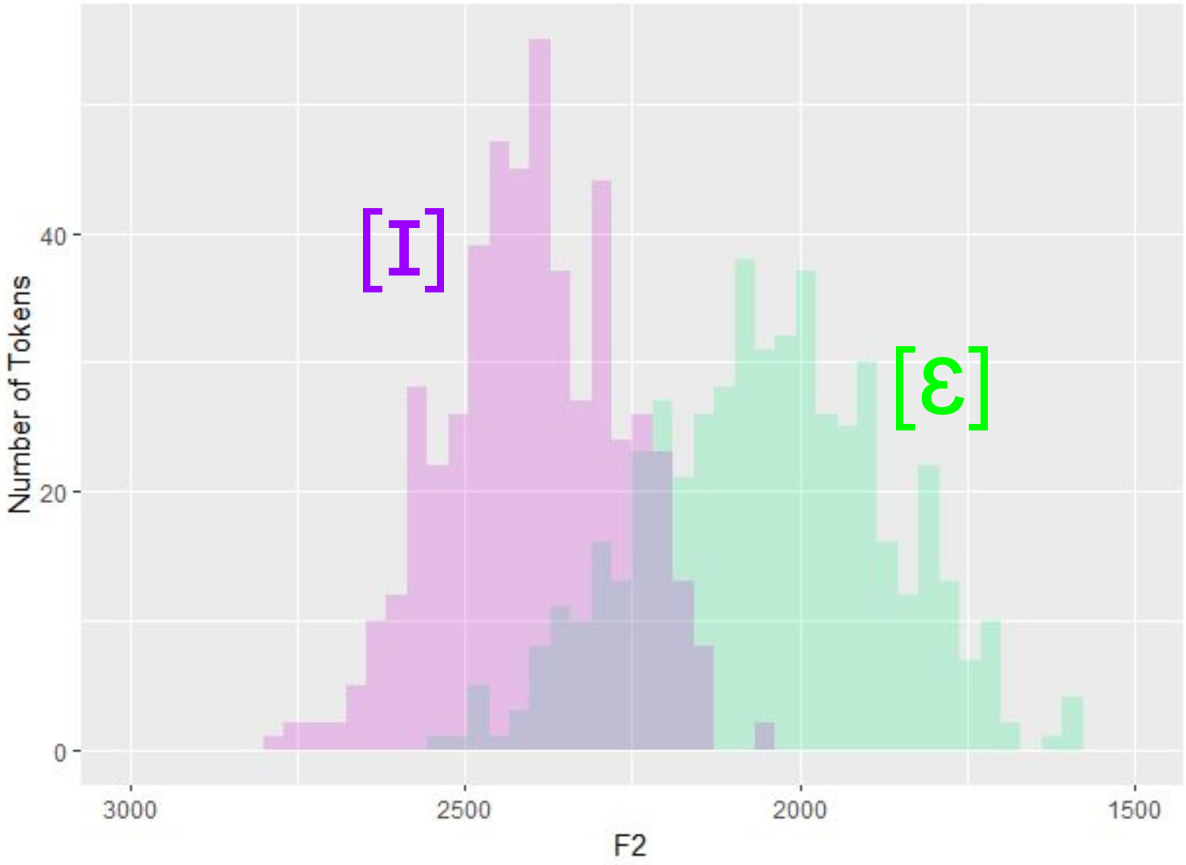
Formants






SECOND FORMANT (Hz)
3400
3000
2600
2200
1800
1400
1000
600
300
450
600
750
900
1050
1200
FIRST FORMANT (Hz)

Focusing on F2, ɪ, ɛ

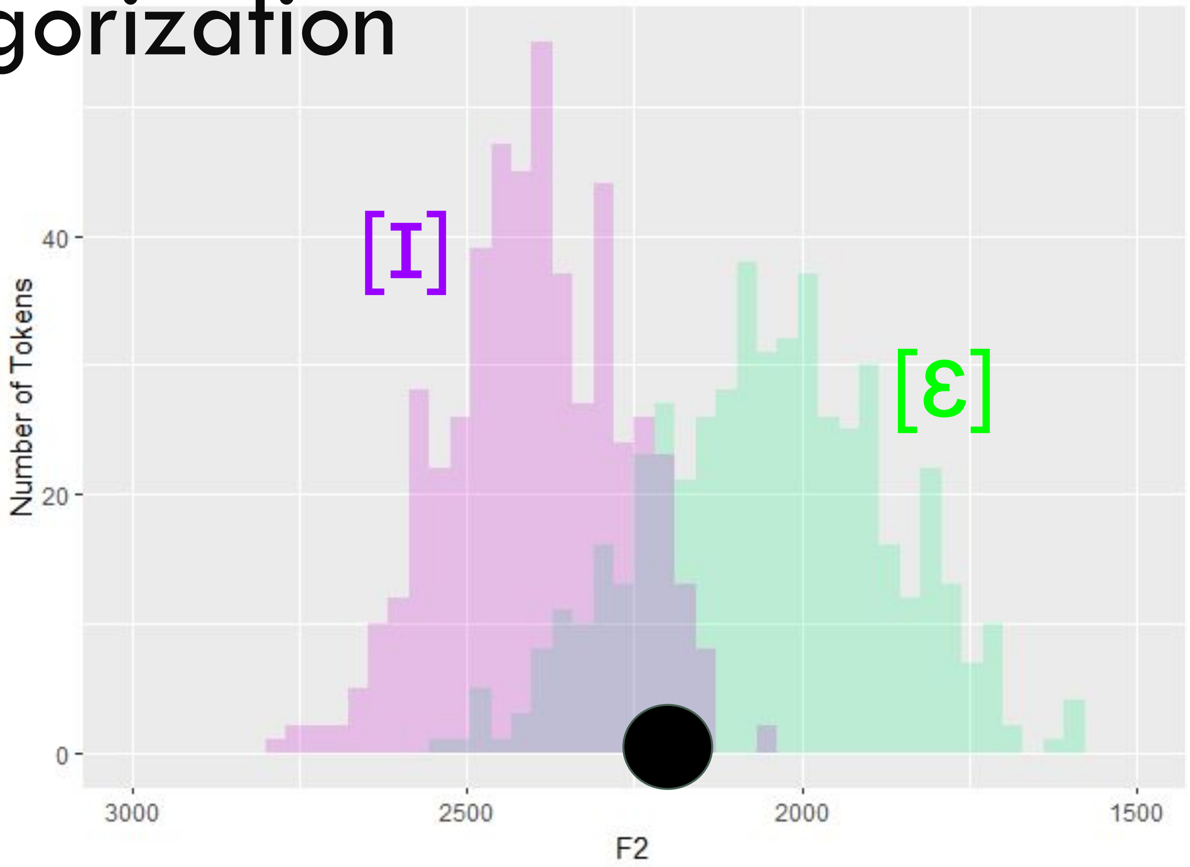


Inspired by Pierrehumbert (2001)

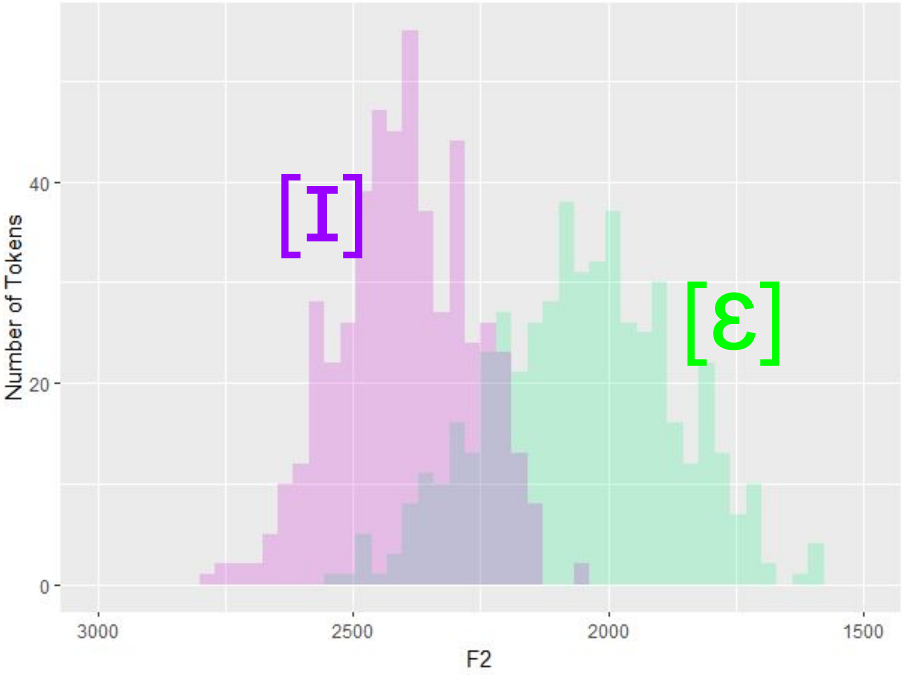
Task 1: Categorization

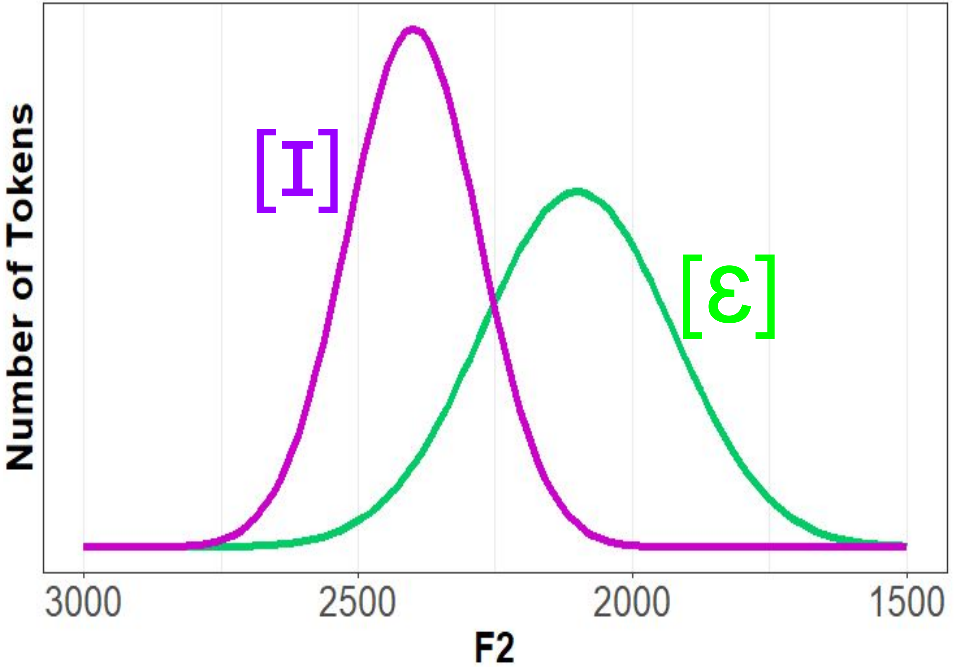
F2: 2200Hz

Did I hear “pin” or “pen”?



Approximating with Gaussians



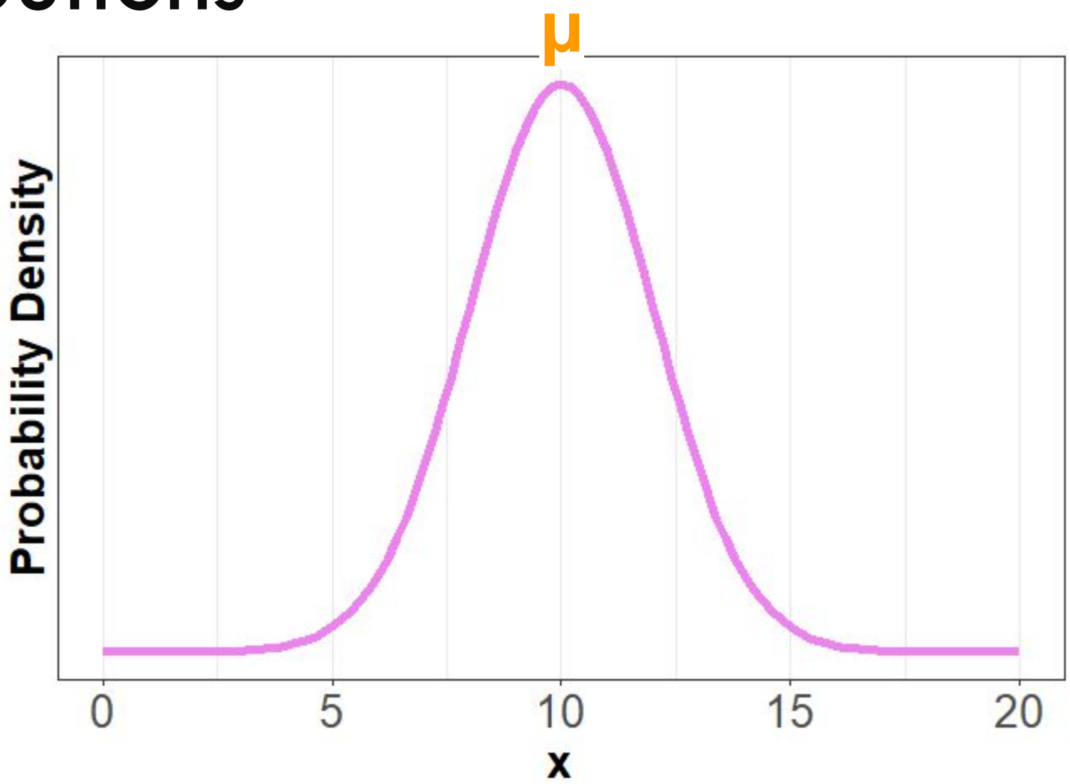


Gaussian distributions

Parameters:

Mean (mu or μ)

$$\sum_{i}^{N} p(X_i) X_i$$



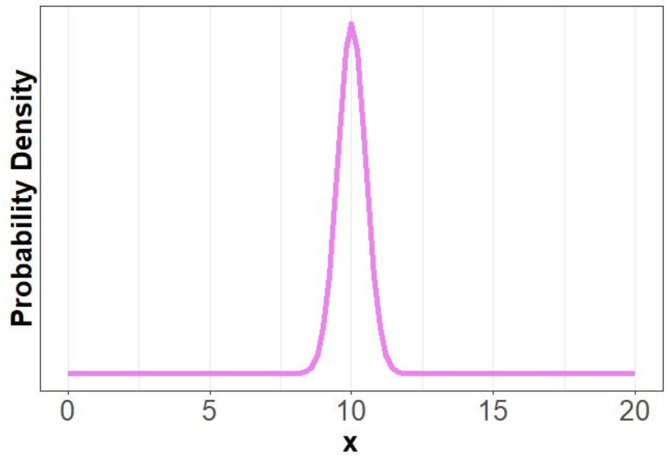
Gaussian distributions

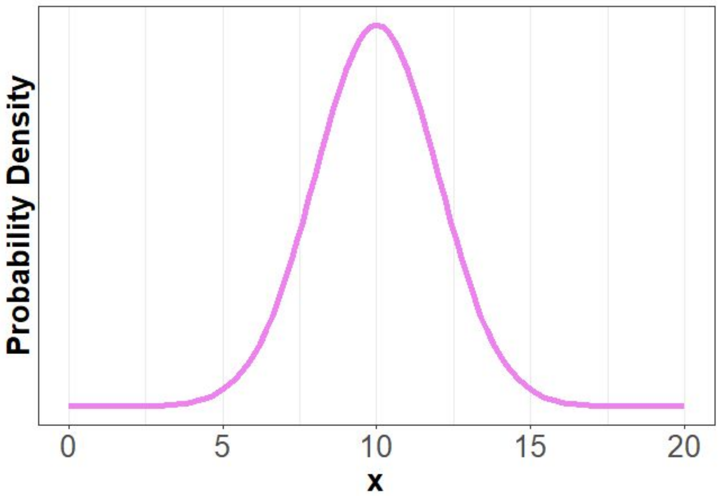
Parameters:

Mean (μ)

Variance (σ^2): average spread from mean

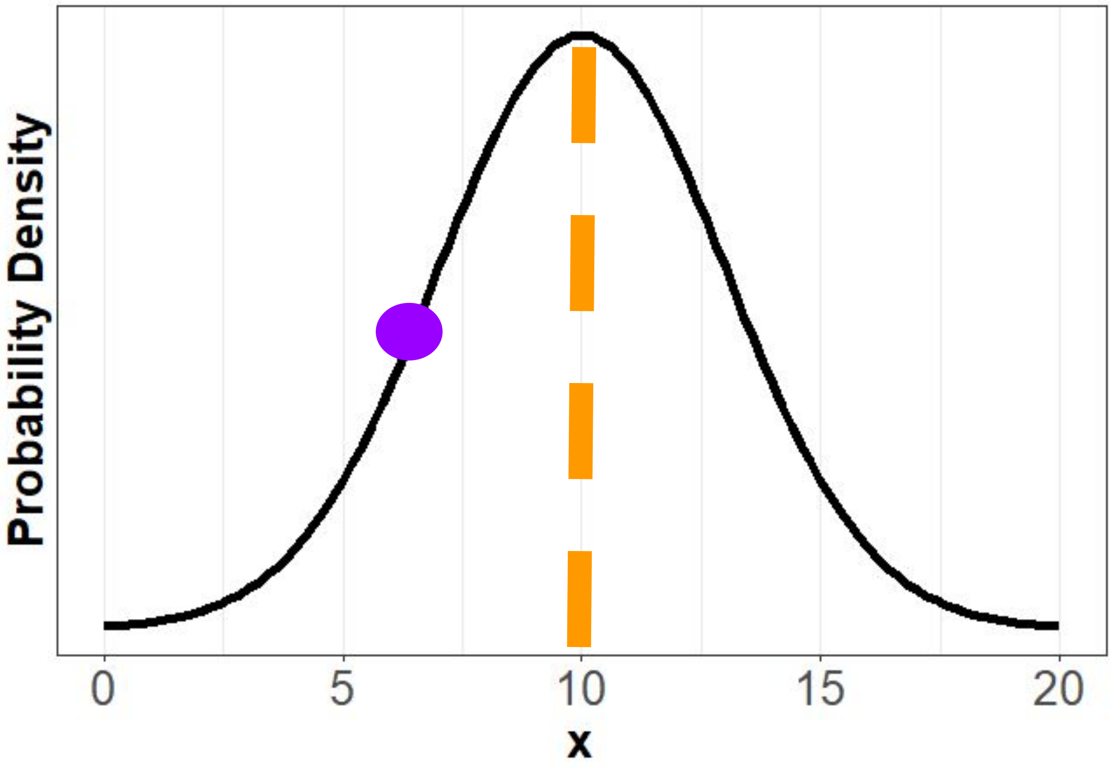
$$\sum_{i}^{N} p(X_i)(X_i - \mu)^2$$





Relative Likelihood

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



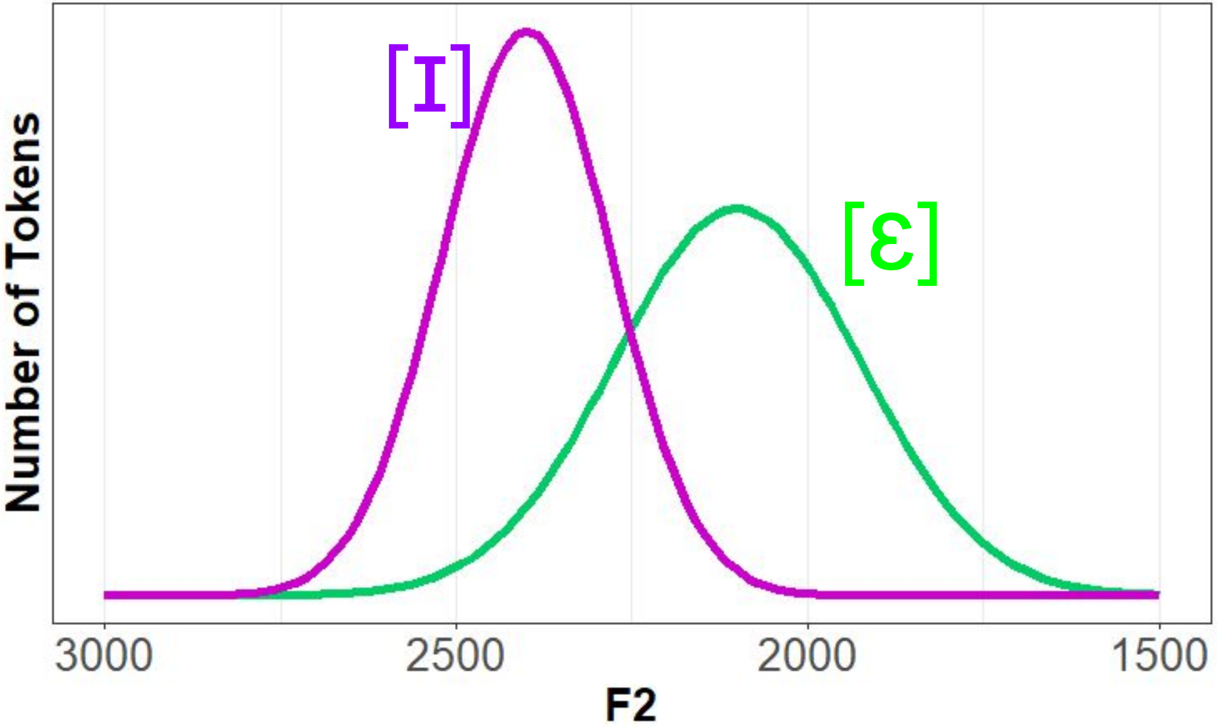
Mixture of Gaussians Parameters

$\mu_{ɪ}$, $\sigma^2_{ɪ}$

P(ɪ) (also called $\pi_{ɪ}$)

μ_{ε}, σ^2_{ε}

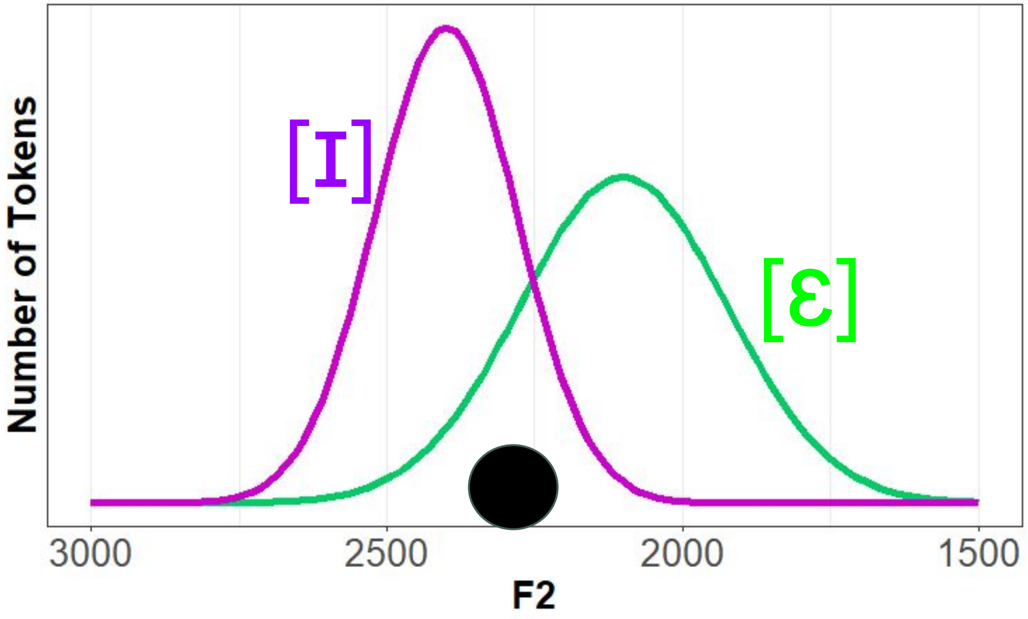
P(ɛ) (also called $\pi_{ɪ}$)



GMM Categorization

Did I hear “pin” or “pen”?

x = 2300Hz



We want:

P([ɪ]|x=2300)

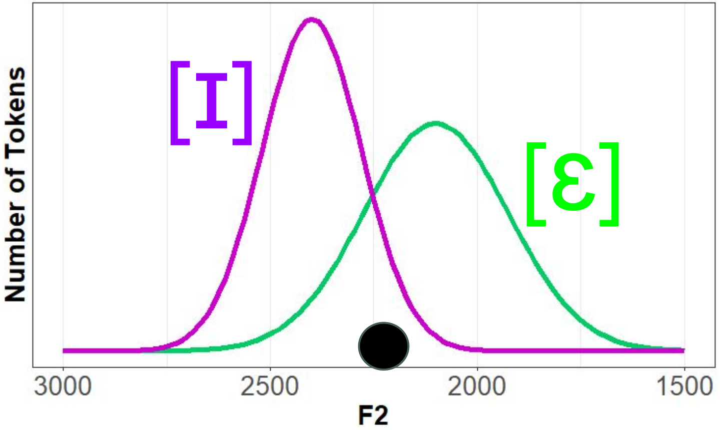
P([ɛ]|x=2300)

Bayes again!

P(A|B) = P(B|A)P(A)/P(B)

P([ɪ]|x) = P(x|[ɪ])P([ɪ])/P(x)

P([ɛ]|x) = P(x|[ɛ])P([ɛ])/P(x)



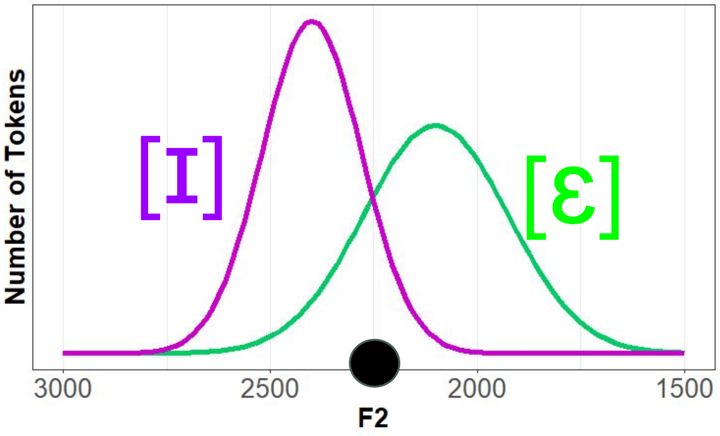
Bayes again!

$P([\text{ɪ}] | x) =$

$P(x | [\text{ɪ}])P([\text{ɪ}])/P(x) =$

$N(x | \mu_{\text{ɪ}}, \sigma^2_{\text{ɪ}}) * P([\text{ɪ}])/P(x) =$

$0.0012 / P(x)$



$\mu_{\text{ɪ}} = 2400$

$\sigma^2_{\text{ɪ}} = 100$

$P([\text{ɪ}]) = 0.5$

$x = 2300$

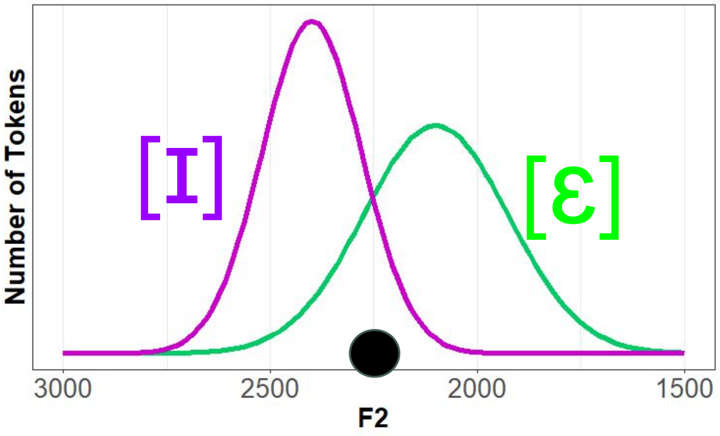
Bayes again!

$P([\varepsilon] \mid x) =$

$P(x \mid [\varepsilon])P([\varepsilon])/P(x) =$

$N(x \mid \mu_{\varepsilon}, \sigma^2_{\varepsilon}) * P([\varepsilon])/P(x) =$

$0.00060/P(x)$



$\mu_{\varepsilon} = 2100$

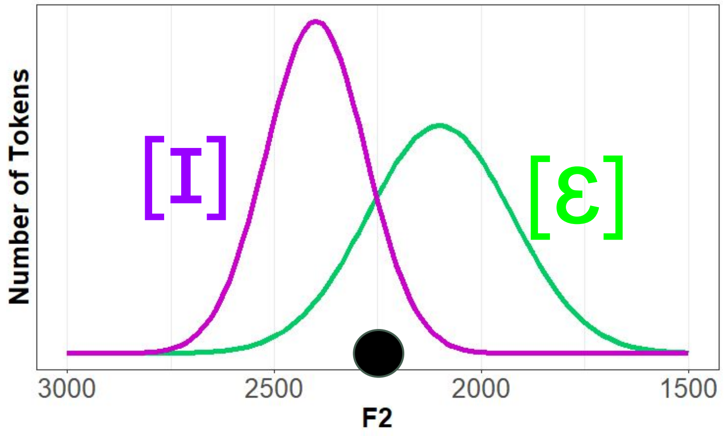
$\sigma^2_{\varepsilon} = 180$

$P([\varepsilon]) = 0.5$

$x = 2300$

Bayes again!

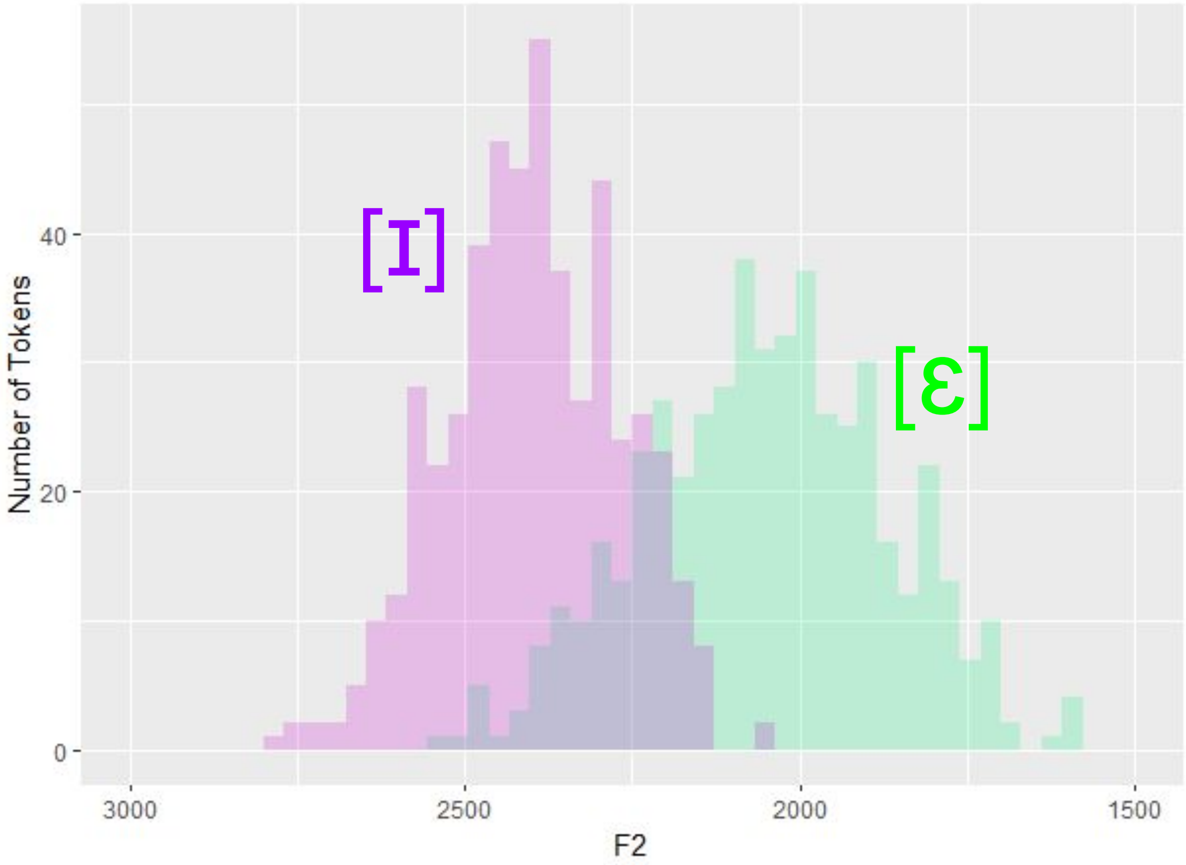
P([ɛ]|x) = 0.00060/P(x)
P([ɪ]|x) = 0.0012/P(x)



P([ɛ]|x) = 0.00060/0.00060 + 0.0012 = .33
P([ɪ]|x) = 0.0012/0.00060 + 0.0012 = .67

P([ɪ]|x) > P([ɛ]|x)

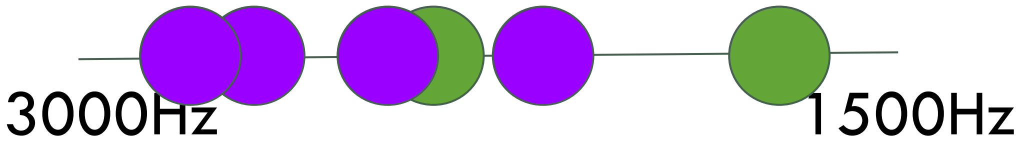
Estimating GMM parameters: labeled data



Estimating parameters: Labeled Data

[ɪ]: 2600Hz
[ɪ]: 2500Hz
[ɪ]: 2300Hz
[ɪ]: 2100Hz

[ε]: 2200Hz
[ε]:1600Hz



$\mu_{ɪ} = ?$

$\sigma^2_{ɪ} = ?$

$\mu_{\varepsilon} = ?$

$\sigma^2_{\varepsilon} = ?$

$P(\varepsilon) = ?$

$P(ɪ) = ?$

Estimating GMM: Labeled Data

[ɪ]: 2600Hz
[ɪ]: 2500Hz
[ɪ]: 2300Hz
[ɪ]: 2100Hz

[ɛ]: 2200Hz
[ɛ]:1600Hz

$\mu_{ɪ} = 2375$

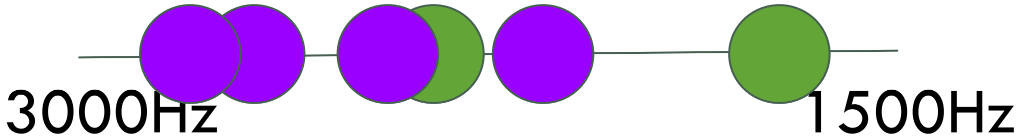
$\sigma^2_{ɪ} = ?$

$\mu_{ɛ} = ?$

$\sigma^2_{ɛ} = ?$

$P(ɛ) = ?$

$P(ɪ) = ?$

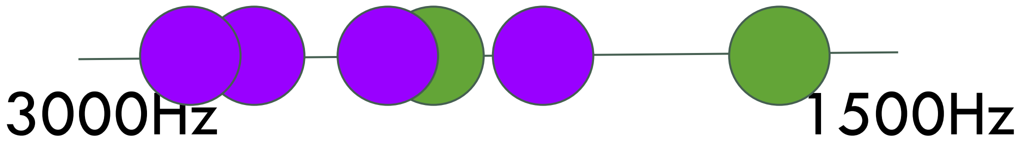


Estimating GMM: Labeled Data

[ɪ]: 2600Hz
[ɪ]: 2500Hz
[ɪ]: 2300Hz
[ɪ]: 2100Hz

[ɛ]: 2200Hz
[ɛ]:1600Hz

$\mu_{ɪ}$ = 2375
$\sigma^2_{ɪ}$ = 221
$\mu_{ɛ}$ = ?
$\sigma^2_{ɛ}$ = ?
P(ɛ)= ?
P(ɪ)= ?



Estimating GMM: Labeled Data

[ɪ]: 2600Hz
[ɪ]: 2500Hz
[ɪ]: 2300Hz
[ɪ]: 2100Hz

[ɛ]: 2200Hz
[ɛ]:1600Hz

$\mu_{\text{ɪ}} = 2375$

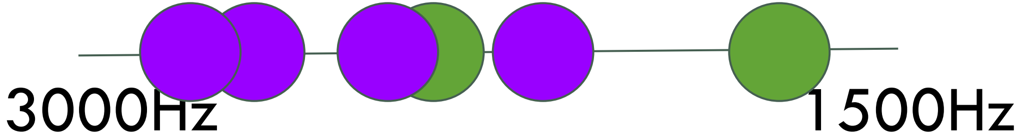
$\sigma^2_{\text{ɪ}} = 221$

$\mu_{\varepsilon} = 1900$

$\sigma^2_{\varepsilon} = ?$

$P(\varepsilon) = ?$

$P(\text{ɪ}) = ?$



Estimating GMM: Labeled Data

[ɪ]: 2600Hz
[ɪ]: 2500Hz
[ɪ]: 2300Hz
[ɪ]: 2100Hz

[ɛ]: 2200Hz
[ɛ]:1600Hz

$\mu_{ɪ} = 2375$

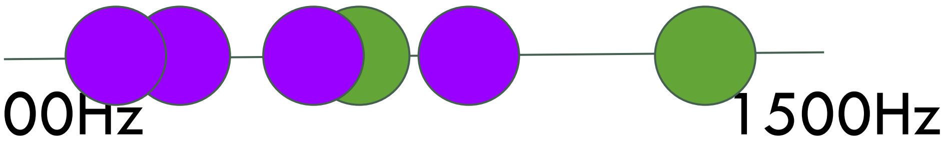
$\sigma^2_{ɪ} = 221$

$\mu_{\varepsilon} = 1900$

$\sigma^2_{\varepsilon} = 424$

P(ɛ)= ?

P(ɪ)= ?



Estimating GMM: Labeled Data

[ɪ]: 2600Hz
[ɪ]: 2500Hz
[ɪ]: 2300Hz
[ɪ]: 2100Hz

[ɛ]: 2200Hz
[ɛ]:1600Hz

$\mu_{ɪ} = 2375$

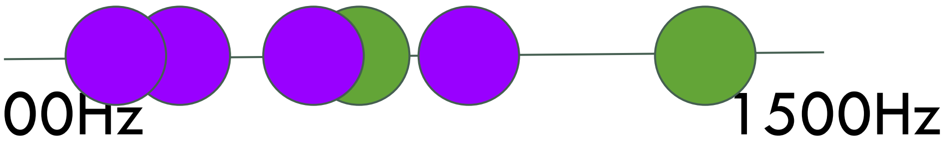
$\sigma^2_{ɪ} = 221$

$\mu_{\varepsilon} = 1900$

$\sigma^2_{\varepsilon} = 424$

$P(\varepsilon) = 2/6$

$P(ɪ) = ?$



Estimating GMM: Labeled Data

[ɪ]: 2600Hz
[ɪ]: 2500Hz
[ɪ]: 2300Hz
[ɪ]: 2100Hz

[ɛ]: 2200Hz
[ɛ]:1600Hz

$\mu_{\text{ɪ}} = 2375$

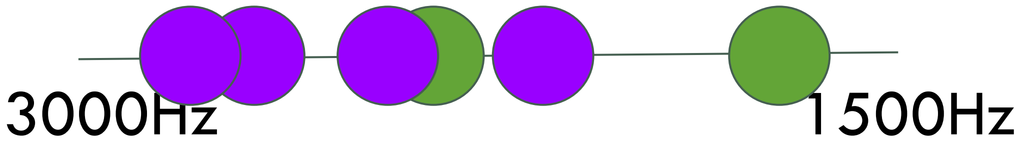
$\sigma^2_{\text{ɪ}} = 221$

$\mu_{\varepsilon} = 1900$

$\sigma^2_{\varepsilon} = 424$

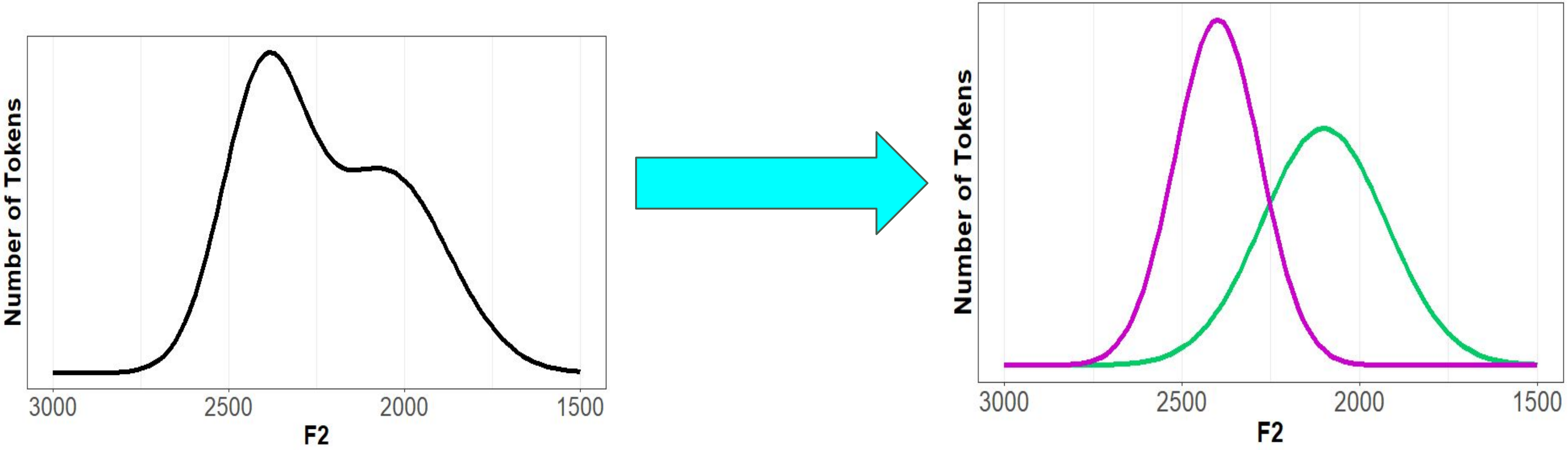
$P(\varepsilon) = 2/6$

$P(\text{ɪ}) = 4/6$



Task #2: Unsupervised Learning

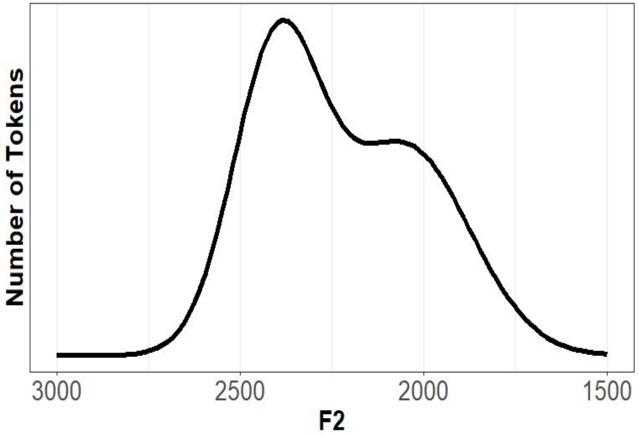
What are the categories? What are their parameters (μ, σ,π)?



Unsupervised learning: cognition

How do infants learn phoneme categories with so much overlap? [Feldman 2009, Vallabha 2007]

Without knowing anything about the categories beforehand, input data looks like this:



GMM Expectation-Maximization

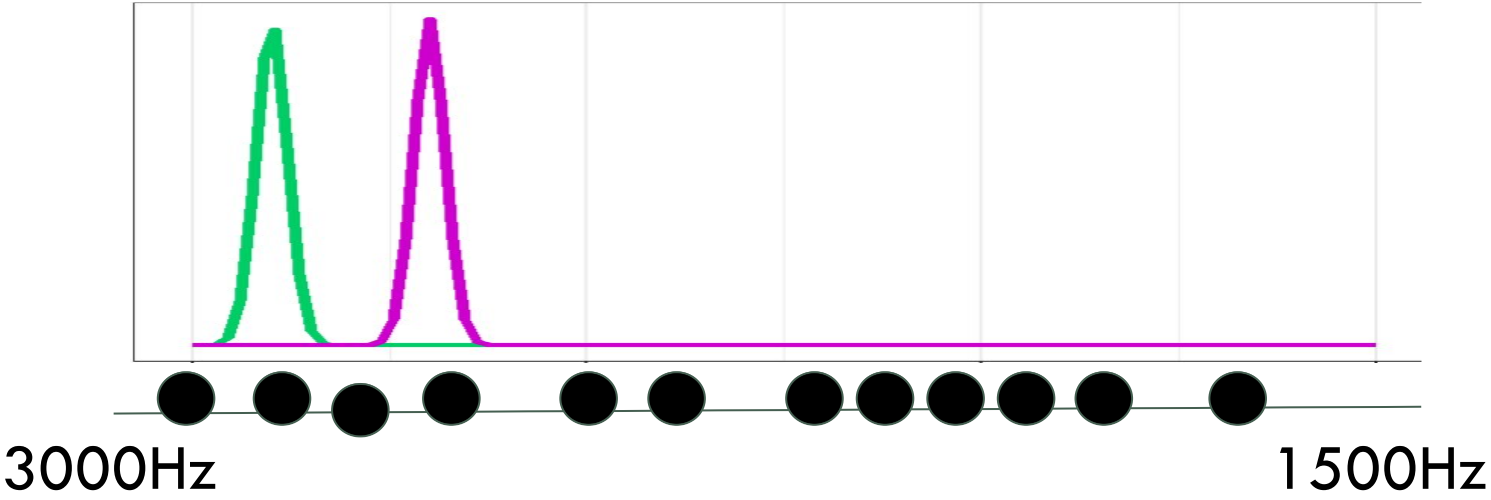
Intuition:

- If we knew the vowel labels, we could estimate mu and sigma for each category
    - But we don't know the vowel labels :(

- If we knew mu and sigma for each category, we could estimate the vowel labels
    - But we don't know the mus and sigmas :(

GMM Expectation-Maximization

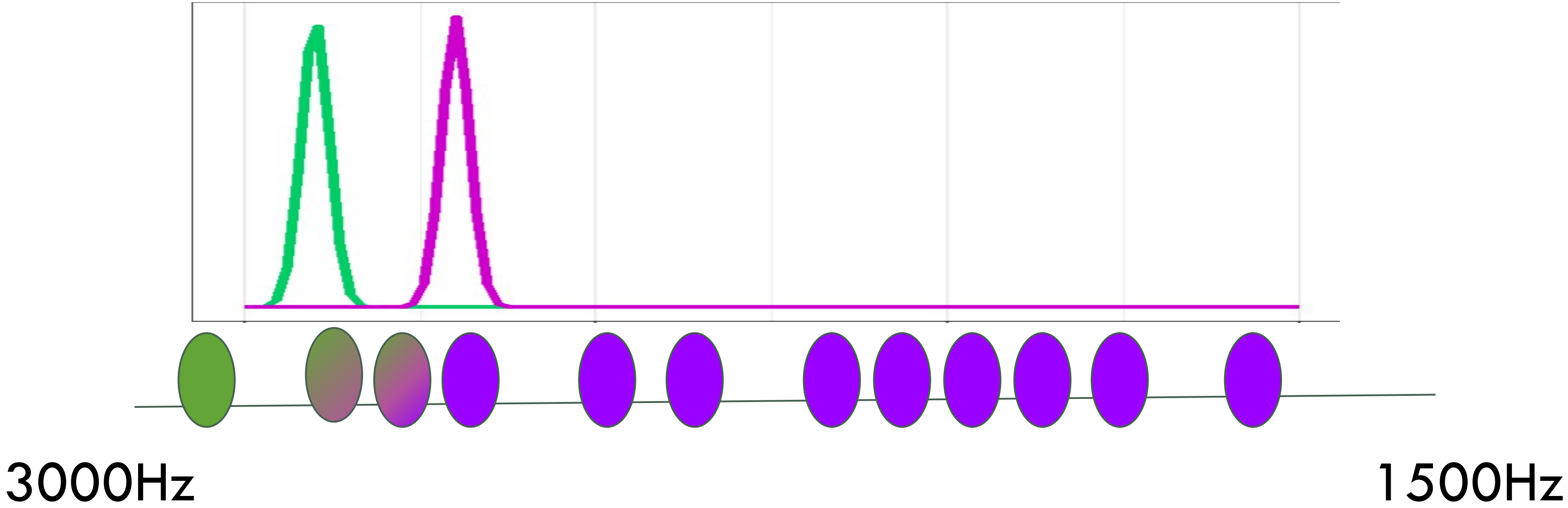
Initialization: Start with k categories with random means and variances [cf. k-means!]



Based on demo from Victor Lavrenko

GMM Expectation-Maximization

Expectation: How likely is each category for each label?

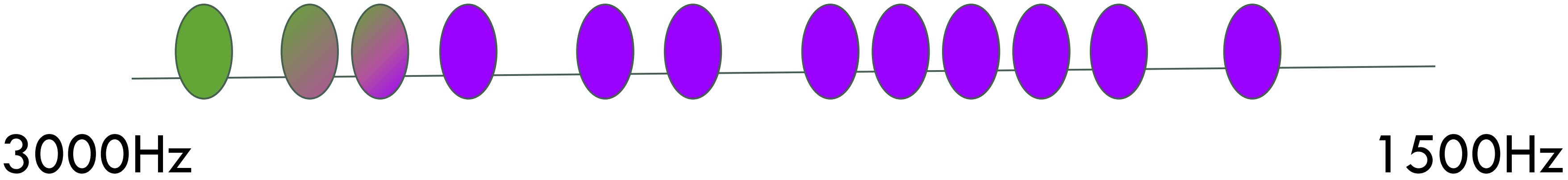


GMM Expectation

Expectation: How likely is each category for each label?

For each observation **x**, for each category c (μ_c, σ^2_c), compute:

$$\mathbf{P}(\mathbf{c}\,|\,\mathbf{x}) = N(x\,|\,\mu_c, \sigma^2) * P(c)/P(x)$$



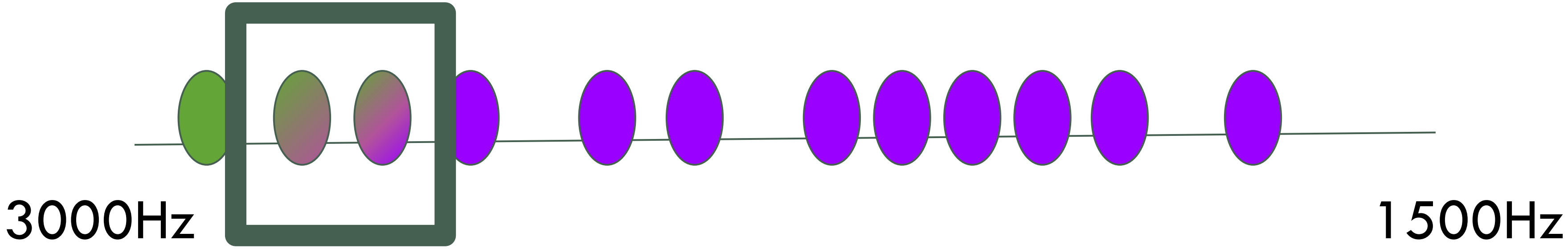
GMM Expectation

Expectation: How likely is each category for each label?

For each observation **x**, for each category c (μ_c, σ^2_c), compute:

$\mathbf{P}(\mathbf{c}|\mathbf{x}) = N(x | \mu_c, \sigma^2) * P(c)/P(x)$

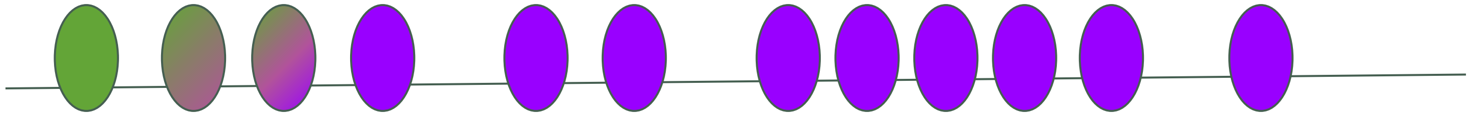
cf. k-means: *soft* categorization



GMM Maximization

Maximization: Update each category's parameters based on the observations

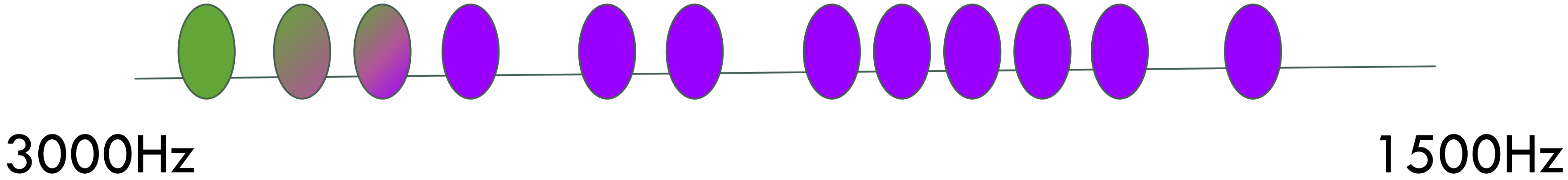
Each observation's contribution to the parameters is **weighed by P(category|observation)**



GMM Maximization

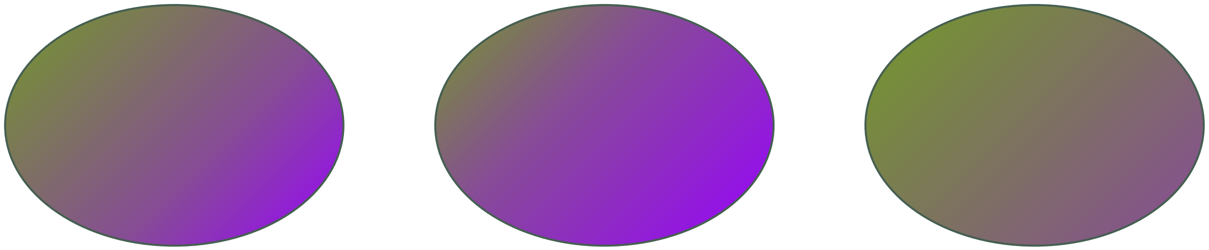
Maximization: Update each category's parameters based on the observations

$$\text{New } \mu_c = \frac{x_1 P(c|x_1) + x_2 P(c|x_2) + \ldots + x_n P(c|x_n)}{P(c|x_1) + P(c|x_2) + \ldots + P(c|x_n)}$$



GMM Maximization

- Think of each category as taking part of the responsibility for each observation
- That responsibility could be really big or small



GMM Maximization

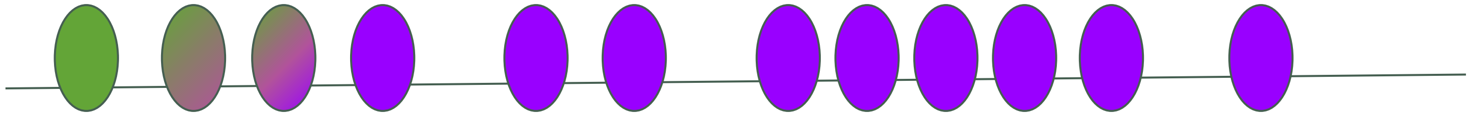
Just like a weighted/soft version of computing category mean

GMM new mean

$$\frac{x_1 P(c|x_1) + x_2 P(c|x_2) + \ldots + x_n P(c|x_n)}{P(c|x_1) + P(c|x_2) + \ldots + P(c|x_n)}$$

K-Means new mean

$$\frac{0 * x_1 + 1 * x_2 + \ldots + 1 * x_n}{0 + 1 + \ldots + 1}$$



3000Hz 1500Hz

GMM Maximization

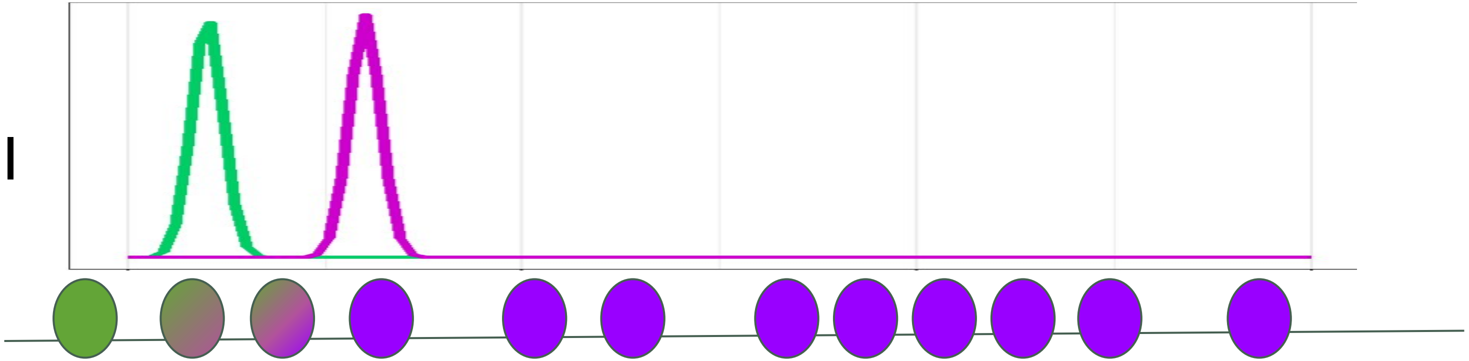
More expectation: do the same weighted estimates for the rest of the parameters

New σ_c^2: $\dfrac{P(c|x_1)(x_1-\mu_c)^2+P(c|x_2)(x_2-\mu_c)^2+\ldots+P(c|x_n)(x_n-\mu_c)^2}{P(c|x_1)+P(c|x_2)+\ldots+P(c|x_n)}$

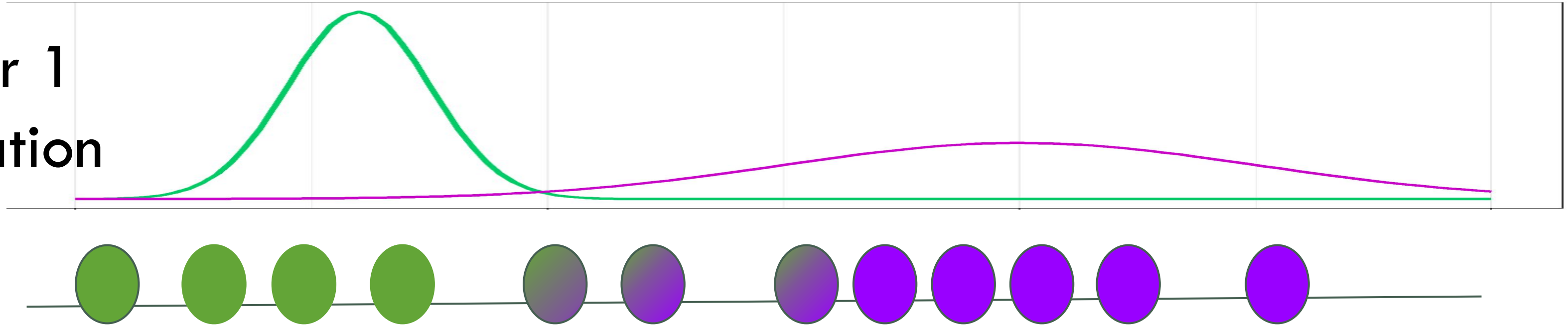
New $P(c)$: $\dfrac{P(c|x_1)+P(c|x_2)+\ldots+P(c|x_n)}{N}$

GMM Expectation-Maximization

Initial



After 1 iteration



Differences from k-means

- Soft instead of hard categorization while learning
- More parameters: prior probability of category, variance
- Guaranteed to increase likelihood of data given model at every step
- Could converge on local instead of global maximum

Multiple dimensions

Beyond just F2: can characterize vowels with F1 and F2 for 2-D Gaussian

+ more! (e.g. length)

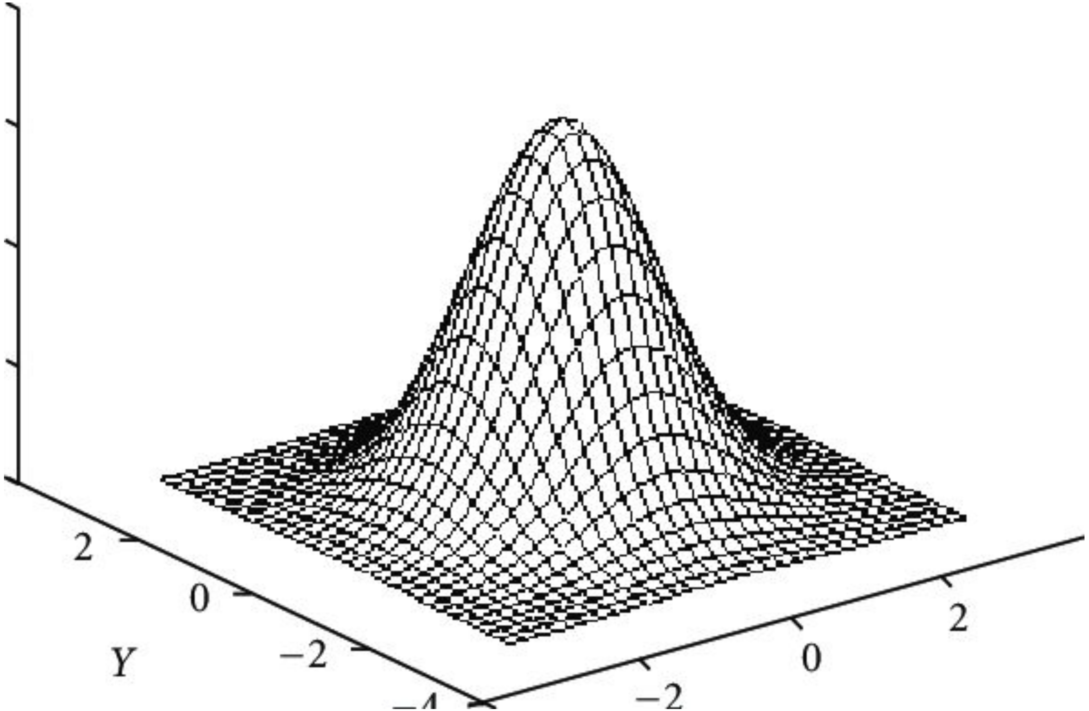


image credit: wikipedia

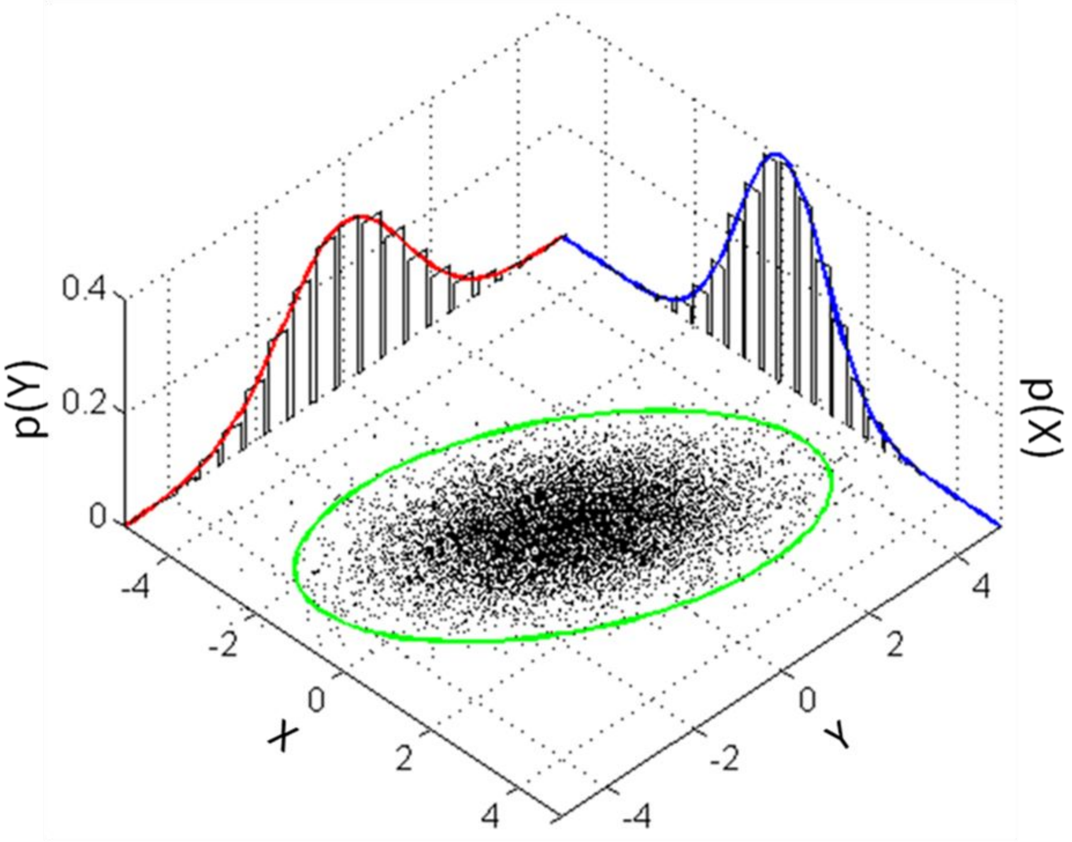
Multiple dimensions

- Category means for each dimension

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

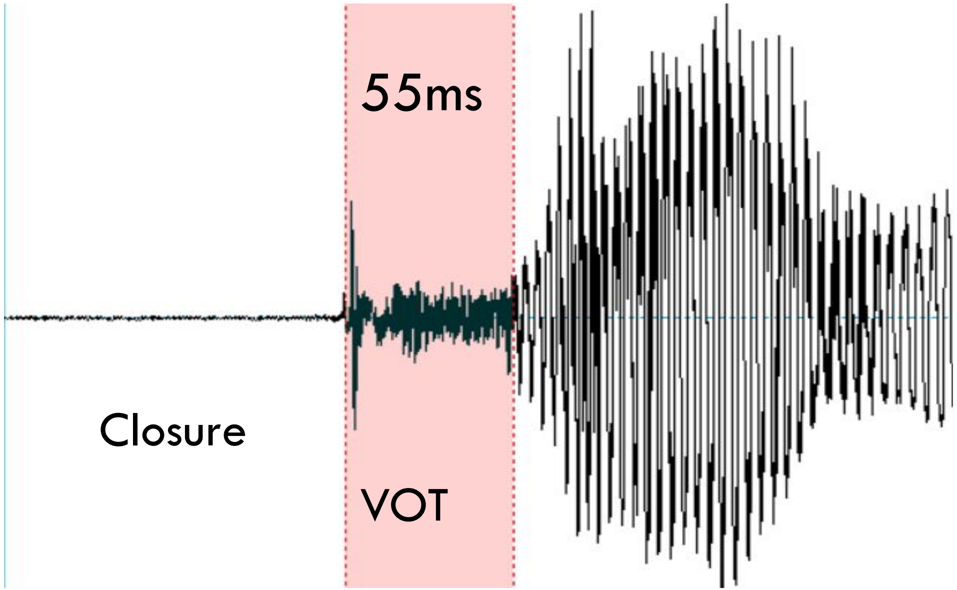
- Instead of just σ^2, covariance matrix

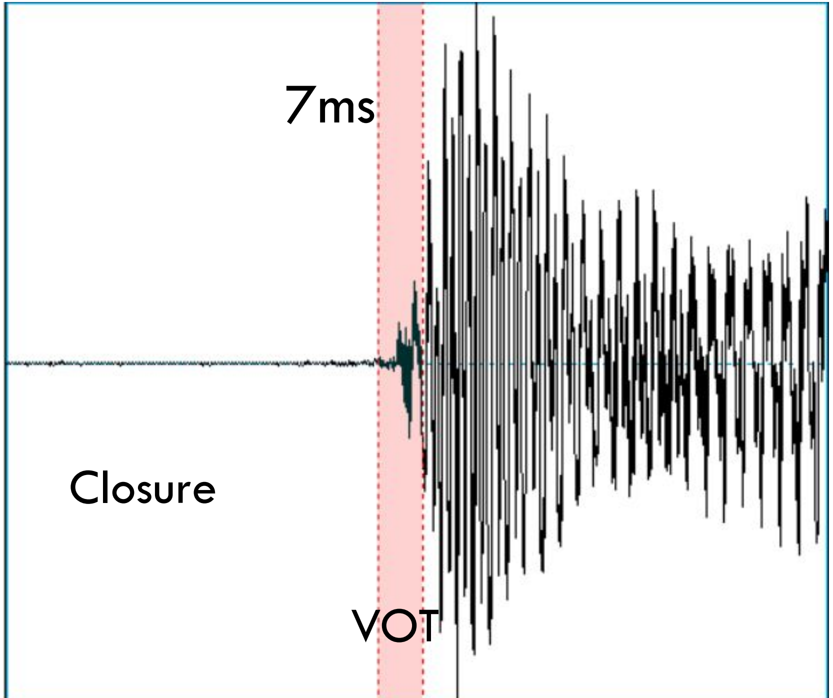
$$\begin{bmatrix} 1 & 3/5 \\ 3/5 & 2 \end{bmatrix}$$



Beyond vowels: Stop voicing

Voice Onset Time (VOT): pIn vs bIn





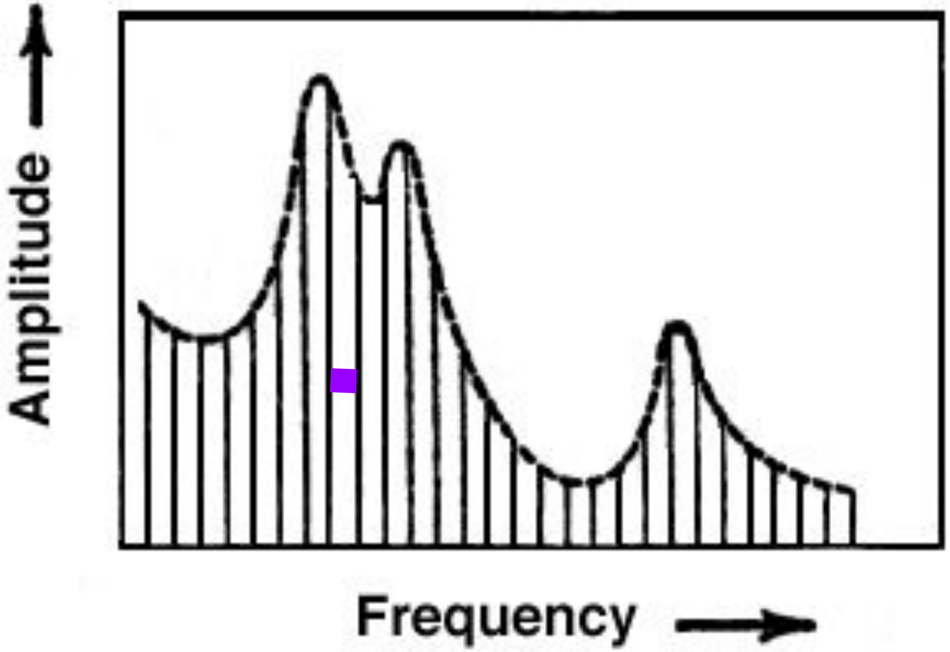
Stop voicing

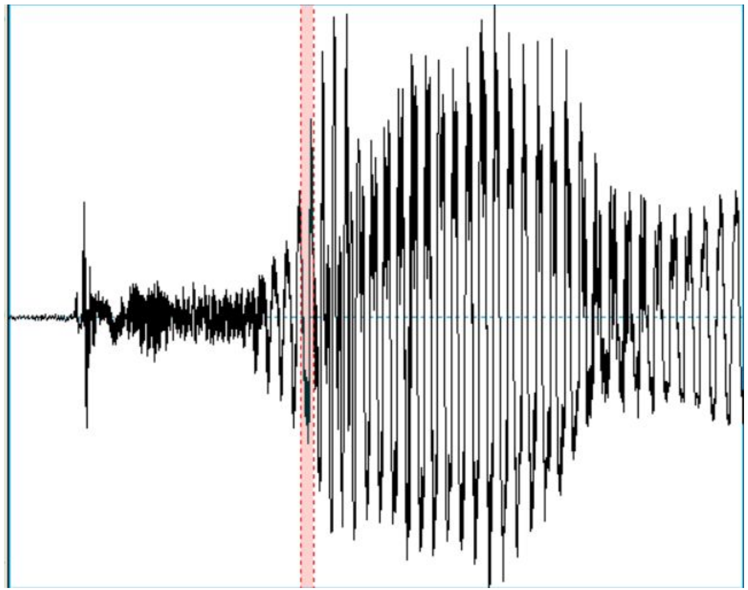
Fundamental frequency (f0) also correlates with stop voicing! f0 is:

- rate at which vocal folds are vibrating
- associated with pitch

- Tend to have lower f0 right next to voiced stops
- Tend to have higher f0 right next to voiceless stops

Measuring f0

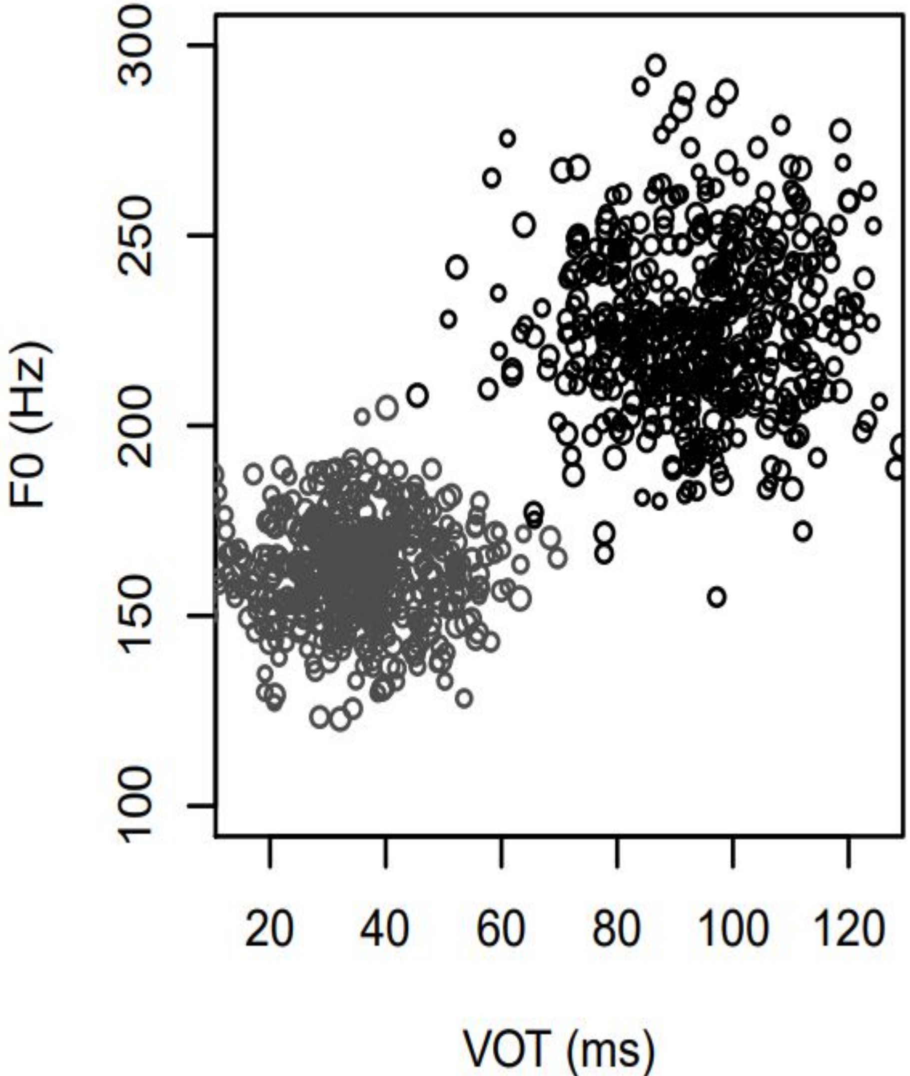




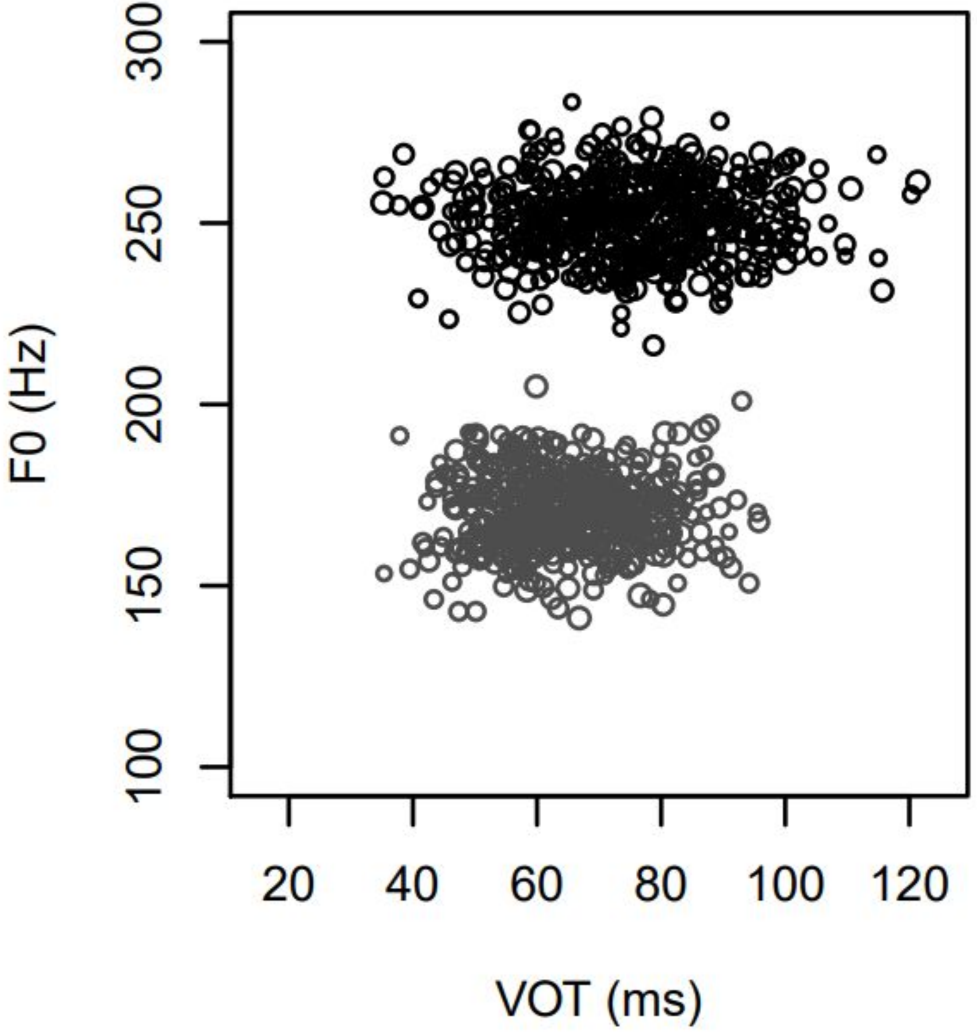
f0 and VOT in Korean stops

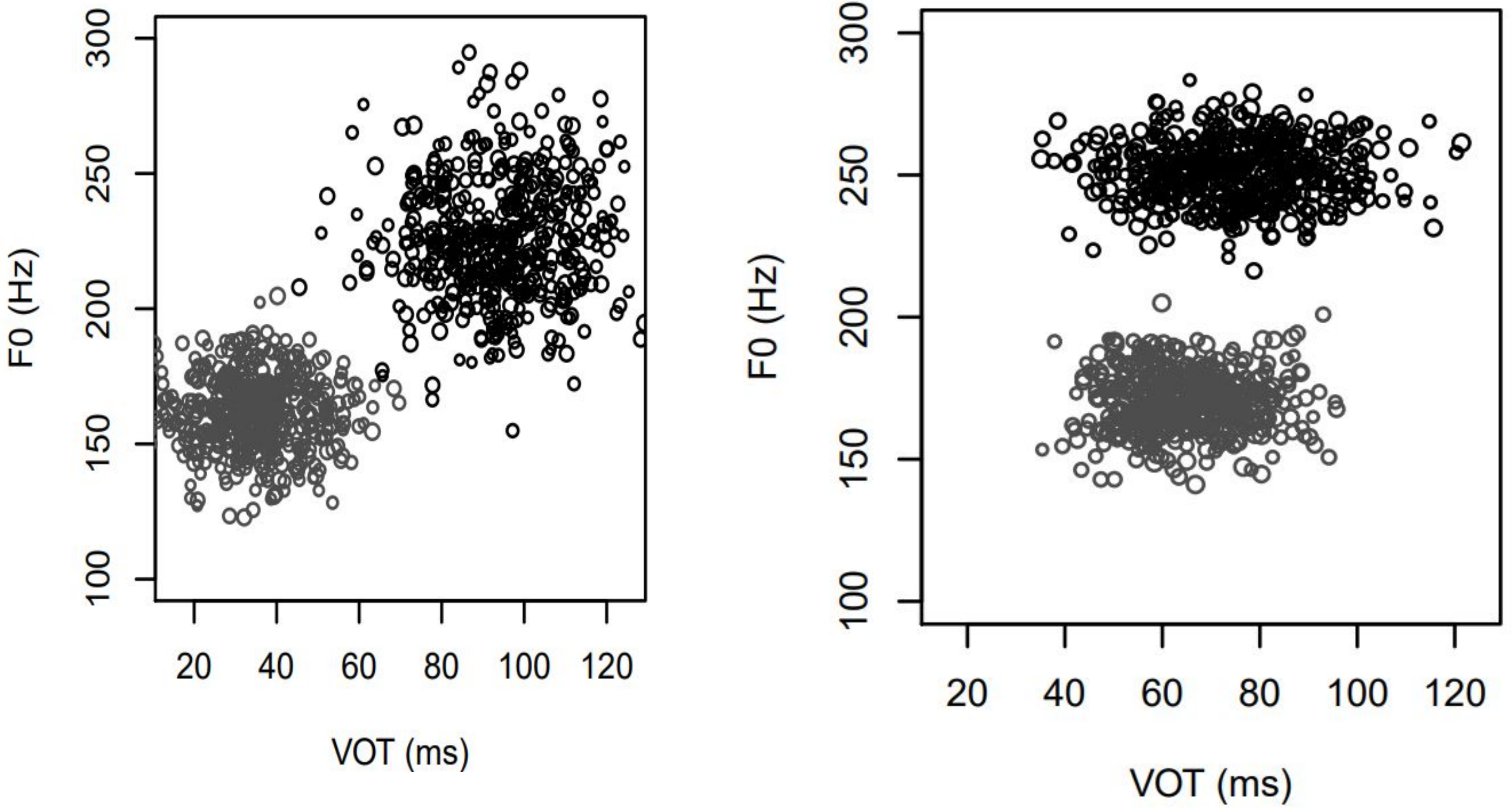
- Kirby (2013)
- Categories changing over time (ongoing!)
- Categories distinguished more and more by f0 than VOT
- Case of tonogenesis

1960



2000s

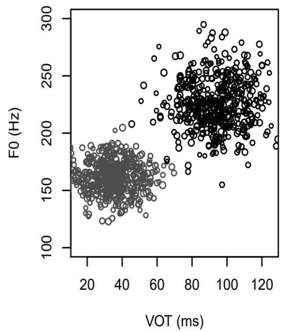




F0 (Hz)
100
150
200
250
300
20
40
60
80
100
120
VOT (ms)
F0 (Hz)
100
150
200
250
300
20
40
60
80
100
120
VOT (ms)

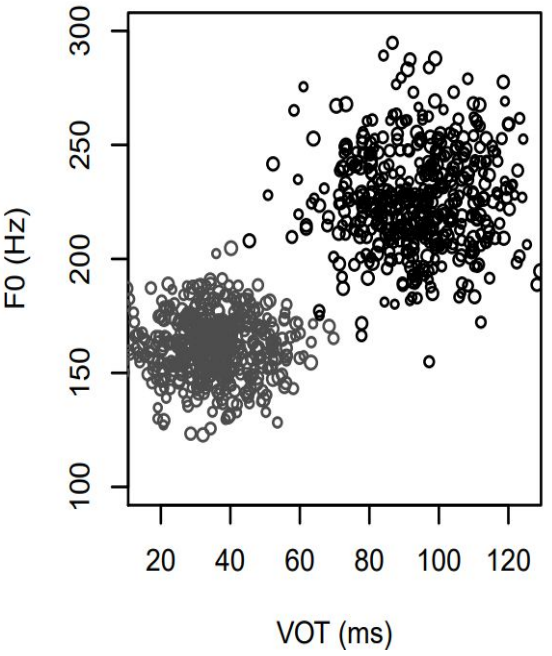
Sound Change with GMM

- "Agents"
  - Have a memory of categorized observations $x_1 \ldots x_n$
  - Each observation has an f0 and VOT value
  - Has a mixture of Gaussians model estimated from memory observations
  - Adds each perceived observation to memory
  - Memory observations decay over time



Sound Change with GMM

- "Agents"
  - Produce: sample from Gaussian mixture model
    - Sample a category from P(c)
    - Sample f0 and VOT values from Gaussian distributions for that category:
      - $N_{VOT}(x \mid \mu_{VOT}, \Sigma)$
      - $N_{f0}(x \mid \mu_{f0}, \Sigma)$

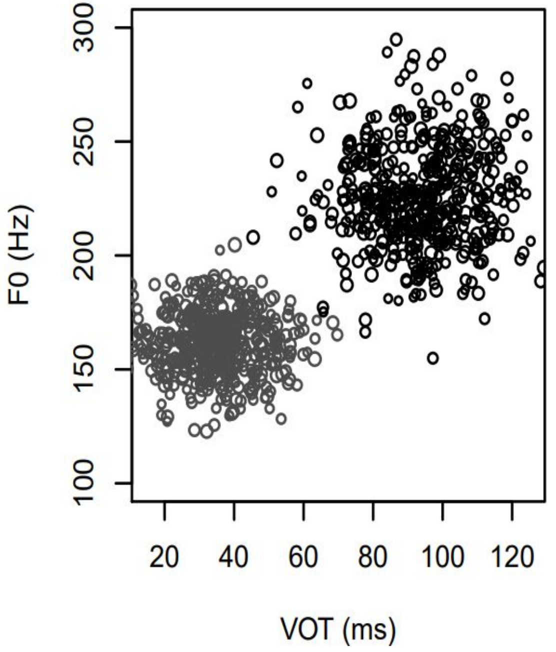


Sound Change with GMM

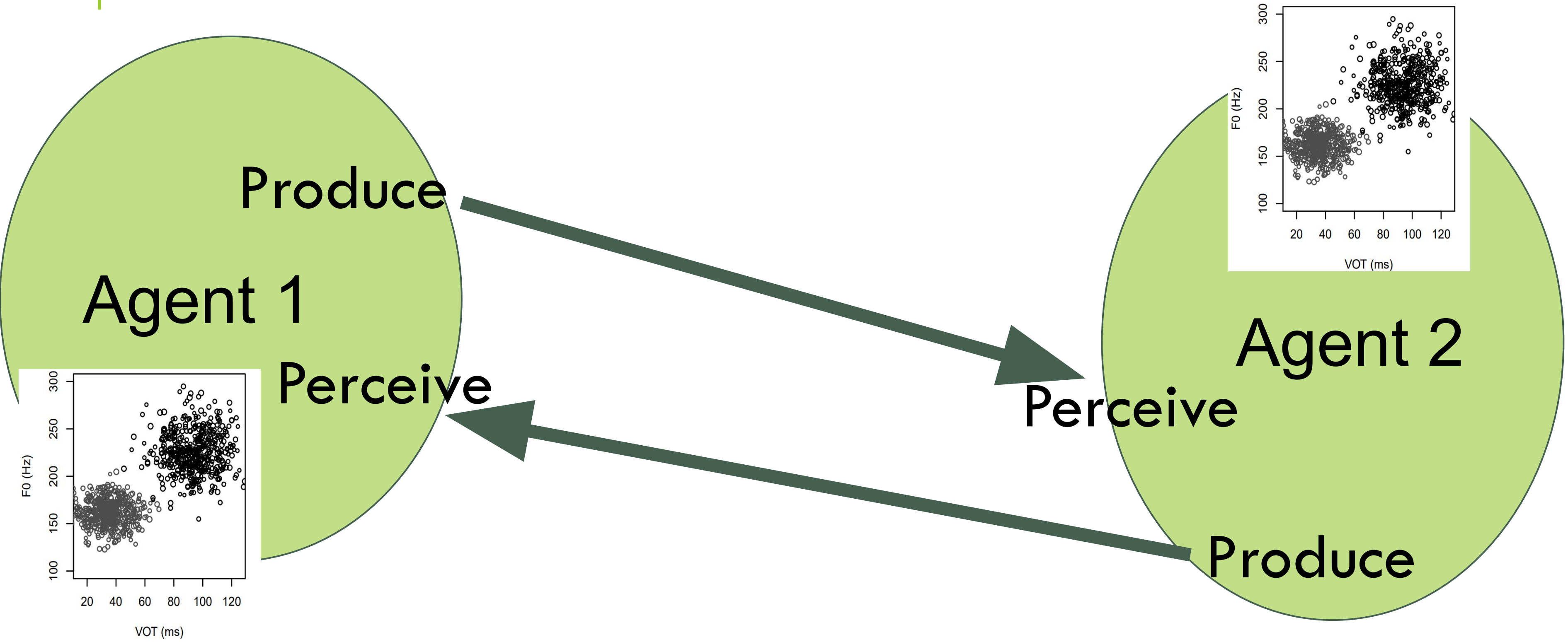
- "Agents"
  - Perceive: receive observation with f0, VOT value
    - Categorize observation:
      - maximize $P(c|x)$
      - $= P(c)P(x|\mu_c,\Sigma)$
      - Ideal observer?

$$P(c|x) = 0.6\ldots$$

    - Add to memory



Sound Change with GMM



What makes distributions move?

Biases in production:

- Alter produced values
- λ: constant values added to one or more dimensions (f0, VOT)
- β: chance of "enhancing" category distinction
  - Move means further apart and reduce variance before sampling

Korean Simulations

- Initialize agent memories to 1960 distribution
- Run perception+production for many iterations
- Manipulate λ and β to eventually produce the 2000s distribution: what kind of bias and enhancement is necessary?

Evaluation

- How to compare simulation distribution to 2000s distribution?
- KL divergence: how much ‘dirt’ to move from one to the other

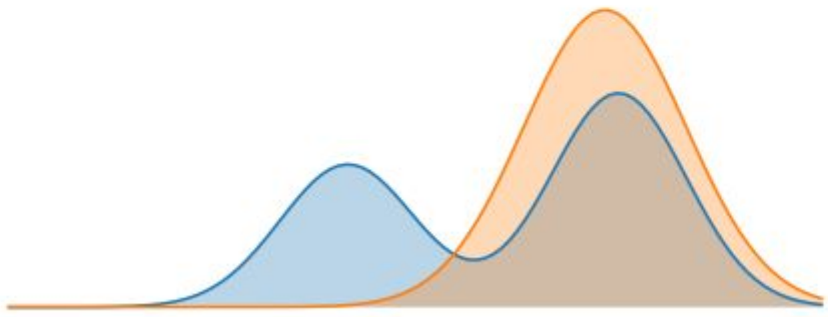


image source:
Dibya Ghosh

Korean Simulations Findings

- Both enhancement and bias influences necessary to produce most 2000s-like distribution

- Other cues involved (spectral tilt, vowel length) - f0 takes over without any bias specifically preferring it