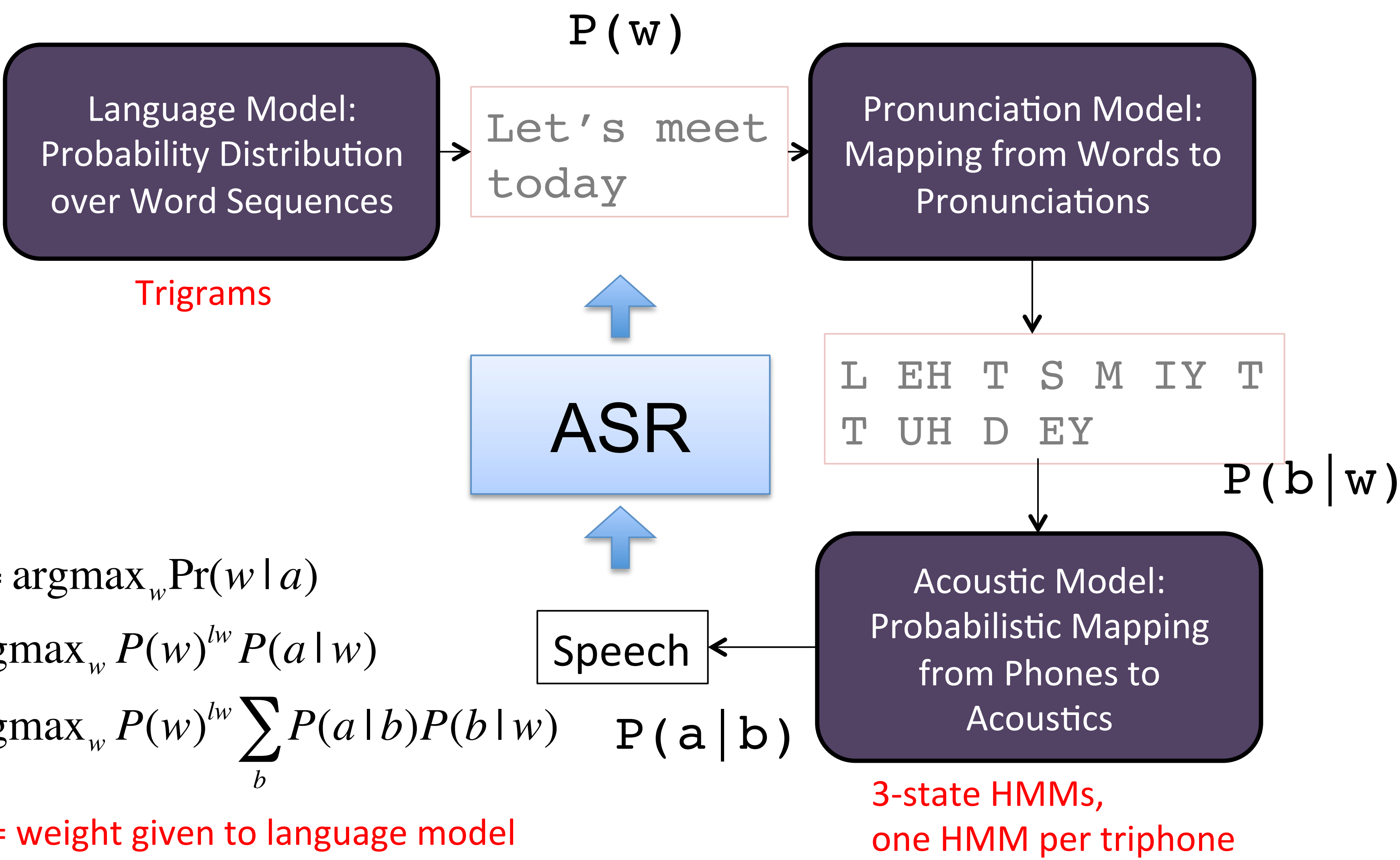
Automatic Speech Recognition

October 30, 2014

Generative Story of Speech



$$w^* = \text{argmax}_w \Pr(w \mid a)$$

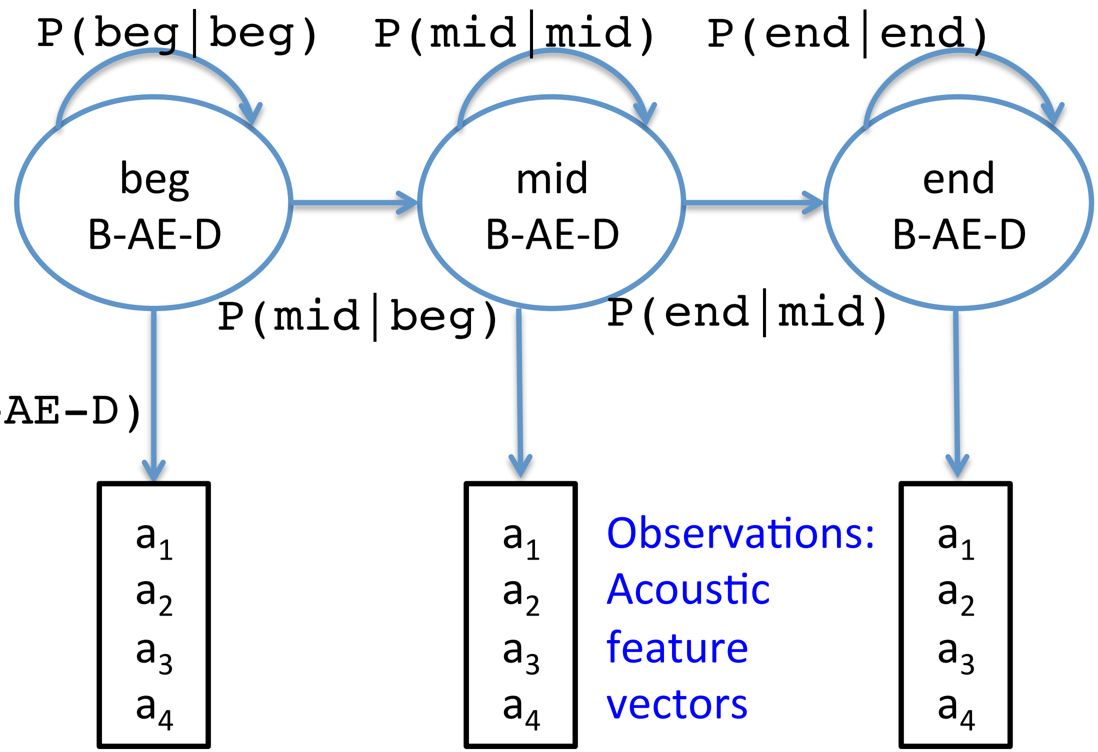
$$= \text{argmax}_w\, P(w)^{lw} P(a \mid w)$$

$$= \text{argmax}_w\, P(w)^{lw} \sum_b P(a \mid b) P(b \mid w)$$

lw = weight given to language model

Acoustic Model

HMM for a single triphone (e.g. AE in the context B-AE-D)



Emission probabilities given by a Gaussian, parametrized by a mean and covariance

Each acoustic feature vector is computed from a 25 ms time slice of speech, every 10 ms (overlapping)

Pronunciation Model

• DREAD	D R EH D
• DREADED	D R EH D IH D
• DREADFUL	D R EH D F AH L
• DREADFULLY	D R EH D F AH L IY
• DREADING	D R EH D IH NG
• DREADNOUGHT	D R EH D N AO T
• DREADS	D R EH D Z
• DREAM	D R IY M
• DREAMED	D R IY M D
• DREAMER	D R IY M ER
• DREAMERS	D R IY M ER Z

Fancier versions:
multiple pronunciations
per word, with probabilities

Language Model

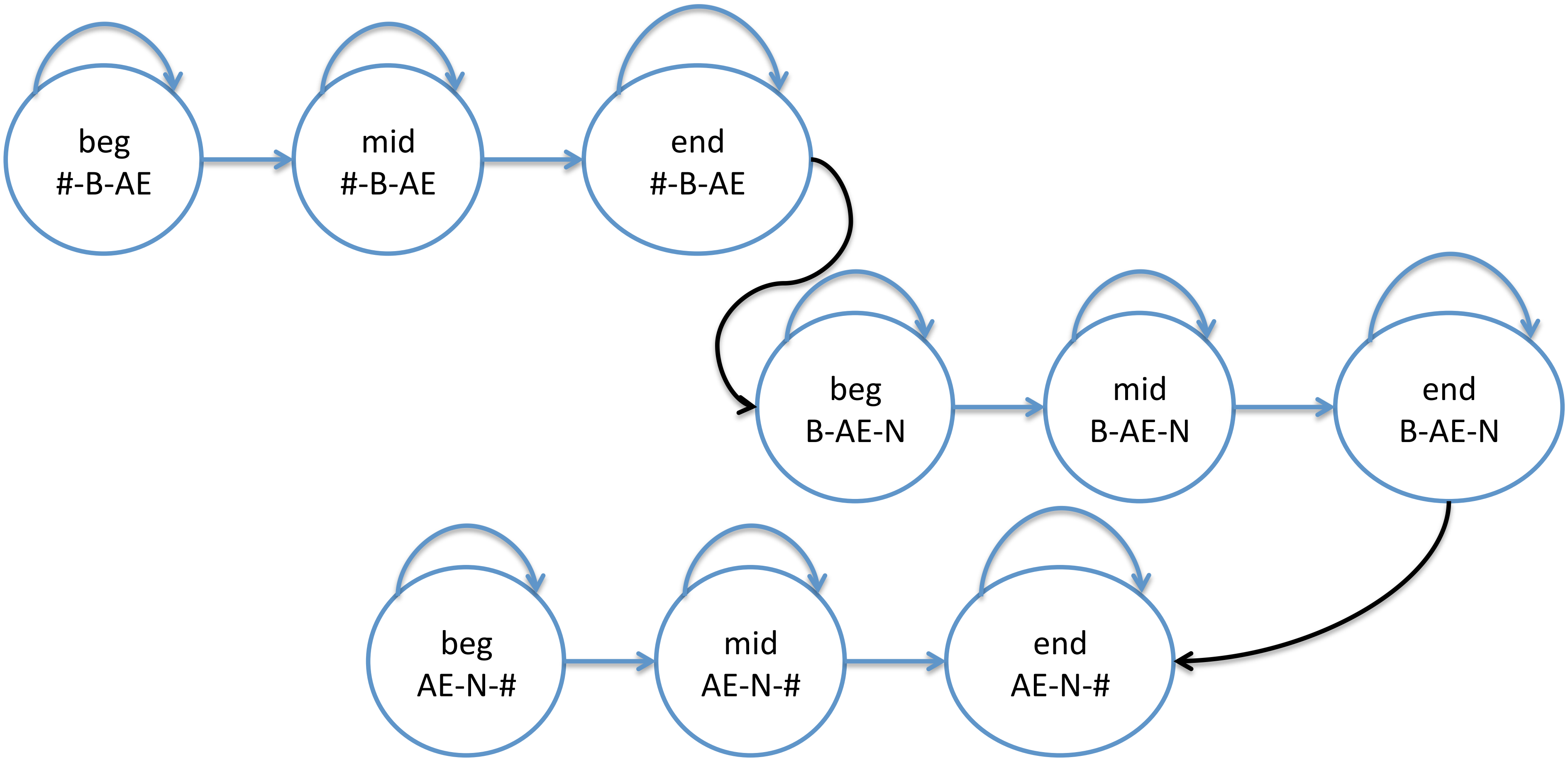
Trigrams over words

```
\data\
ngram 1=64001
ngram 2=9382014
ngram 3=13459879

\1-grams:
-2.2801 <UNK> -0.0796
-4.4211 'CAUSE -1.2221
-4.5633 'EM -0.7278
-5.3040 'N -1.1561
-5.1095 'S -0.5186
-5.2887 'TIL -0.8268
-1.2258 </s> -7.0258
-99.0000 <s> -0.7635
-1.6818 A -1.3696
```

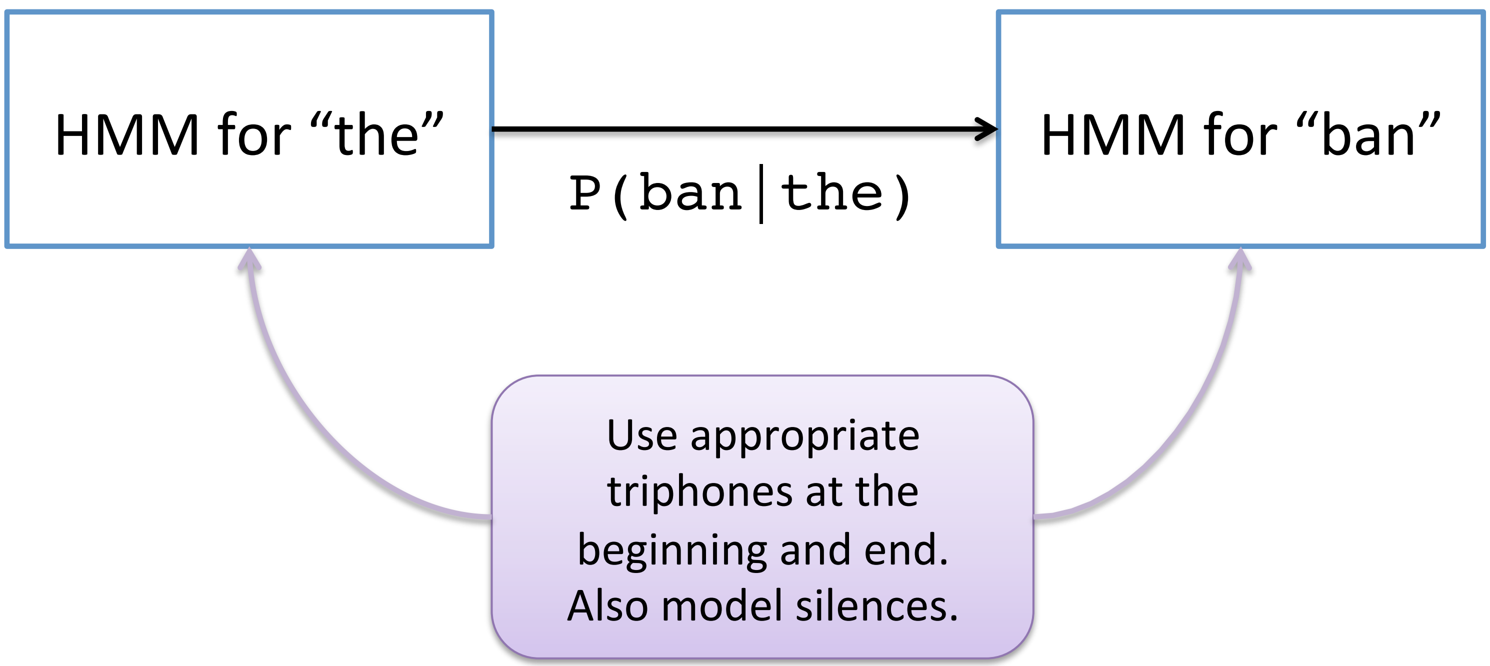
HMM for the word "ban"

Look up pronunciation dictionary and string the triphone HMMs together



HMM for a sentence

String word HMMs together using language model n-gram probabilities



One gigantic HMM

- Triphone HMMs combine to form
- Word HMMs which combine to form
- Sentence HMMs

- Desired result: best sequence of words that produced speech
- Find the best path through the HMM states using Viterbi

One gigantic HMM

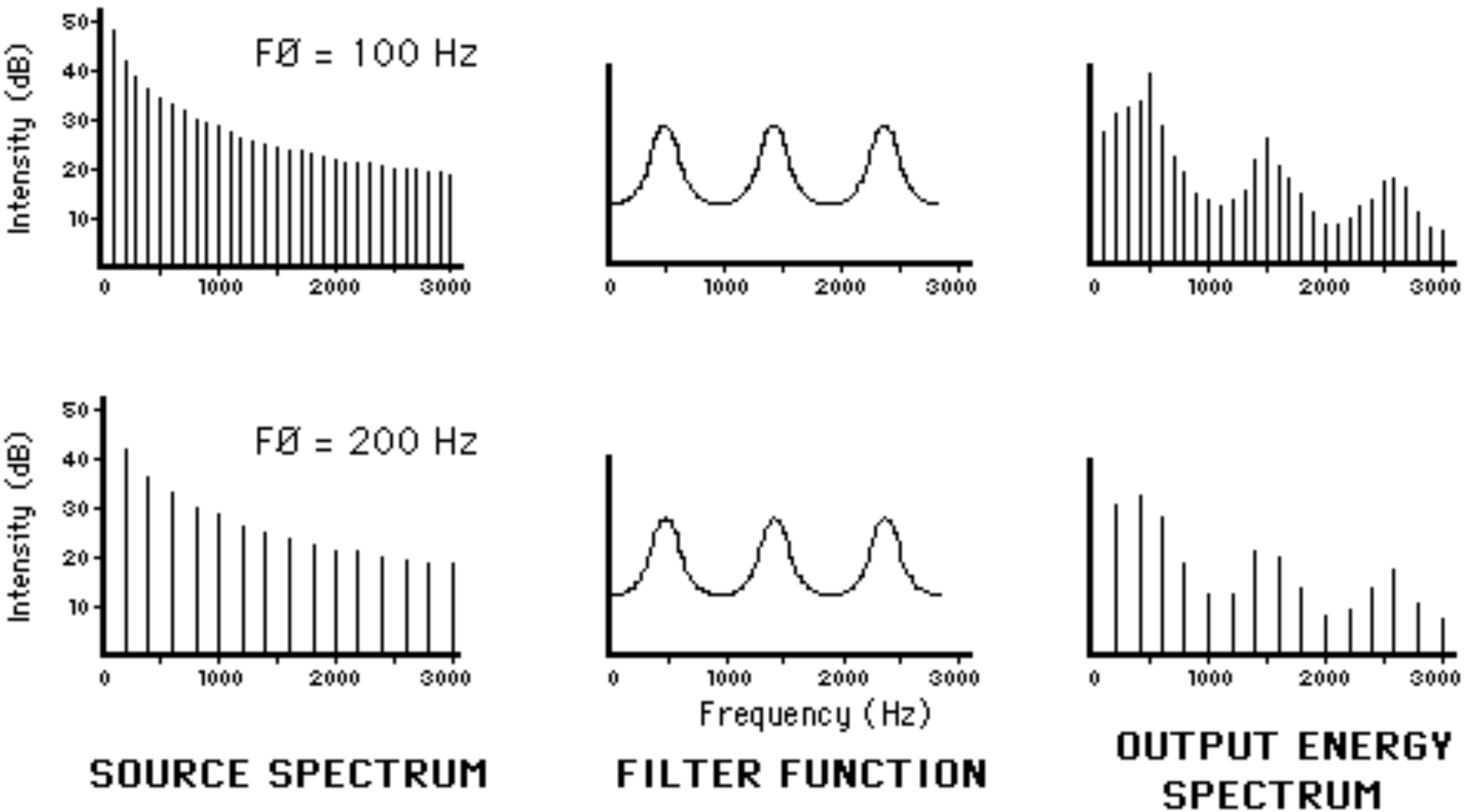
- Desired result of ASR: best sequence of words that produced speech
- Find the best path through the HMM states using Viterbi
  - Can be prohibitively expensive in memory and time!
  - Constrain Viterbi by doing beam search: prune a certain number of low probability states at each time step

Training

- Language Model: Train n-grams from text just as usual (aiming for the appropriate domain)
- Pronunciation Model: usually a dictionary
- Acoustic Model:
  - Train from a large corpus of speech and word-level transcriptions
  - Unknown: transition and emission probabilities of triphone HMMs
  - Learn these probabilities with Expectation Maximization

What are the acoustic features?

- Formants for vowels are a good start
- How do we extract formants?
- Think back to source-filter model: vocal cords produce complex wave, vocal tract shapes filter them according to resonances



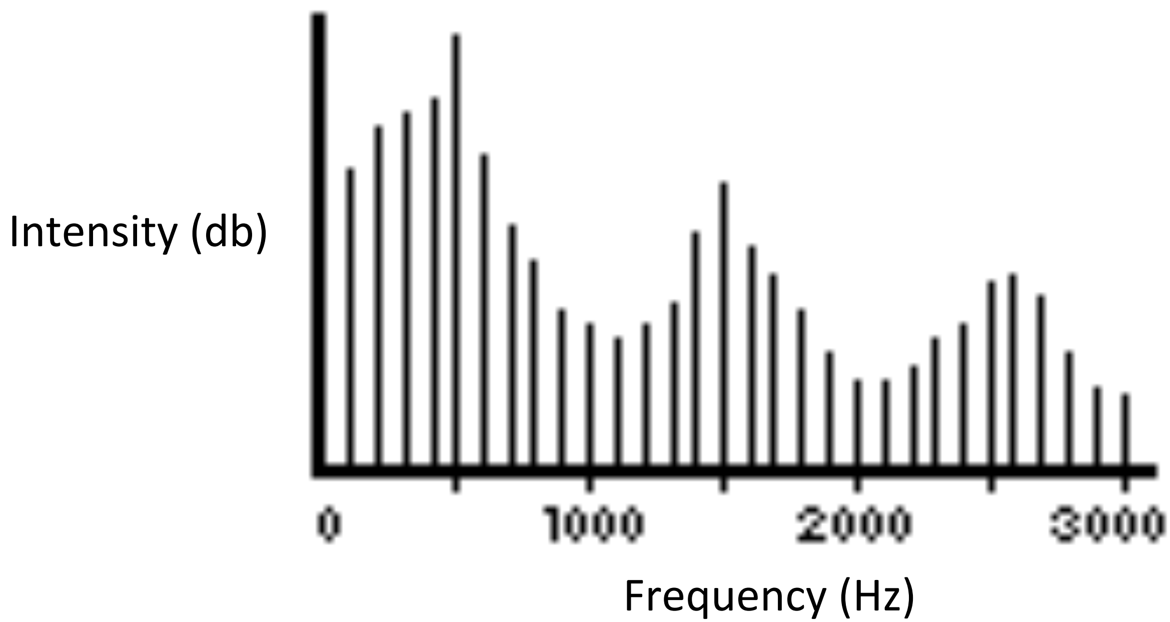
50
40
30
20
10
Intensity (dB)
FØ = 100 Hz
0
1000
2000
3000
FØ = 200 Hz
Frequency (Hz)
SOURCE SPECTRUM
FILTER FUNCTION
OUTPUT ENERGY SPECTRUM

Source interaction with Filter

- Sopranos singing at high frequencies
- Harmonics are spaced too far apart to hit the resonant frequencies

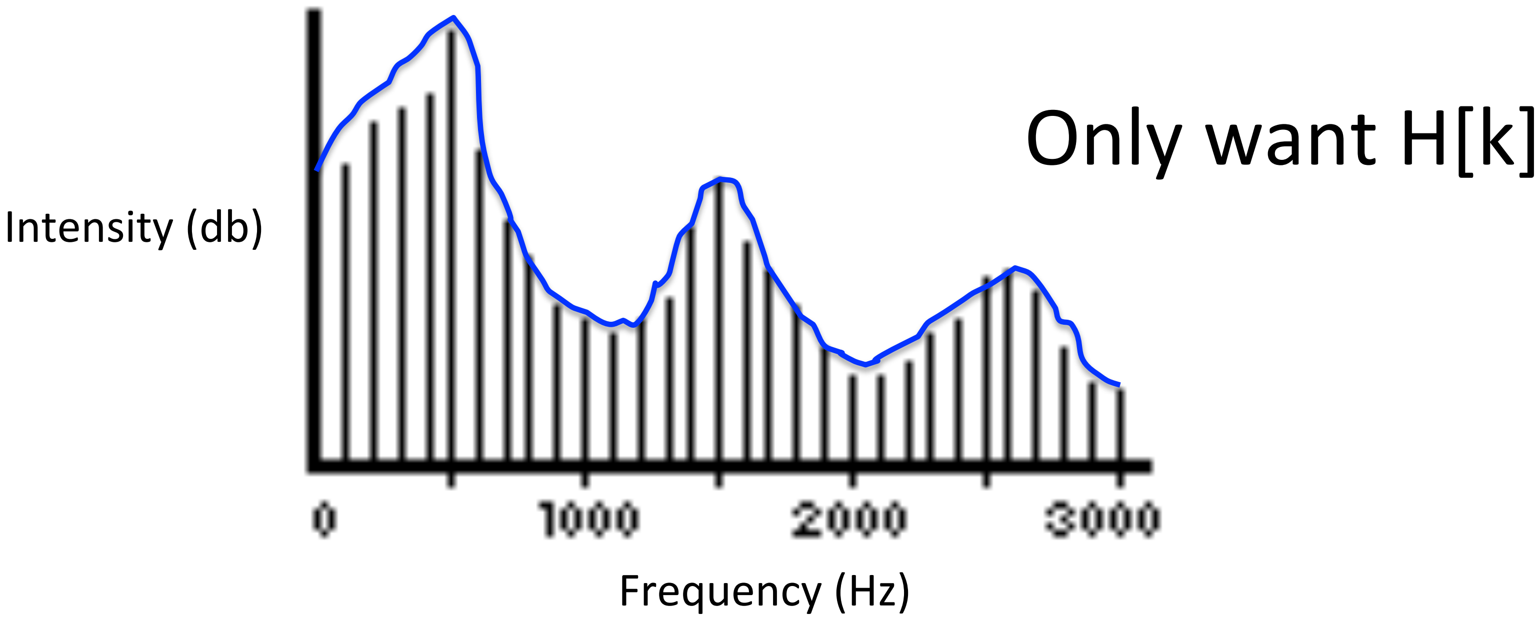
Acoustic Features

- Compute spectrum for a given time window by taking the Fourier transform of speech wave



Acoustic Features

- By source-filter model, this is a convolution of the voice E and the tract H
- Spectrui



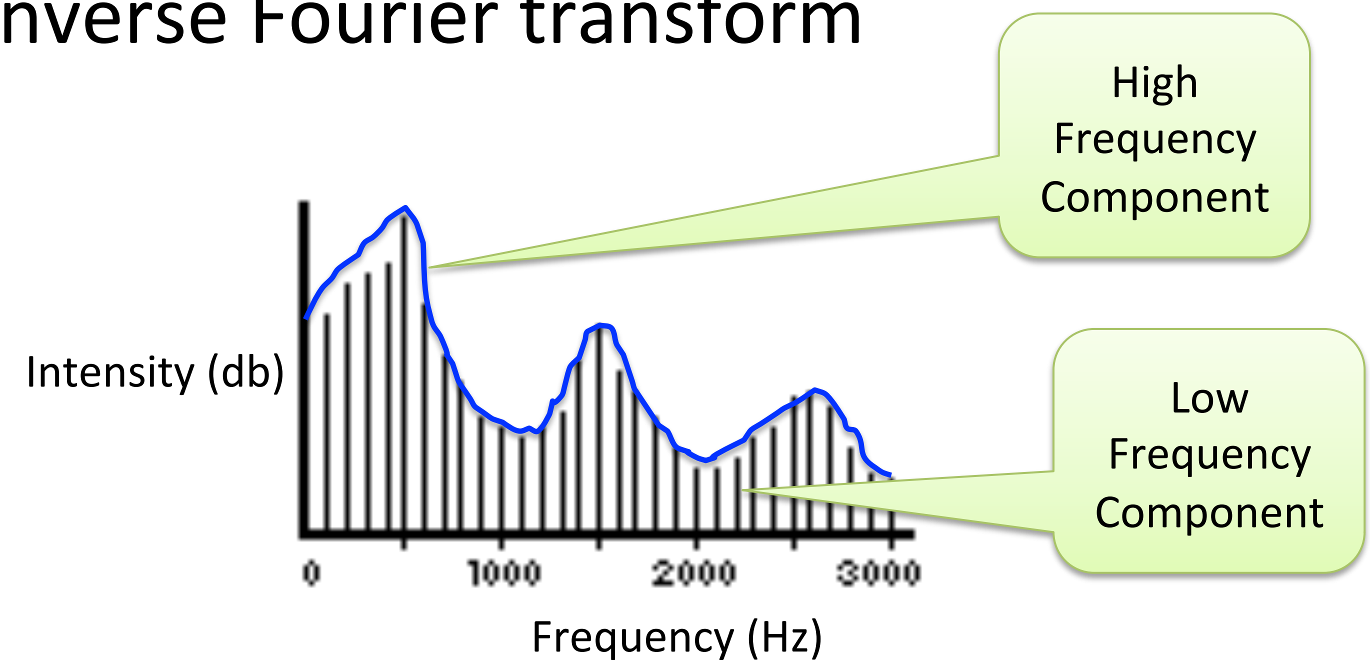
Acoustic Features

- Spectrum X[k] = E[k] * H[k]
- Take log X[k] for two reasons:
  - Intensity variation is more on log scale than linear
  - Allows us to write the convolution as a sum

    log X[k] = log E[k] + log H[k]

Acoustic Features

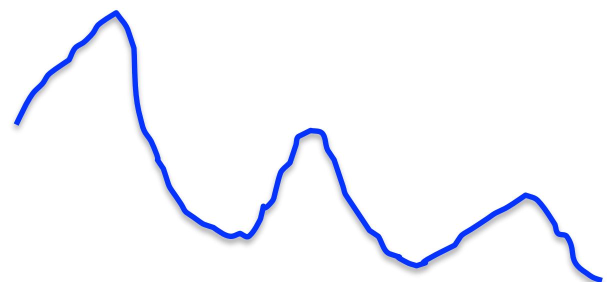
- We see log X[k], and want to compute this separation to get log H[k]
- Play a neat trick: treat log X[k] as wave and take the inverse Fourier transform



Acoustic Features

- Play a neat trick: take the inverse Fourier transform of log X[k]
- Transform ends up separating low and high frequency regions

Low Frequency E[k]



High Frequency H[k]

Acoustic Features

- One last step: human ear does not perceive frequencies linearly
- We are less sensitive to differences in high frequency ranges than in low ranges
- By running perceptual experiments, we come up with the "Mel scale":

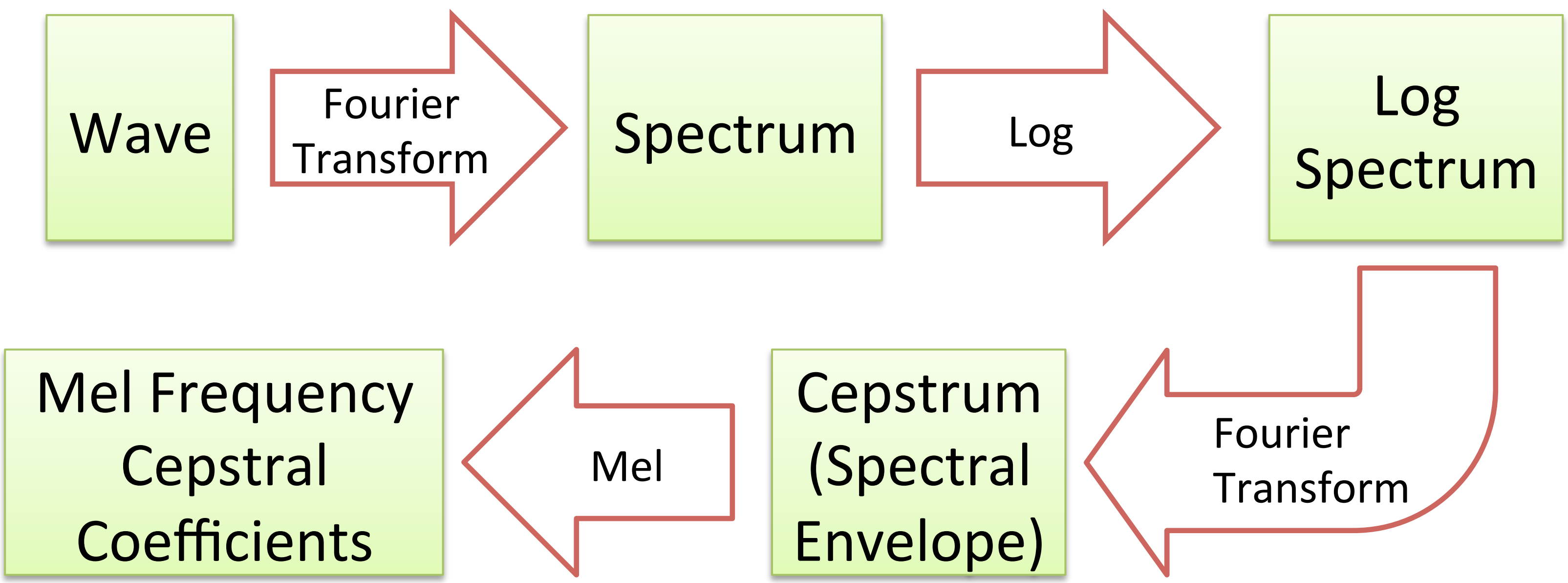
$$f_{mel} = 2595 \log_{10}(1+f/700)$$

Acoustic Features

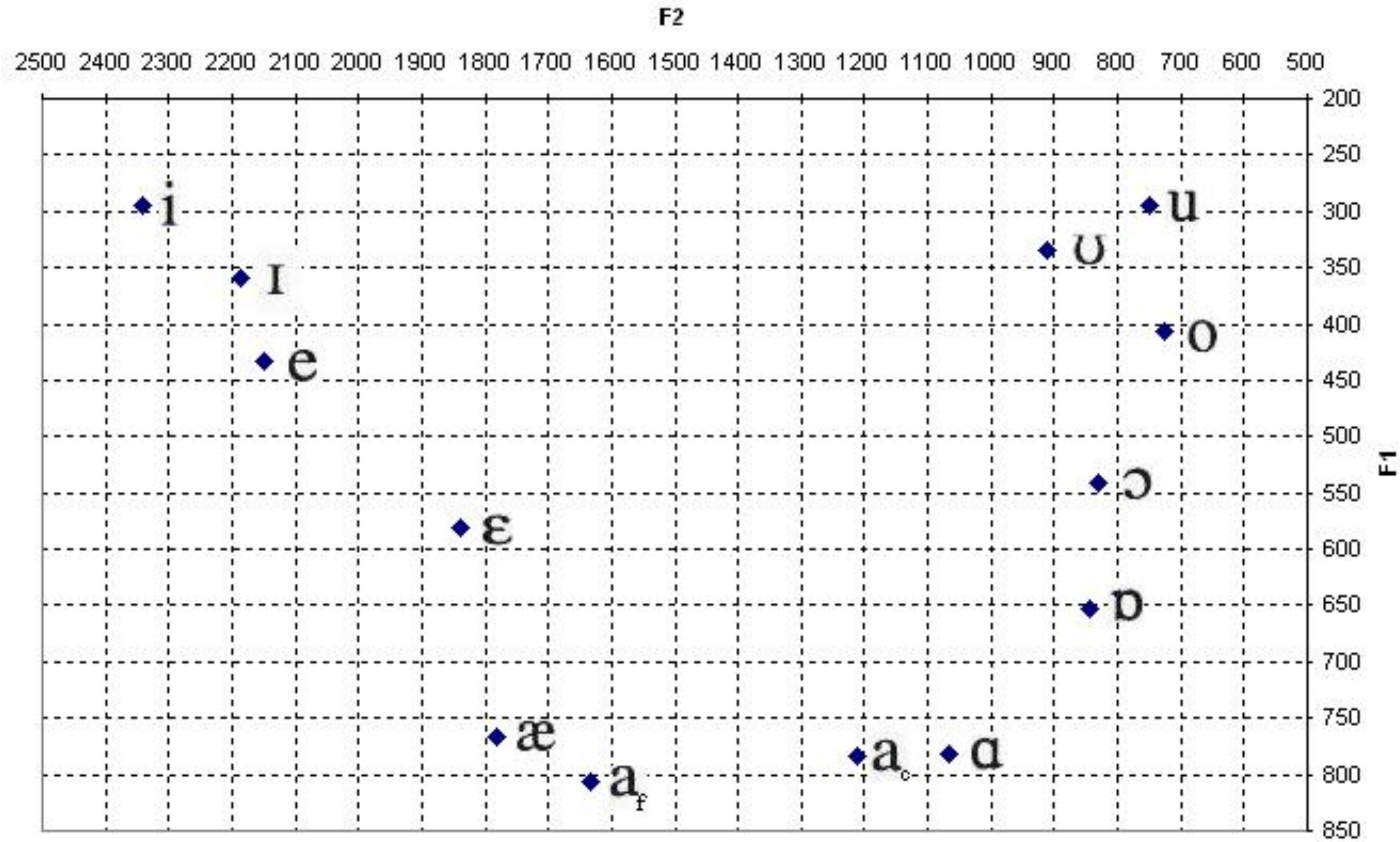
- Map the extracted cepstral peaks onto the Mel scale
- Take the highest 13 peaks
  - More or less correspond to formants
  - 2-3 peaks may be enough for vowels, but we need the remainder for consonants, resistance to noise, etc.

Acoustic Feature Extraction: Recap

- Divide speech signal into 25 ms windows, every 10 ms (overlapping windows)
- At each window:



More Phonetics



More Phonetics

- Vowels = F1 and F2
- Stops: short release
- Fricatives: turbulence
- Nasals: faint formants
- Voice onset time: time between stop release and start of voicing
- Cues also come from formant transitions