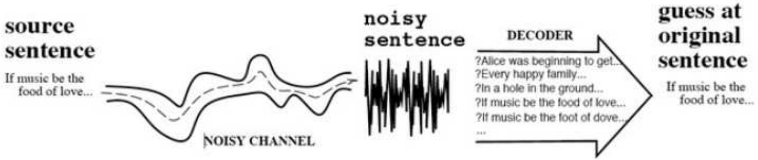
Speech Recognition

[J&M'00 Ch. 9]

Speech Recognition

The **noisy channel** approach: treat the input sound stream as a noisy version of the string of words.



Then the question becomes: what is the most likely sequence out of all sentences in the language L given some acoustic input Q?

Speech Recognition

Enter Bayes' Rule again:

Let $O = o_1 ... o_n$ (consecutive 10/15/20ms slices of the input)
Let $W = w_1 ... w_n$ (string of words)

$$\hat{W} = \text{arg max}_{W \in L} \frac{P(O|W)P(W)}{P(O)}$$

which we can simplify in the usual way because $P(O)$ does not change with W:

$$\hat{W} \approx \text{arg max}_{W \in L} P(O|W)P(W)$$

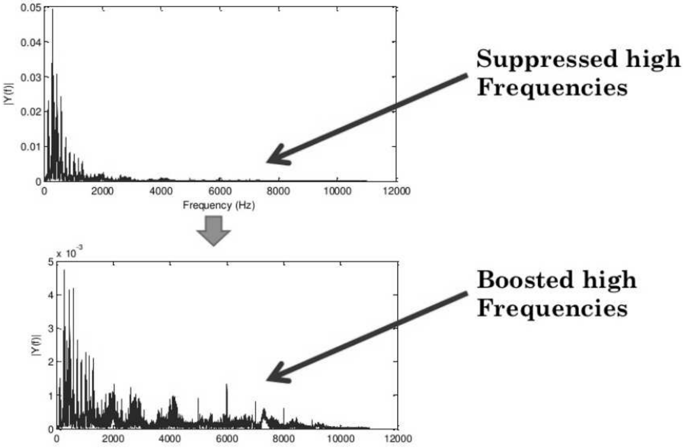
- $P(W)$ (the prior) is easy to estimate: n-gram language model.
- $P(O|W)$ (the likelihood) is a bit harder to estimate...

Speech Recognition

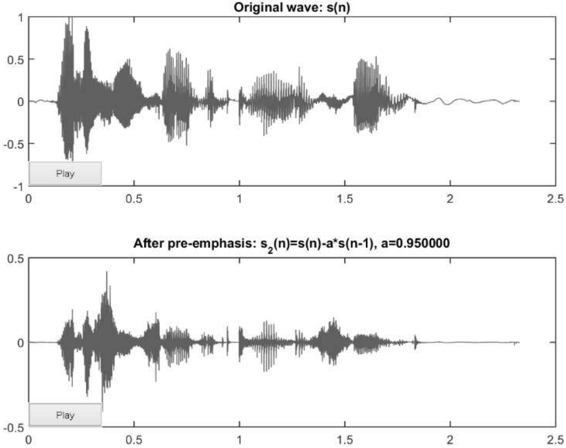
To estimate $P(O|W)$ we need:

- *Feature extraction*
  (sample the waveform and extract spectral features)
- *Phone recognition*
  (using classifiers, usually Gaussian Mixture models)
- *Decoding*
  (word pronunciation HMM)

In many cases, there is more energy at the lower frequencies than at the higher frequencies. In order to help the the model, the latter is boosted. This is called **pre-emphasis**



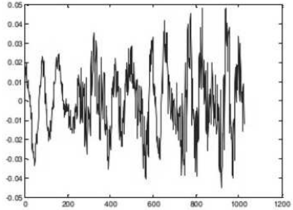
Pre-emphasis

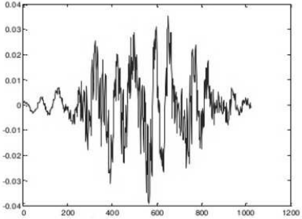


This is a first-order high-pass filter. In the time domain, with input $x[n]$ and $0.9\alpha 0.98$, with equation $y[n] = x[n] - \alpha x[n-1]$

Windowing

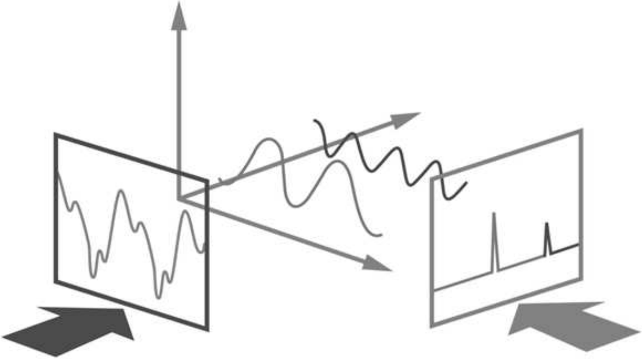
The Hamming window shrinks the values of the signal toward zero at the window boundaries, avoiding discontinuities.





$$w[n] = \begin{cases} 0.54 - 0.46cos(\frac{2\pi n}{L}) & 0 \leq n \leq L-1 \\ 0 & \text{otherwise} \end{cases}$$

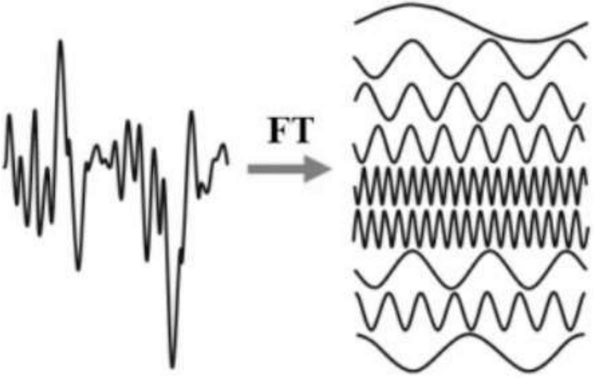
Discrete Fourier Transform (usually the CooleyTukey algorithm)



(for details see this)

Speech Recognition

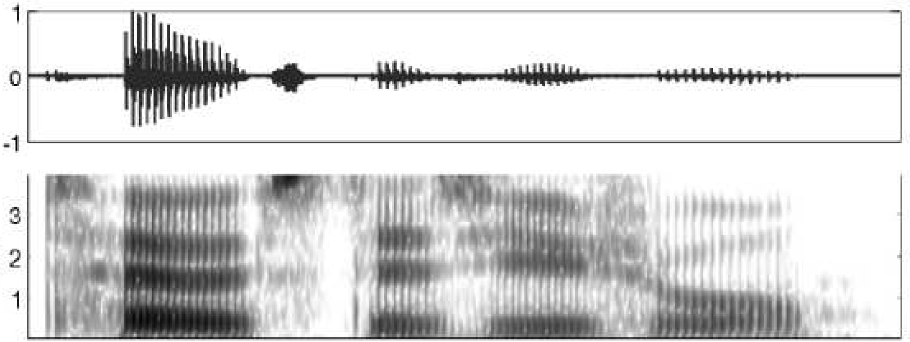
Feature extraction



See this and this and this.

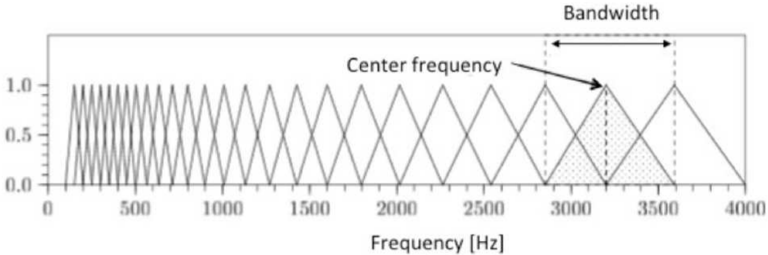
Speech Recognition

Feature extraction



We now have information about how much energy is there in each frequency band, at each time step.

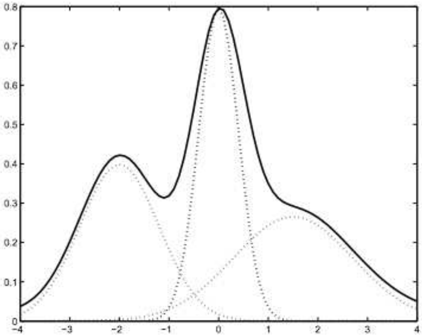
Mel Filter Bank and Log: human hearing is less sensitive to frequencies higher than 1000Hz, and speech recognition performance improves it this aspect is modeled. A mel is a unit of pitch, and the mel scale is linear 1000Hz and logarithmic above 1000Hz.



Next, take the logs of the powers at each of the mel frequencies, and take the discrete cosine transform of the list of mel powers.

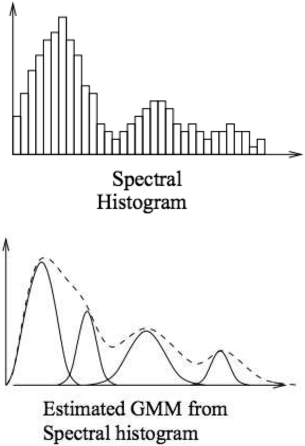
A classifier estimates the probability of a particular HMM state j generating the respective phoneme.

Each phoneme is identified in terms of the mix of several gaussians:



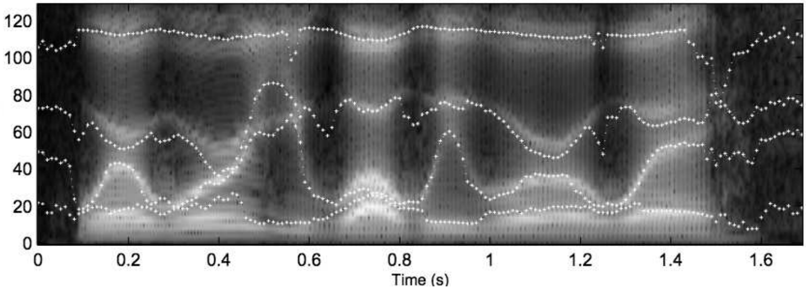
At each time frame we have a sequence of probability vectors containing the likelihoods that each phone unit generated the acoustic feature vector observation at that time.

Gaussian Mixture Model: unsupervised clustering



Speech Recognition

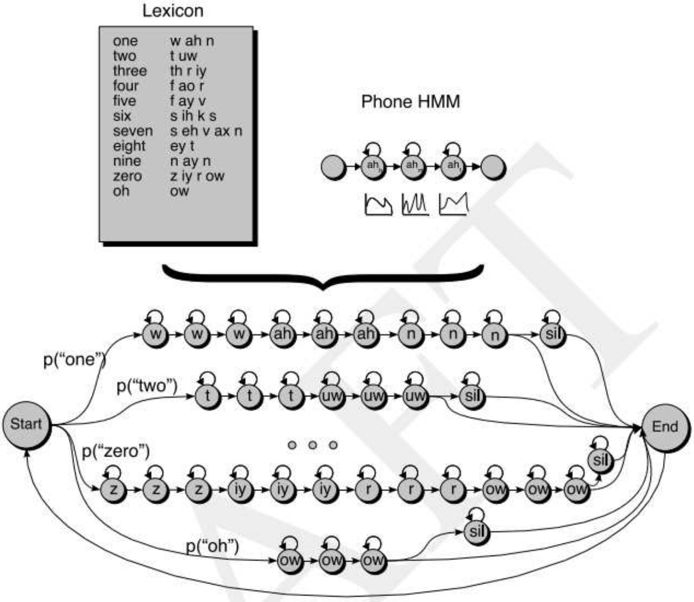
Phone recognition



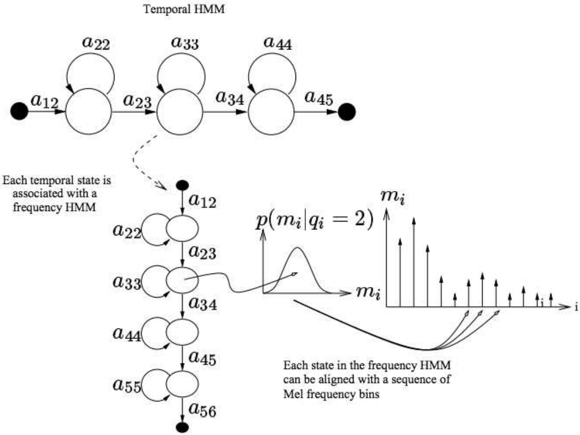
A HMM with:

- A set of states Q corresponding to phones
- A transition probability matrix (the σs representing the phone sequences, using a pronunciation dictionary, e.g. CELEX2)
- A set of observation likelihoods (the τs representing the probability of a feature vector being generated at a given phone state)

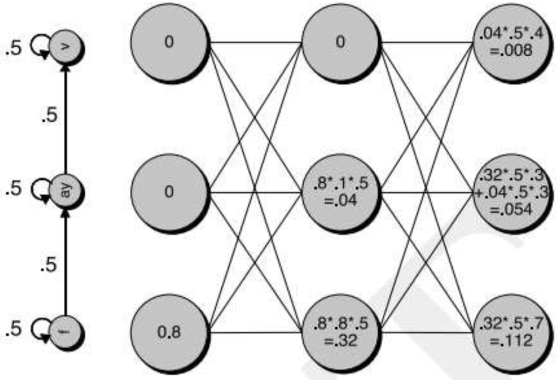
Speech Recognition



Speech Recognition



First 3 time-steps of the forward trellis computation for the word *five*:



(J&M draft page 322, printed version 320)