Chapter 2
Automated Speech Recognition

a) Problem Classification
b) Data Analysis
c) Processing Units
d) Evaluation
e) Applications

Literature

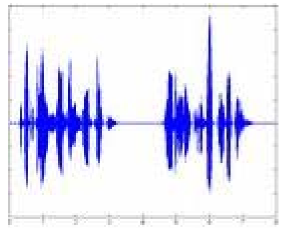
- Rabiner, Juang, Fundamental of Speech Recognition, Prentice Hall, ISBN 0-13-015157-2
- http://www.cs.ubc.ca/~murphyk/Bayes/rabiner.pdf
  - A tutorial on hidden Markov models and selected applications in speech recognition
- D. Balentine, D.P. Morgan, How to Build a Speech Recognition Application, Enterprise Integration Group Inc., ISBN 0-9671278-1-5.
- Skript “Automatische Sprachdienste”, Urs-Viktor Marti
- R.O. Duda, P. E. Hart, D.G. Stork, Pattern Classification, Wiley Verlag, ISBN 0-471-05669-3.

Brief History

- 1946 Sound spectrograph
  - Made analysis easier
  - No time consuming Fourier analysis needed for analysis of sound
- 1948 Theory for human speech developed
- 1975 First word recognizer
- 1980s novel methods developed (HMMs)
- 1986 Large databases acquired
- 2000 VoiceXML defined by W3C

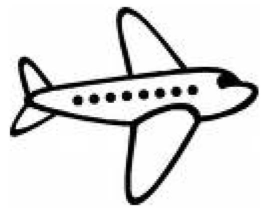
Linguistic Terms

- Acoustics: analysis from the physic‘s point of view
- Phonetics: how do humans produce speech
- Phonology: divide speech into basic units (phonemes)
- Morphology: compose words from morphemes
- Syntax: define allowed word sequences
- Semantic: extract information
- Pragmatics: consider the context

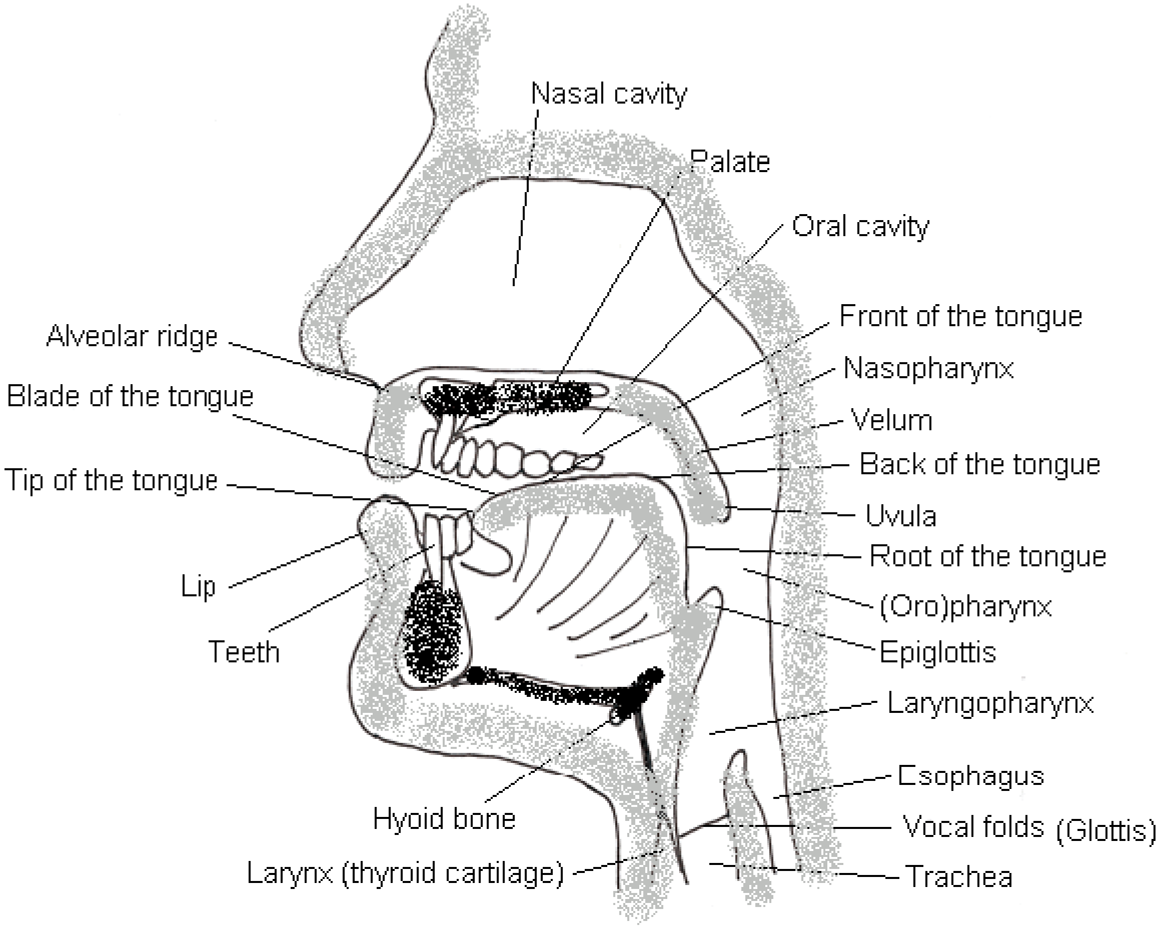


ˈeərəʊpleɪn

airplane

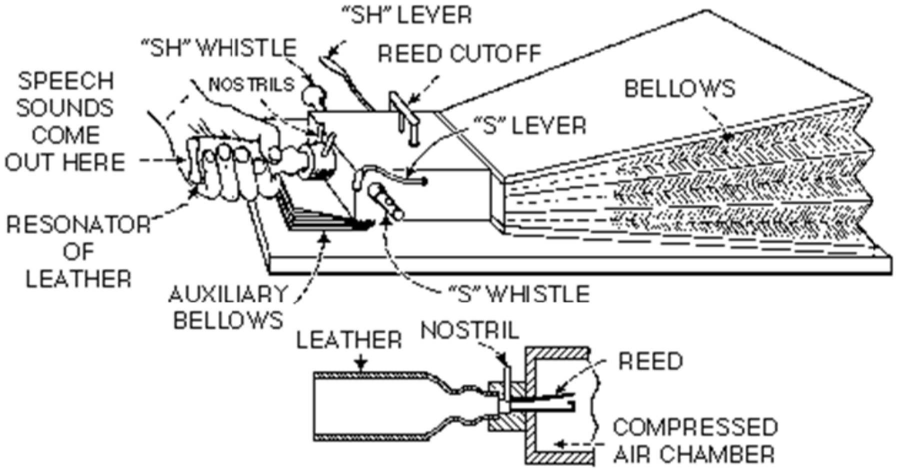


How do WE Speak?



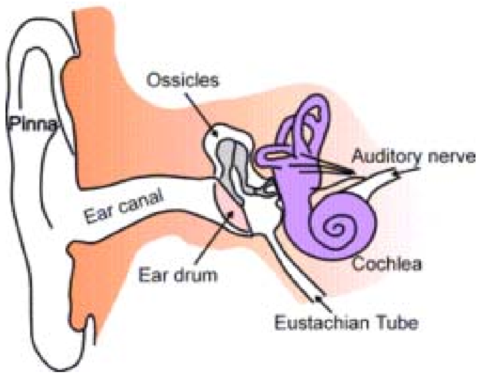
- Vocalization
- Articulation
  - Active
  - Passive
- Resonance

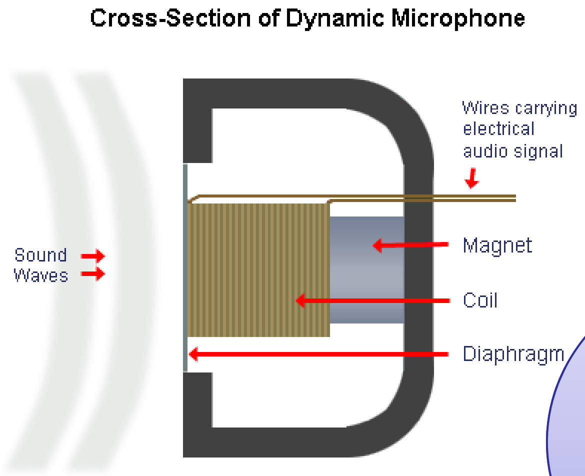
- Wolfgang Kempelens speaking machine 1791



Voder 1939

How do WE Hear?



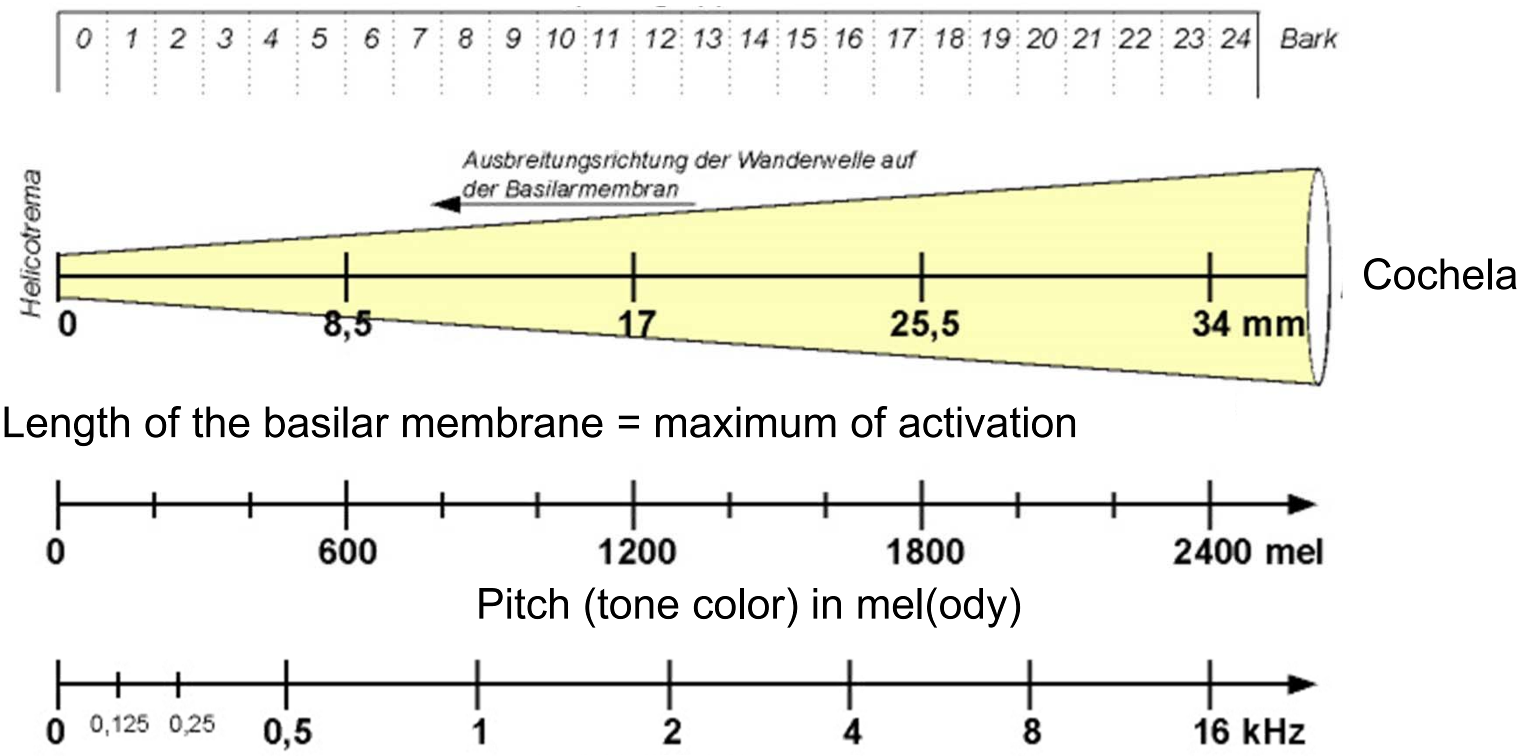


- Frequency
  - In Cochlea there are basilar membranes
  - Nerves are activated for different frequencies
- Pitch
  - Subjective
- Loudness
  - Number of impulses per frequency band
  - Measured in dB
    - Logarithmic to pressure (+3dB=2xPa)

Is this all?
Listen!

Critical Bands and Frequencies

Critical bands



Frequency f

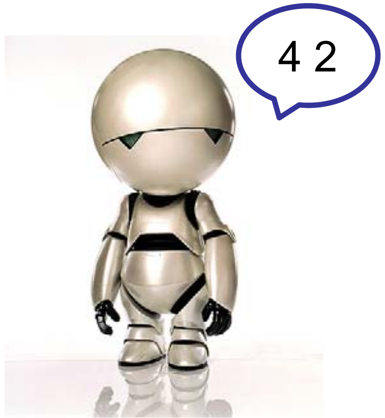
- Logarithmic scale (except until 500 Hz)

Problems in ASR

- Not everyone speaks the same language
- Speech is ambiguous
- Speech evolves
- Word segmentation is difficult
- Environmental noise
- Different input devices
- Multi-lingual content

Classify the Problem

- Spoken data
- Sentences vs. words
  - Less restricted but harder to implement
- One person vs. multiple persons (speaker dependent)
  - Male/female speaker
  - Consider emotions?
- Speaker identification and verification
- Number of classes (words)



- Example: Speaker-dependent isolated digit recognition
  - Fully accessible, deterministic, episodic, static, continuous

Analyze the Data

- Decode the audio format
  - Discrete representation of continuous signal
- Improve signal
  - Remove echo and noise
  - Filter out high range sounds
  - Normalize sound level
  - Normalize sampling rate

Divide the Problem into Sub-Tasks

Speech signal

Feature extraction

Features

Classification

Alternates

Post-processing

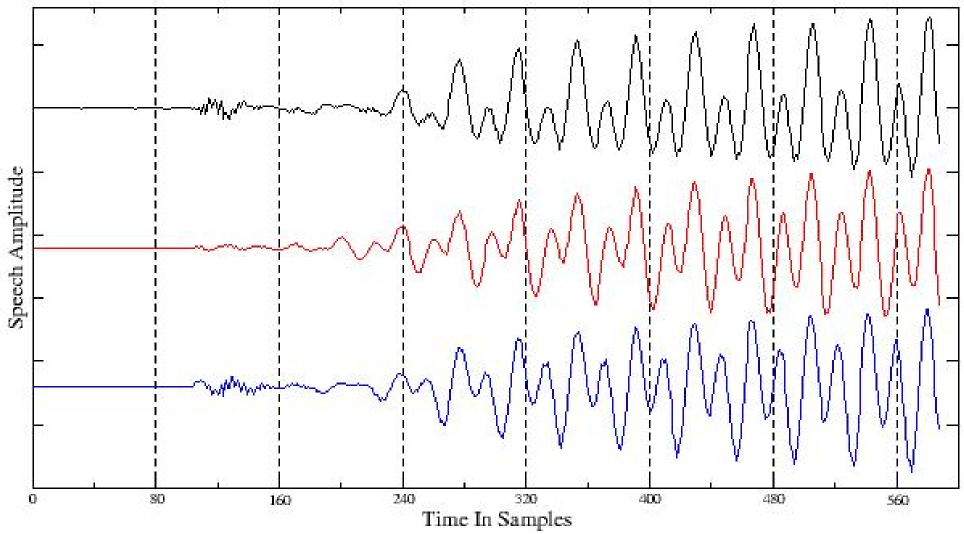
Text

- Wave-form
- Transform raw data into real numbers
- Apply ANN, kNN, GMM, HMM, …
- Use language information

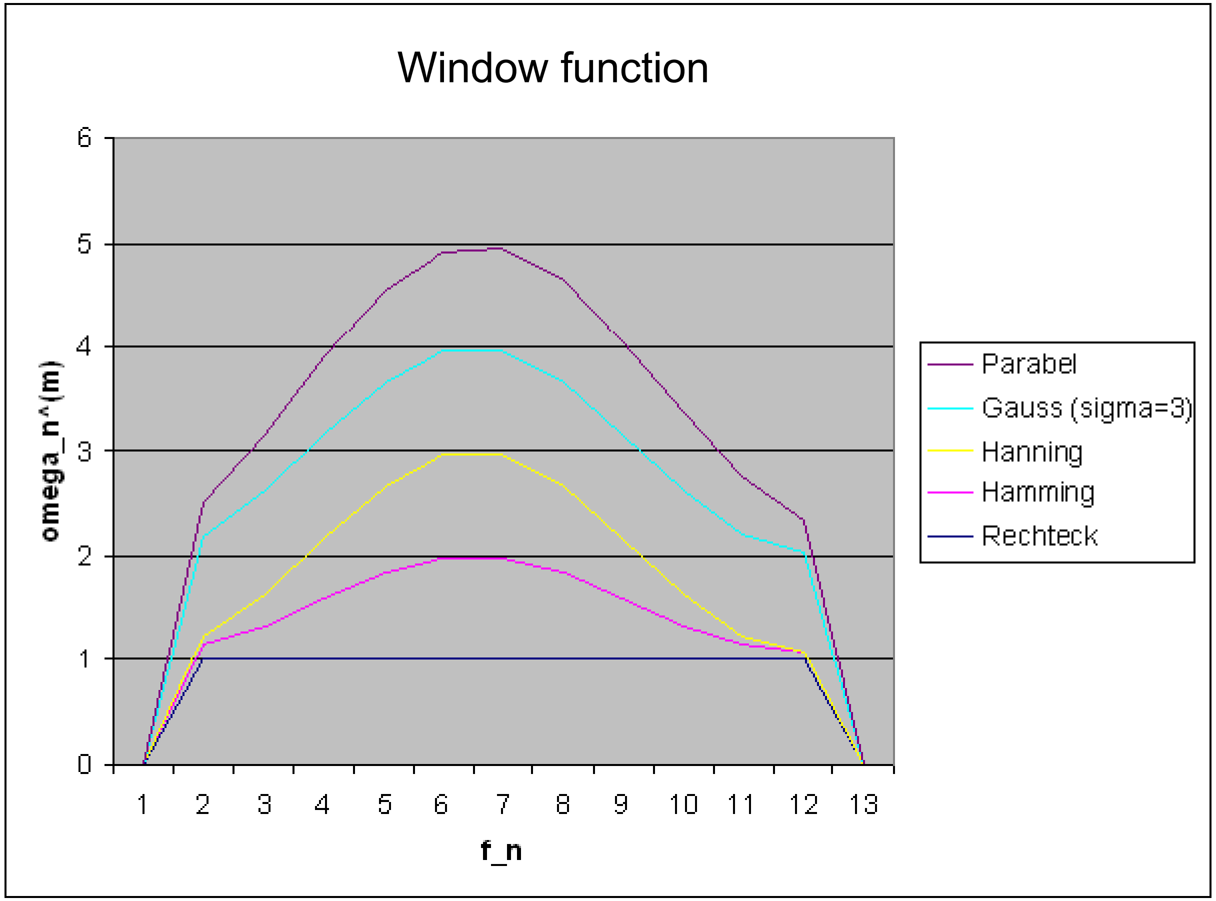
UNI FR

Short-Time Speech Measurements

- Speech signal changes over time
- Spectral properties change
- Idea: divide signal into several segments
  - Assumption signal is in a steady state in small time frames (windows)
  - Generally agreed, but not 100% correct



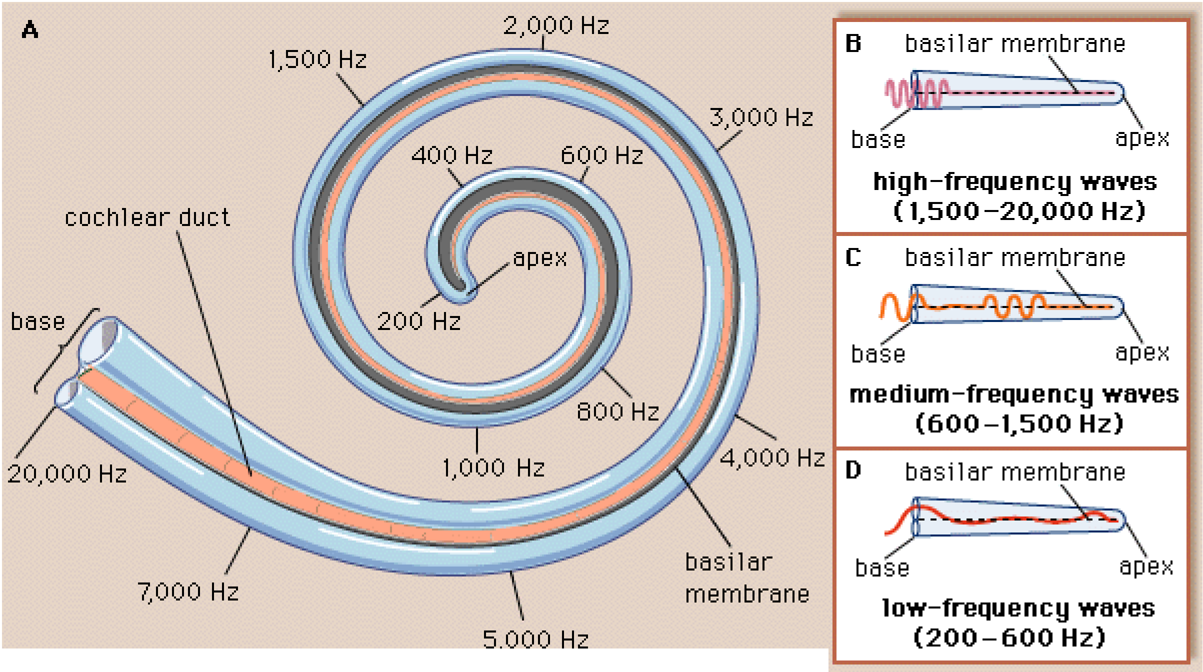
Window functions



- Often, Hamming-window is used $\omega_n^N = 0.5 - 0.5\cos\left(\frac{2\pi n}{N-1}\right)$

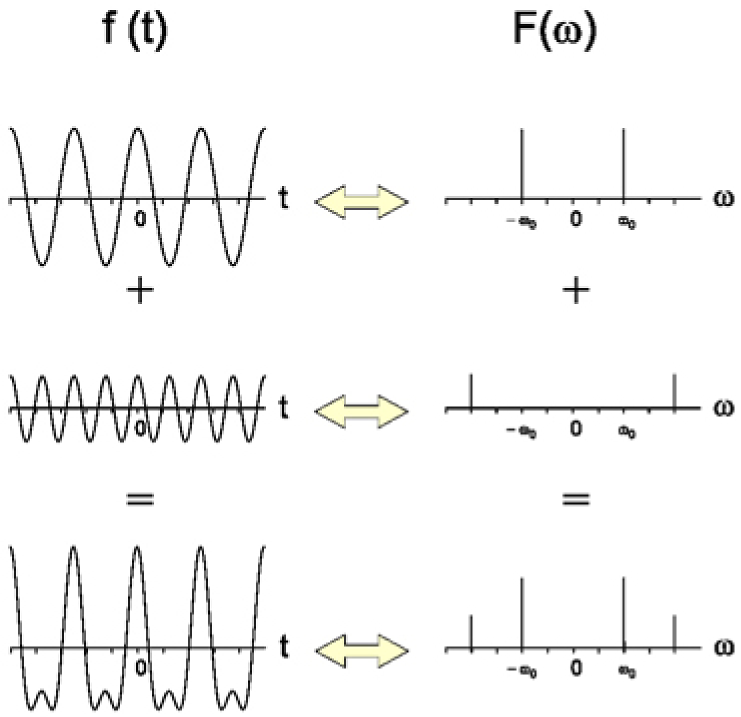
Extract the Spectrum

- It is wise to transform the wave-signal into frequencies
  - Similar to the human basilar membrane
  - Apply Fourier transform (FFT)



Fourier Transformation

Important



Remember: $\omega = 2\pi f$ and i is complex

$$f(t): R \to C$$

- Extracts the frequencies from the spoken signal

$$F_\omega = F(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt$$

- Inverse: Just for information

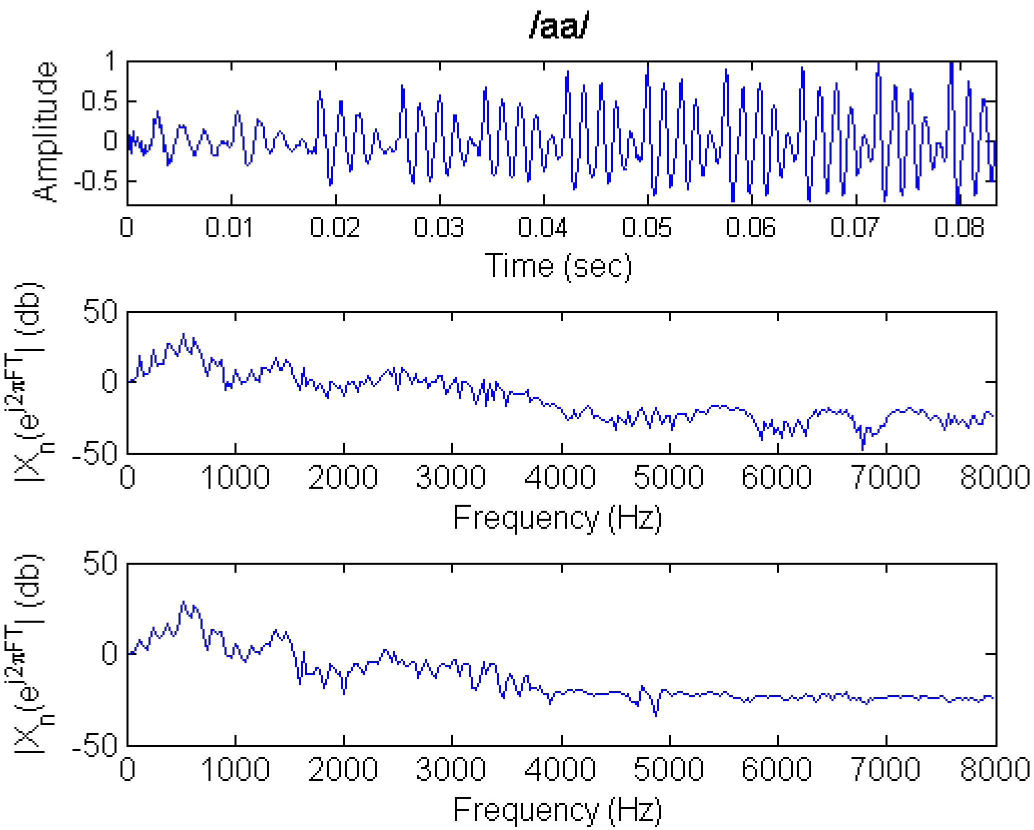
$$F_t^{-1} = f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(\omega) e^{-i\omega t} dw$$

- Discrete:

$$\hat{x}_k = T \sum_{n=-M}^{N-M-1} x_n e^{-i\omega_k t_n}$$

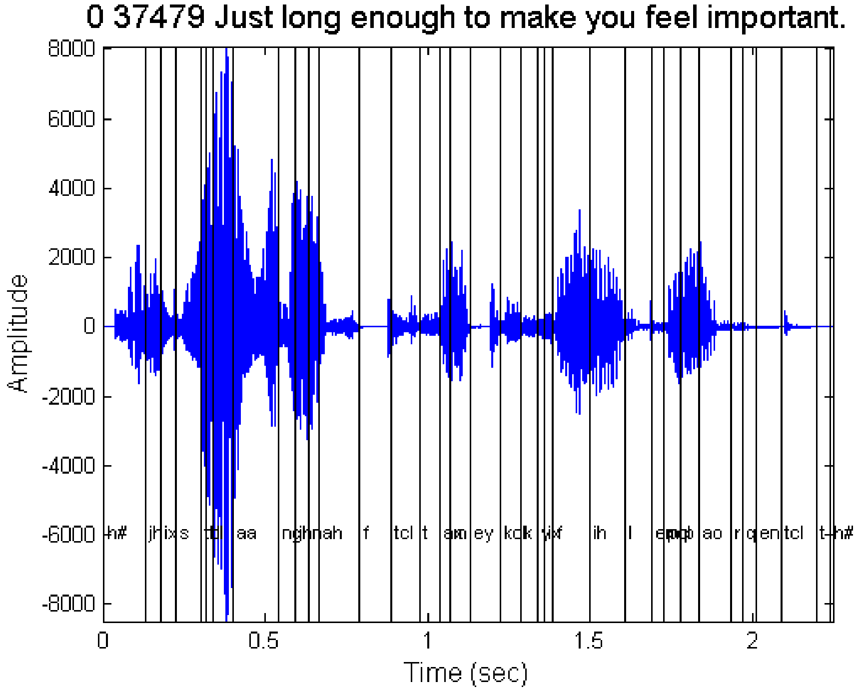
  - T positive, M – shift, $t_n = n*T$

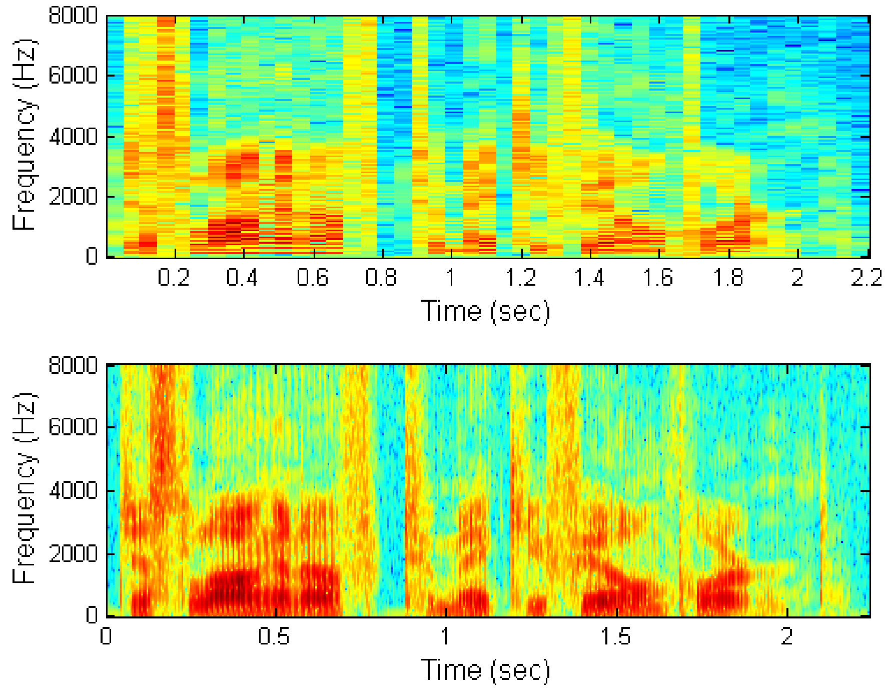
Example for a Short Phrase



- Signal
- FFT
- FFT with Hamming

Define a frame width

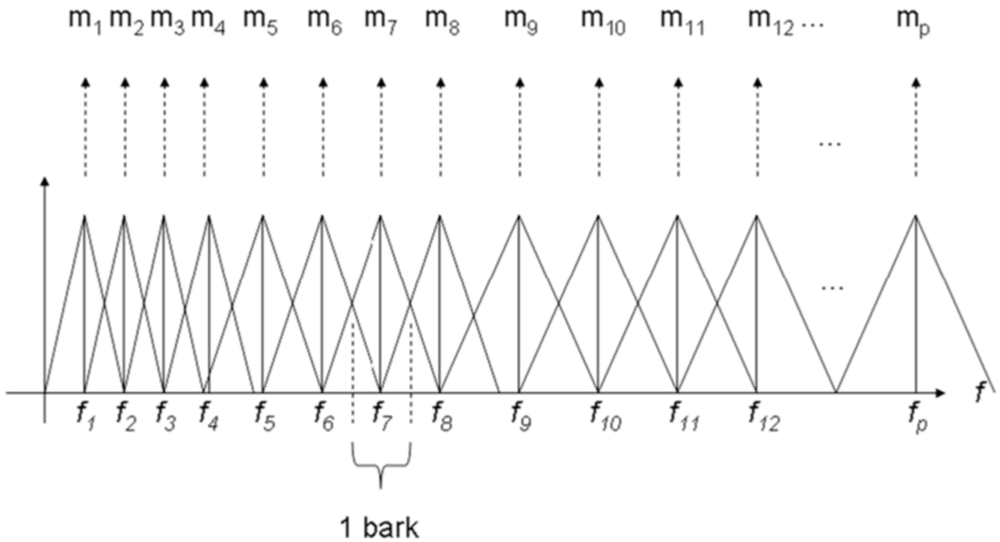




- Wideband or Narrowband (0.05s vs. 0.005s)
  - Better frequency resolution vs. better time resolution
  - Usual: frame width of 25ms, step size of 10ms

Filter out the Useful Frequencies

- Humans group frequency groups in one neuron
- Idea: apply a bank of frequency bands (triangular)
  - Usually done at (Hz): 150, 200, 250, 300, 350, 400, 450, 500, then exponentially until 4,000 (6 bands)



Features

- Often used: Mel-frequency cepstral coefficients (MFCCs)
  - Also used for mp3-compression
- Speech seen as a periodic activation signal of chords e_n is linearly filtered by mouth, tongue, etc. h_n Important
- Mathematically seen as folding $f_n = e_n * h_n$
- For ASR the filter is interesting (and its change)
  - It is separated from the activation signal
  - Cepstral coefficients are commonly calculated using the discrete cosine transformation on the logarithms

Just for information

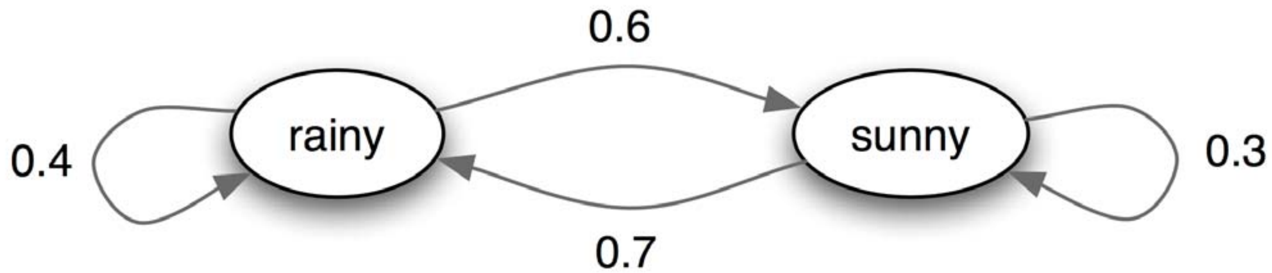
$$c_q^{(m)} = \sum_{k=1}^{K} \log e_k^{(m)} \cos \frac{\pi q(2k+1)}{2K}, q = 1,2,\ldots, N/2$$

  - c_0 is the sound level

MFCC summary

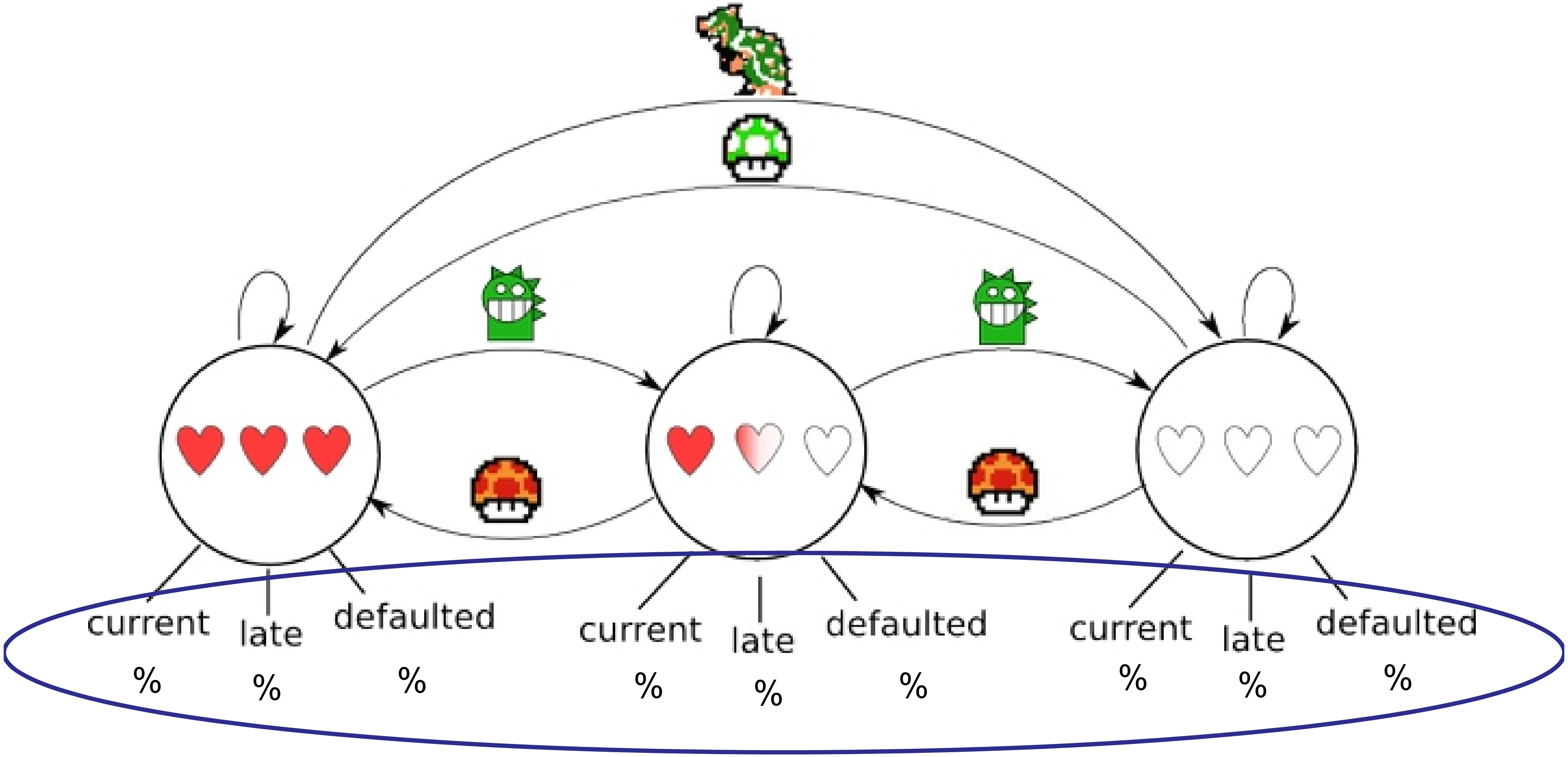
1. Divide the signal into overlapping windows (Hamming)
2. Apply discrete Fourier transformation
3. Generate bank of frequency bands
4. Use the logarithm (transform mult. into add.)
5. Reduce number of frequency bands
6. Decorrelate using the discrete cosine transformation

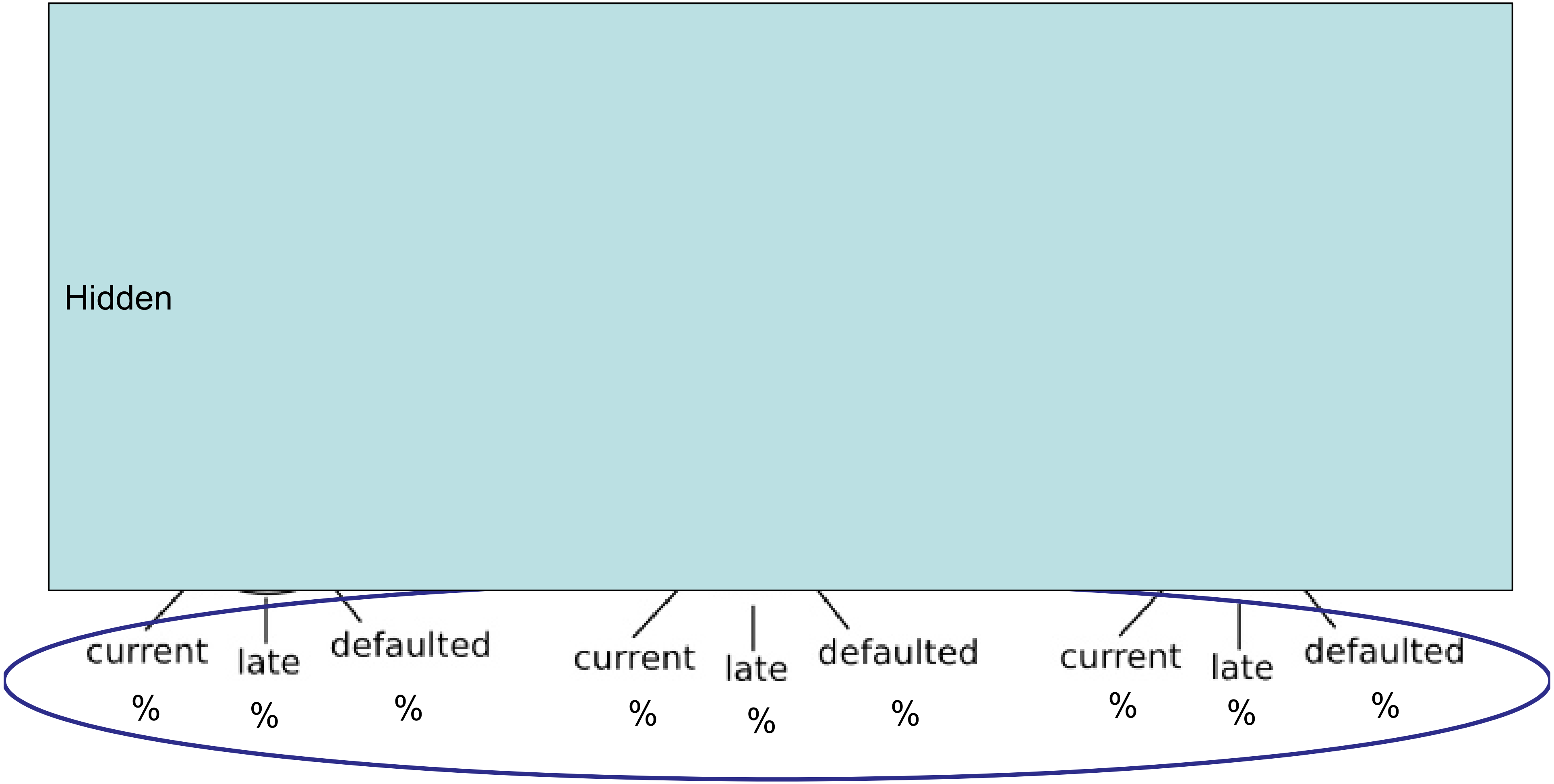
- Typically with Hidden Markov Models
- Extension of Markov Chains
  - Markov Chain example:



  - Typically the task is to calculate the probability that a sequence of states (e.g., rainy, rainy, sunny) appears.
- In HMMs the states are hidden and only observations are visible

Markov Model Idea

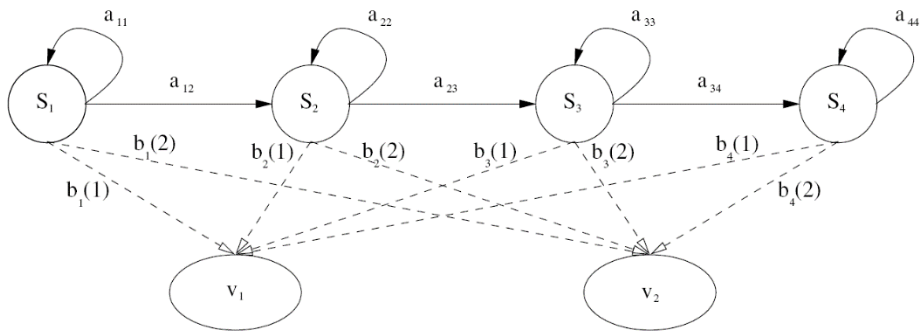



Hidden
current
late
defaulted
%
%
%
current
late
defaulted
%
%
%
current
late
defaulted
%
%
%

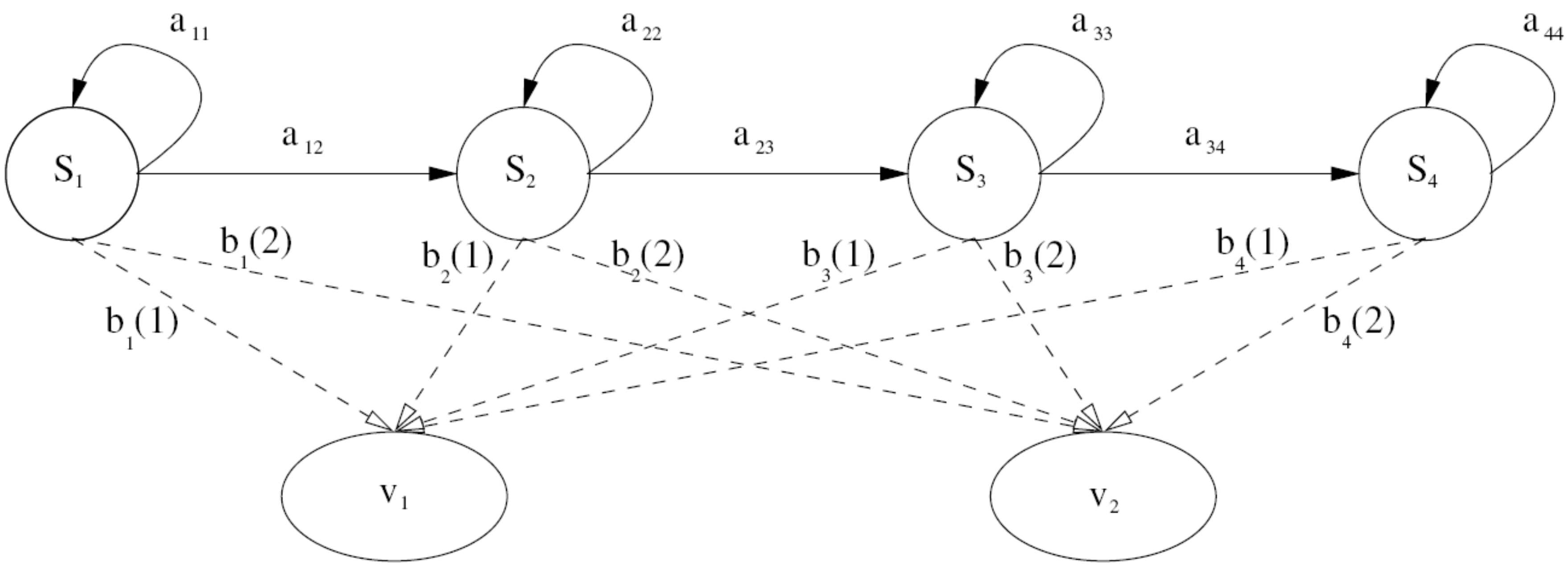
Hidden Markov Models (HMM)

Important

- (S,V,A,B,π), where
  1. $S = \{S_1,\ldots, S_N\}$ set of hidden states. q_t at time t
  2. $V = \{v_1,\ldots, v_M\}$ distinct observation symbols (alphabet)
  3. $A = \{a_{ij} \mid 1<=i, j <=N\}$ probability distribution of state transitions
     - $a_{ij} = p(q_{t+1} = S_j \mid q_t = S_i)$
  4. $B = \{b_j(k) \mid 1<=j<=N, 1<=k<=M\}$ observation symbol probability distribution
     - $b_j(k) = p(v_k \text{ at } t \mid q_t = S_j)$
  5. $\pi = \{i \mid 1<=i<=N\}$ initial state distribution



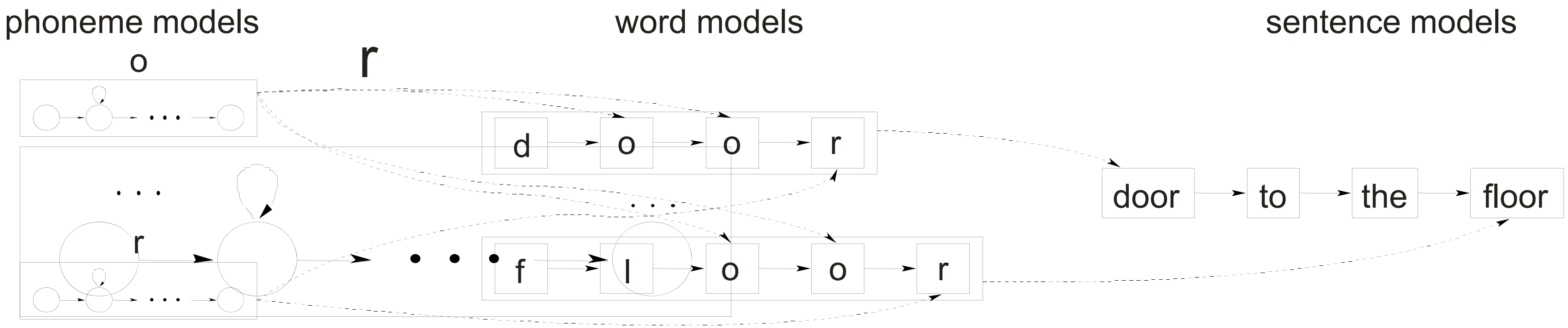
HMM (continued)



- Four states
- Two observations
- Linear topology

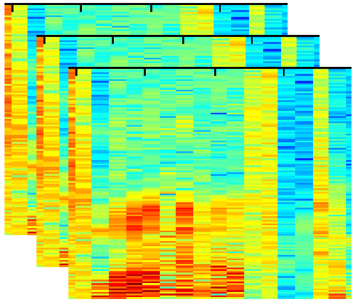
HMM Based Classification

- One HMM for each phoneme (Gaussian probabilities)
- Phoneme models are concatenated to word models
- Then further concatenated to sentence models
- Training using Baum-Welch algorithm
- Testing with Viterbi algorithm

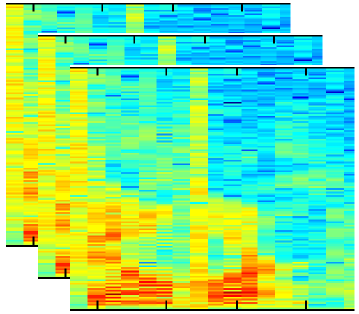


How to use HMMs for ASR

- Training *n* models
  - Maximize the probabilities that an HMM produces the features

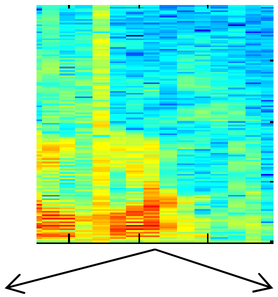


→ HMM_1 …



→ HMM_n

- Testing
  - Calculate the probabilities that all models produce a given sequence *X*
  - The one with maximum probability is the winner
    $argmax_i(P(X|HMM_i))$



HMM_1 % HMM_n %