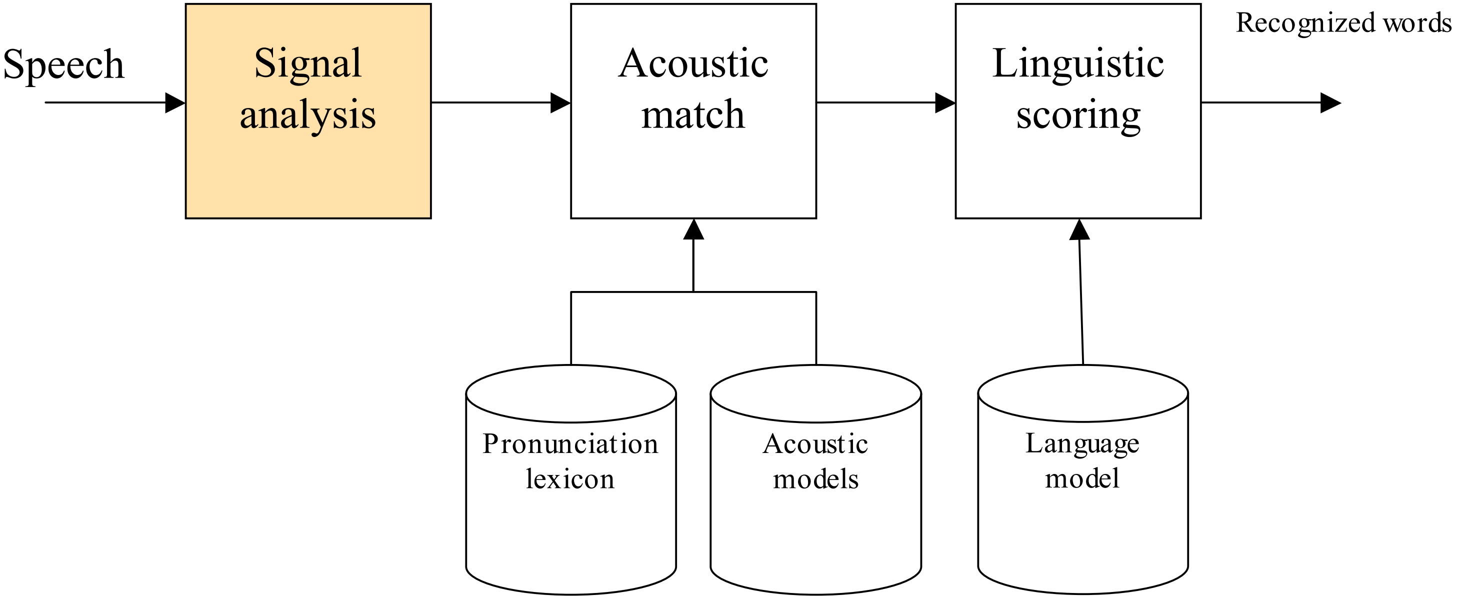
Feature extraction - outline

- Desirable properties of features
- LPC based analysis
- Non-linear frequency scale
- Cepstrum
- Dynamic features

ASR step-by-step: Feature extraction

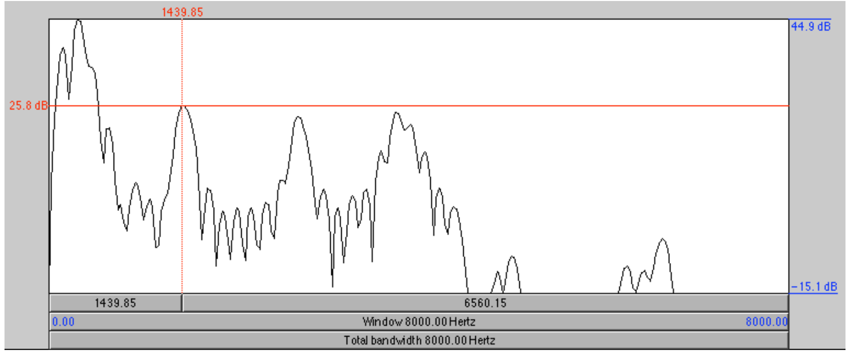


Feature extraction/signal analysis

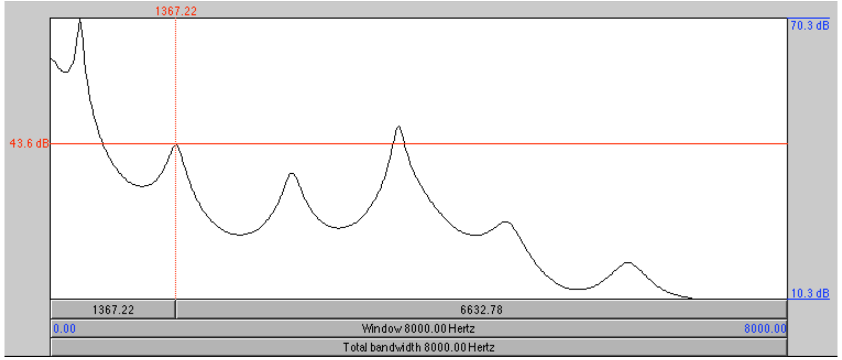
- Waveform is inappropriate for recognition
  - Variability
  - Dimensionality
- Need for a speech representation that is
  - Suitable for discriminating phonetically different sounds
  - Invariant to intra- and interspeaker variations
  - Robust against noise
  - Compact
  - Suited for pattern classification method
- Speech production and perception are closely linked
  - We do not make an articulary effort if difference cannot be perceived
  - We do not listen for differences that are never produced

Speech analysis

- Hearing: Ear performs short-time spectral analysis of sounds
- Source-filter model
  - – Sound discrimination: ”excitation information is not required”
    - What about tonal languages?



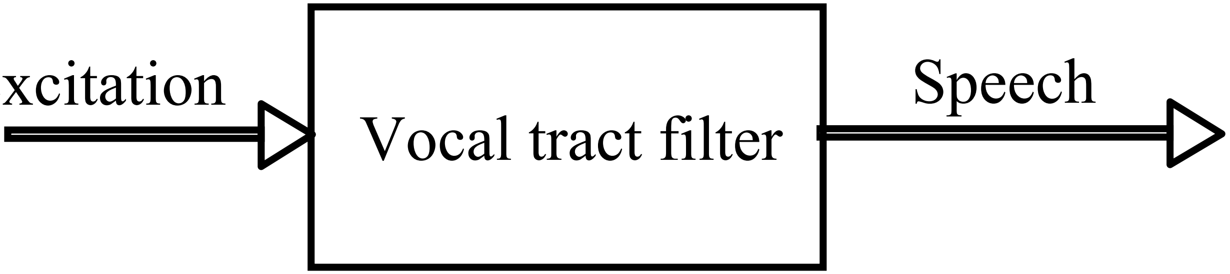
Power spectrum



Spectral envelope

- Spectrum estimation requires sufficently large time slice to be reliable
- Speech is time varying - find suitable compromise
- Typical: 25ms time window, analysis performed every 10 ms

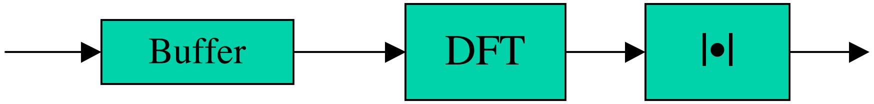
Analysis based on speech production



- Vocal tract model of uniform, lossless tube sections lead to a vocal tract filter wich is an all-pole filter
  - I.e. can model resonances well, but not well suited to modeling spectral valleys, e.g. in nasals
  - Inverse filter exists
  - Speech spectrum estimated as the power transfer function of the VTS
- Simple mathematical formulation
  - Linear Prediction, LPC coefficients
  - Many equivalent representations of the coefficients that are well suited for recognition purposes
    - Reflection coefficients, line spectral frequencies, log area ratios, ...

Alternative analyses

- LPC assumes specific model of speech production
  - Parametric spectral estimation
  - Model includes assumptions and justifications
- Non-parametric spectral estimation
  - Periodogram
    - Magnitude of short-time Fourier transform

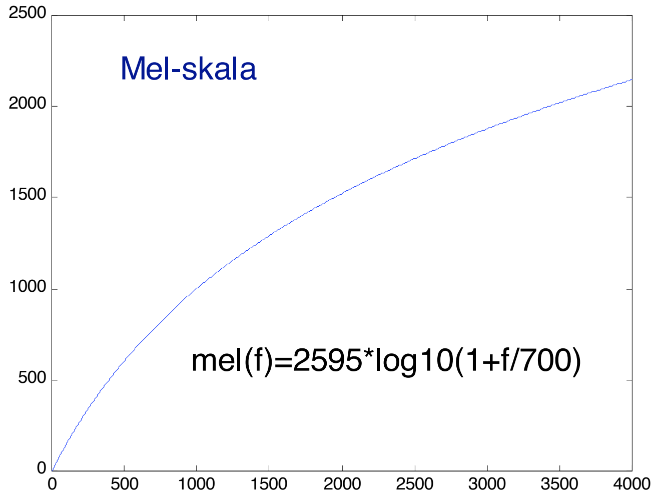


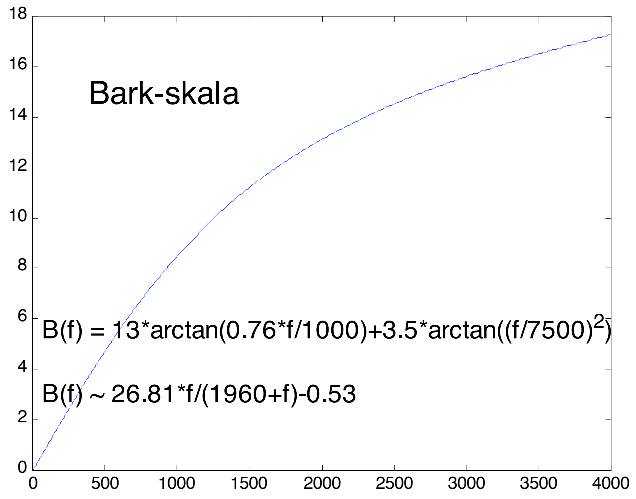
$$|S(\omega,m)| = \left|\sum_{n=-\infty}^{\infty} s(n)\cdot w(m-n)\cdot e^{-jn\omega}\right| \quad (|\mathrm{S}(\omega_{\mathrm{i}},m)| = \left|\sum_{n=m}^{m+N-1} s(n)\cdot e^{-j(n-m)\omega_i}\right| \text{ ; DFT)}$$

- Similar to a time slice in the spectrogram
- Not a very good spectrum estimate
- Can be interpreted (and implemented) as a filter bank

From linear to perceptual frequency scale

- Hearing/perception:
  - Frequency dependent temporal resolution
  - Frequency dependent loudness sensitivity
  - Non-linear frequency resolution
- Non-linear frequency scale:





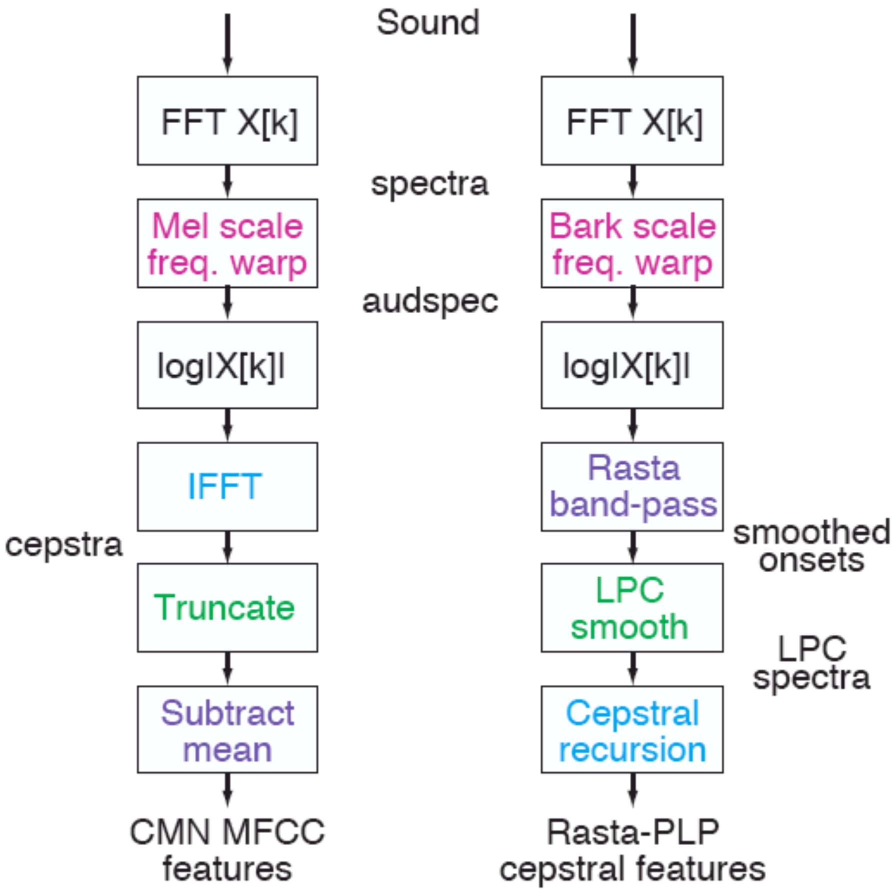
Cepstrum

- Inverse discrete Fourier transform of the log magnitude spectrum

$$c(n) = IDFT\{\log\left|S\ (e^{j\omega})\right|\}$$

- Efficient for decoupling source and filter due to the log operation
- Performs a decorrelation of the parameters
  - Desirable for compactness
  - In statistical pattern recognition, the correlation matrix of the parameter vector is often used. Decorrelated parameters makes this matrix diagonal, i.e. defined by *N* parameters instead of *NxN*
- The log magnitude spectrum is real and symmetric
  - The inverse DFT can be implemented as the less computationally demanding Discrete Cosine Transform

MFCCs and PLPs



- Perceptually based frequency scale
- Perception based power compression (log or cubic root)
- Spectral smoothing (truncation of cepstrum or LPC)
- ~ Decorrelated parameters
- Increased robustness through mean subtraction or ”Rasta” filtering

Dynamic features

- Feature vectors corresponding to a short time spectral estimate represent a snap-shot of the speech signal
- Important information is contained in the temporal evolution of the signal (cfr. spectrograms)
- Dynamic features are approximations to the time derivatives of the spectrum/cepstrum
- Delta-coefficients (Furui):

$$\Delta c_n = \left(\sum_{i=-W}^{W} i c_{n+i} \right) / \left(\sum_{i=-W}^{W} i^2 \right)$$

$$\underset{W=2}{=} \frac{1}{10} \left[\left(c_{n+1} - c_{n-1} \right) + 2(c_{n+2} - c_{n-2}) \right]$$

- Similarly for acceleration
- RASTA-filtering
  - Time derivative + filtering of band energies