Specialization Module

Speech Technology

Timo Baumann
baumann@informatik.uni-hamburg.de

UNIVERSITÄT HAMBURG, DEPARTMENT OF INFORMATICS
NATURAL LANGUAGE SYSTEMS GROUP

Speech Parametrization

Audio vs. Speech

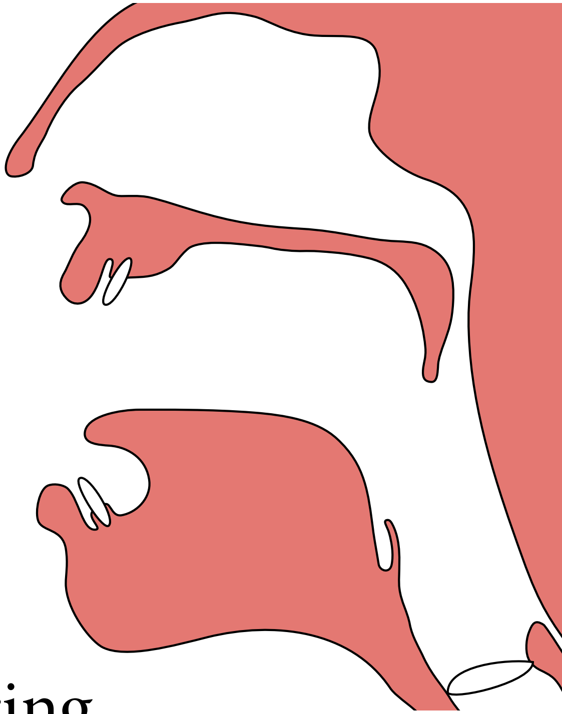
- speech audio is sampled and discretized by sound card
  - often 32 or 48kHz (i.e., 32000 samples per second)
  - often 16 bit or 24 bit samples
- ~64 Kilobyte of data per second

- spontaneous speech:
  - <8 phonemes per second
  - <64 phonemes in the phoneme set
- 9 bit/second should be enough!

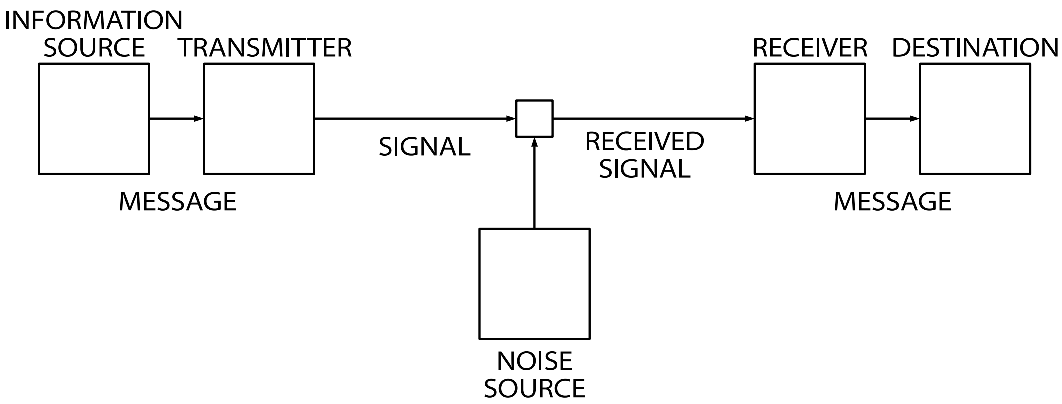
Goals

- reduce signal redundancy
  - simplify speech recognition/synthesis (e.g. MFCCs)
  - easier transmission (e.g. mp3, ...)
- increase signal-noise ratio
  - abstract away from irrelevant signal properties
  - keep relevant properties
    - depending on the goal: speech content, colouring, speaker properties,...
- (loosely) based on insight about human auditory processing:
  - semi-stationarity, spectral analysis, phase dropping, frequency binning, source-filter separation

Idea: Inverse Filtering



- the glottal folds produce a primary (saw-tooth-like) signal
  - rich in overtones/harmonics
- the vocal tract acts as a (frequency) filter
  - mostly attenuation
- aim: separate primary signal from vocal tract filtering
  - problem: source is not additive noise but convolutional



Cepstral Analysis

- method to separate signal source from filter
  - filter parameters determine signal envelope → phones
  - glottal source parameters unimportant to distinguish phones
- idea of cepstral analysis:
  - fourier transform turns convolution into multiplication
  - logarithm reduces multiplication to addition
  - another transformation (**spec**trum->**ceps**trum) results in parameters describing the signal envelope

Formal Problem Statement

- given
  - $s(t)$: the source signal
  - $v(t)$: the vocal tract
- we get
  - $x(t) = s(t) \otimes v(t)$ (convolution operator)
- transform to frequency domain:
  - $X(f) = S(f) \times V(f)$ (standard multiplication)
- now what?

Properties of s(t) and v(t)

what differentiates your primary signal (over time) from your vocal tract modifications (over time)

Deconvolution by Cepstral Analysis

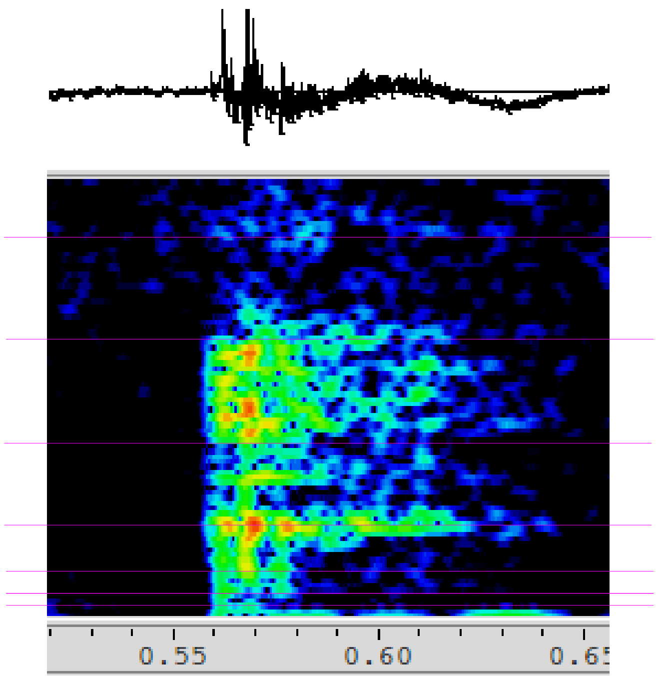
- make use of the fact that source dominates high frequencies and vocal tract filtering dominates lower frequencies:
  - $log|X(f)| = log|S(f)| + log|V(f)|$
    - taking the absolute values implies that we disregard phase information;
    - we've reduced convolution to multiplication to addition, nice!
- so far, frequencies are still correlated, but we can apply another round of (inverse) Fourier transform on the logarithmized spectrum (from the spectrum to the <u>ceps</u>trum):
  - $c(n) = c_s(n) + c_v(n)$
  - lower <u>quefre</u>ncy coefficients: vocal tract (i.e., signal envelope), higher <u>quefre</u>ncy coefficients: glottal excitation (including pitch)

Sliding Window

- window selects a stretch of audio (often 25 ms)
  - windowing function: Hamming/von Hann/…
- shift window by 10 ms (5 ms)
- 1 second of audio → 100 windows
  (~3-30 windows for one phoneme)
- perform signal analysis and parameterization on individual windows

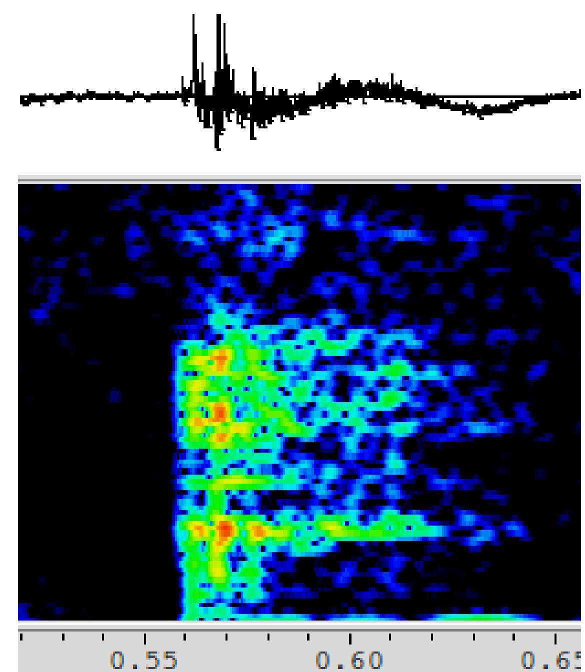
Mel-binning

- human auditory resolution differs with frequency
  - high resolution for low frequencies
  - low resolution for high frequencies
- add energy within frequency bins
  - small bins for low frequencies
  - increasingly larger bins
- often 16 or 32 bins



Fixed Windows vs. Landmark Detection

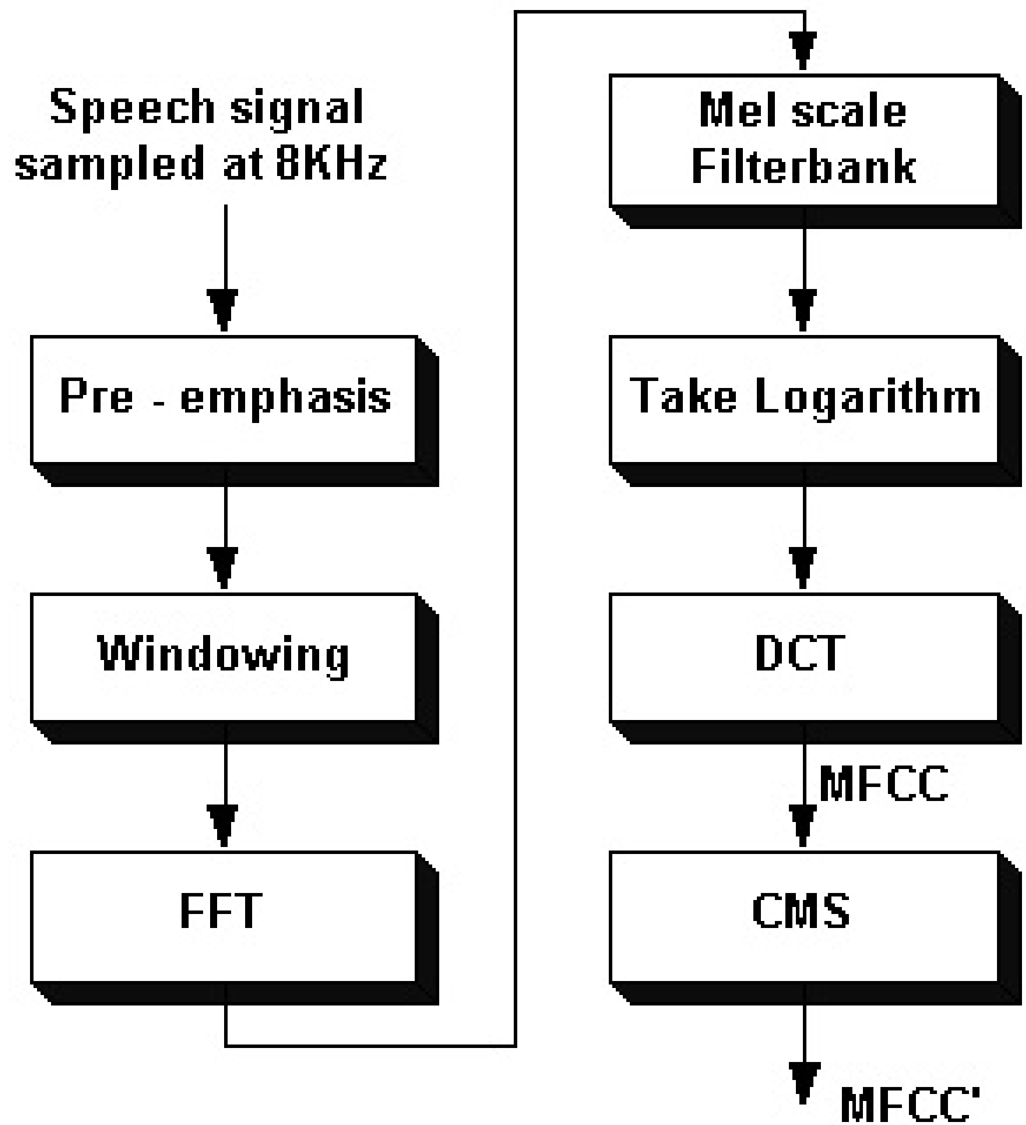
- some sounds (especially [p,t,k]) are badly described using fixed windows
  - the lip opening of plosives is very aprupt, not a slow change (as assumed by quasi-stationarity)
- fixed windows will hardly ever coincide with phoneme changes → blurring of details
- find landmarks, find stretches of speech in-between



How about (originally) additive noise?

Cepstral Mean Normalization

- noise will usually end up in all cepstral components
- not all components will center around 0
  - Z-normalization of individual components
    - compute mean and stddev
    - subtract mean, multiply by stddev
  - often just mean subtraction, no full normalization
- often performed locally used a sliding window to estimate mean and stddev
- many more advanced techniques to reduce the impact of noise


Speech signal
sampled at 8KHz
Pre - emphasis
Windowing
FFT
Mel scale
Filterbank
Take Logarithm
DCT
MFCC
CMS
MFCC'

Summary

- speech is a quasi-stationary signal
  - analyze sliding windows
- signal analysis
  - deconvolute signal filter from source using cepstral processing
  - use Mel-binning to model human frequency sensitivity
- phonemic information is largely contained in lower quefrency components
- prosody is largerly contained in higher quefrency components

Thank you.

baumann@informatik.uni-hamburg.de

https://nats-www.informatik.uni-hamburg.de/SLP16

Further Reading

- accessible introduction to signal parametrization:
  - P. Taylor (2009): *Text-to-Speech Synthesis*. Cambridge Univ Press. ISBN: 978-0521899277. InfBib: A TAY 43070.
  - D. Jurafsky & J. Martin (2009): Speech and Language Processing. Pearson International. InfBib: A JUR 4204x
- in-depth mathematical approach:
  - Rabiner & Juang (1993): *Fundamentals of Speech Recognition*. Prentice Hall. Stabi: A 1994/994.

Notizen

Desired Learning Outcomes

- understand the task of speech parametrization and how it relates to the source-filter model and the general model of communication
- know the processing steps to produce parameters and be able to discuss alternatives