Analysis of Sequential Data

Analysis of Digital Signals

Prof. Dr. H.-P. Hutter, HII-Group, InIT/ZHAW,
hans-peter.hutter@zhaw.ch

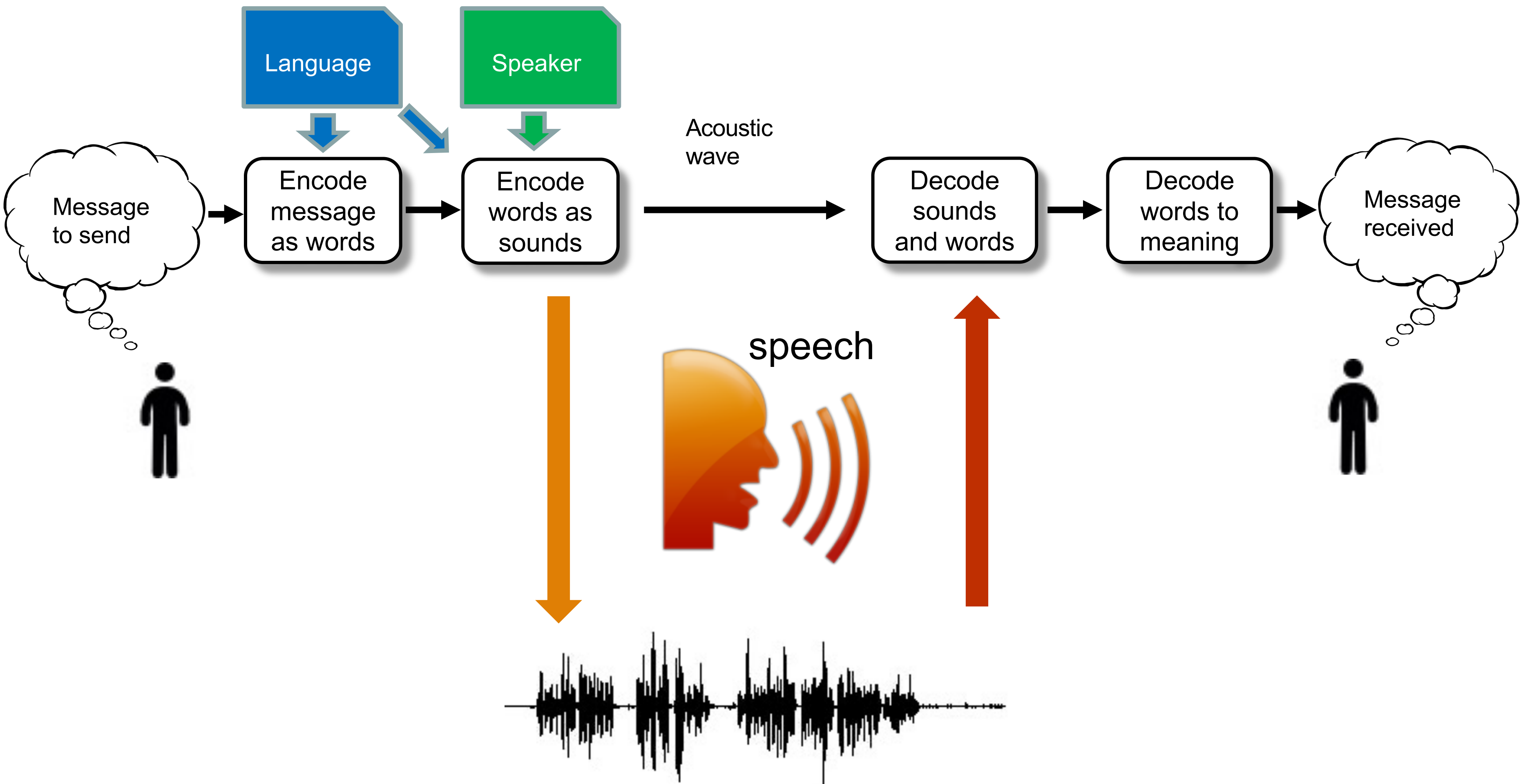
Objectives

- You know different techniques to analyse digital signals in different domains
- You know the most important characteristics of a speech signal
- You know the basic units of speech and their characteristics
- You know the most important features used in speech recognition

Speech Signal

- Speech Signal
  - Is one of the most challenging temporal signals to analyse
  - Produced by humans for humans to transmit a message
  - It has a lot of interesting applications
    - Speech recognition
    - Speaker recognition
    - Language recognition
    - Speech classification
    - Speech synthesis
    - ...
    - Conversational Assistants
    - Natural Language Processing

What is Speech?



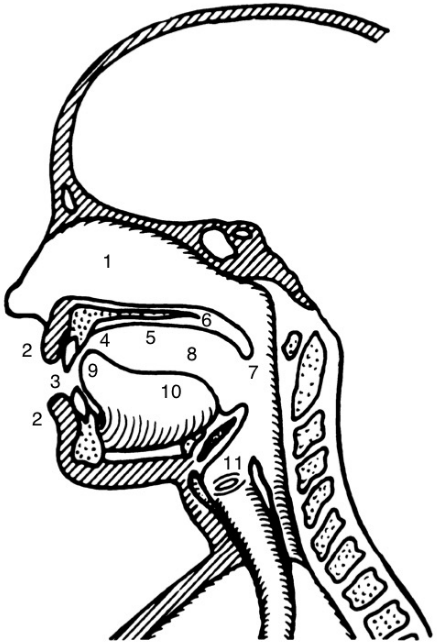
Speech Signal
What we know

- Speech signal
  - Longitudinal pressure wave
  - Very large power range
    - 0 dB : faintest audible sound
    - 120 dB: loudest sound, human ear can tolerate (10^6 times as loud)

- Is transformed by microphone into
  - Analog eletrical signal that is later on sampled
  - or directly sampled to a digital signal

Speech Signal
What we know

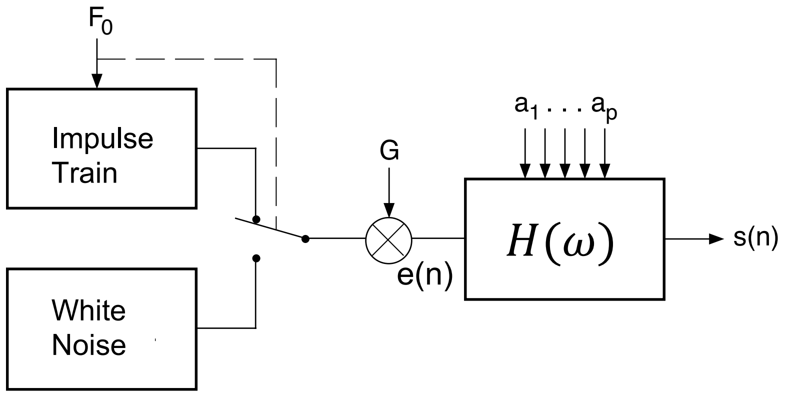
- Humans produce it
  - Vocal tract
  - For humans



1 Nasal Cavity
2 Lips
3 Teeth
4 Tooth-ridge
5 Hard palate
6 Velum
7 Uvula
8 Cavum Oris
9 Tongue tip
10 Tongue middle
11 Vocal Cords (Glottis)

Speech Signal
What we know

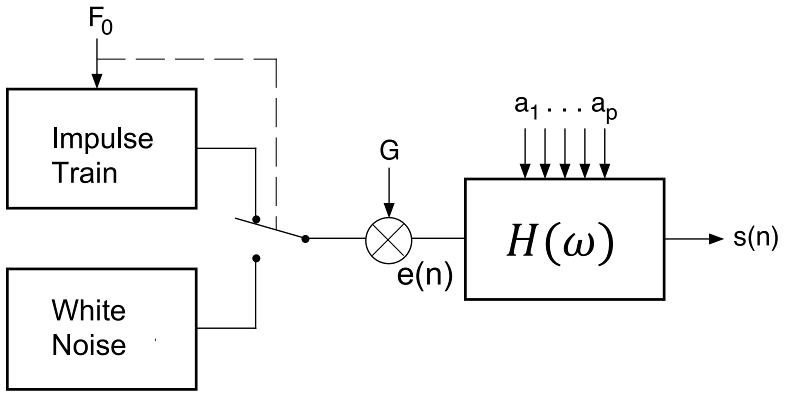
- How do human produce speech?
- Simple speech production model
  - Glottis either produce an impulse train with fundamental frequency F_0 or white noise
  - Excitation function e(n)



Speech Signal What we know

- Simple speech production model
  - Vocal tract filters this signal e(n) with filter H(ω) (impulse response $h(n)$)

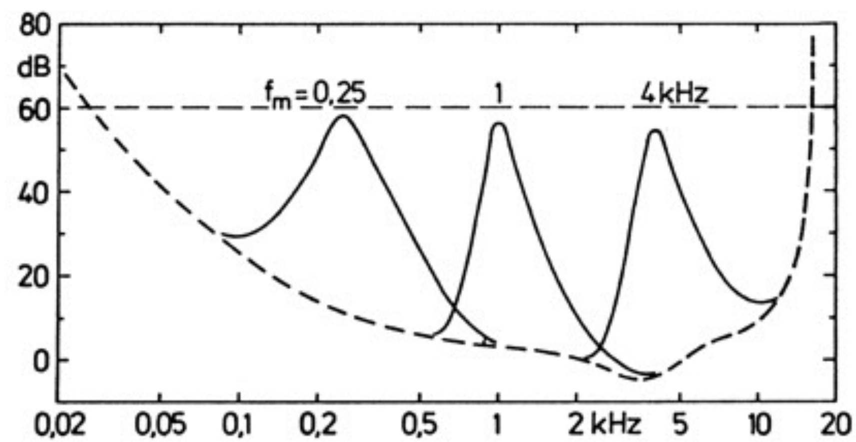
$$s(n) = h(n) * e(n)$$



  - Filter characteristic H(ω) is determined by the coefficients $a_1, a_2, \dots, a_p$
  - $a_1, a_2, \dots, a_p$ can be estimated from the signal s(n) with LPC (Linear Prediction Coding)
    - $a_1, a_2, \dots, a_p$ are therefore called LPC-parameters

Speech Signal
What we know

- Humans understand it
  - Ear
  - Brain
- Main characteristics of human ear
  - Does some kind of spectral analysis of a sound
  - Hears sounds from ca. 20 Hz-18 kHz
  - Logarithmic sensitivity
    - Sound pressure
    - Sound frequency
  - Energies over neighboring frequencies are intergrated



Speech Signal
What we know

- Phonemes
  - Smallest sound units that distinguish different words
  - Speaker independent
  - Notation: IPA (SAMPA)
- Phones
  - Acoustic representation of the phoneme
  - Speaker dependent
- Examples of phonemes
  - b, d:
    - Bad [bæd] <-> Dad [dæd]
  - æ, e:
    - Bat [bæt] <-> bet [bet]

Speech Signal
What we know

- Which phonemes exist in English but not in German and vice versa?
  - English:

    Your answer?

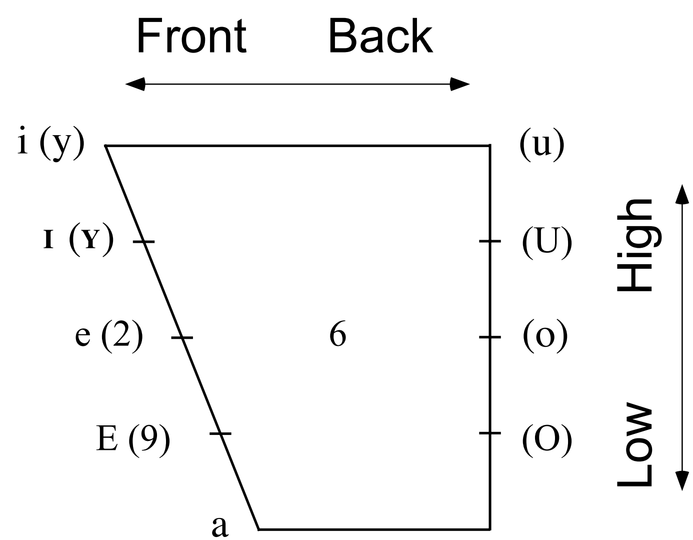
  - German:

    Your answer?

Speech Signal
What we know

- There are 20-60 phonemes in western languages
- German:
  - 48 phonemes
- Two major classes
  - Vowels
  - Consonants

- Vowels



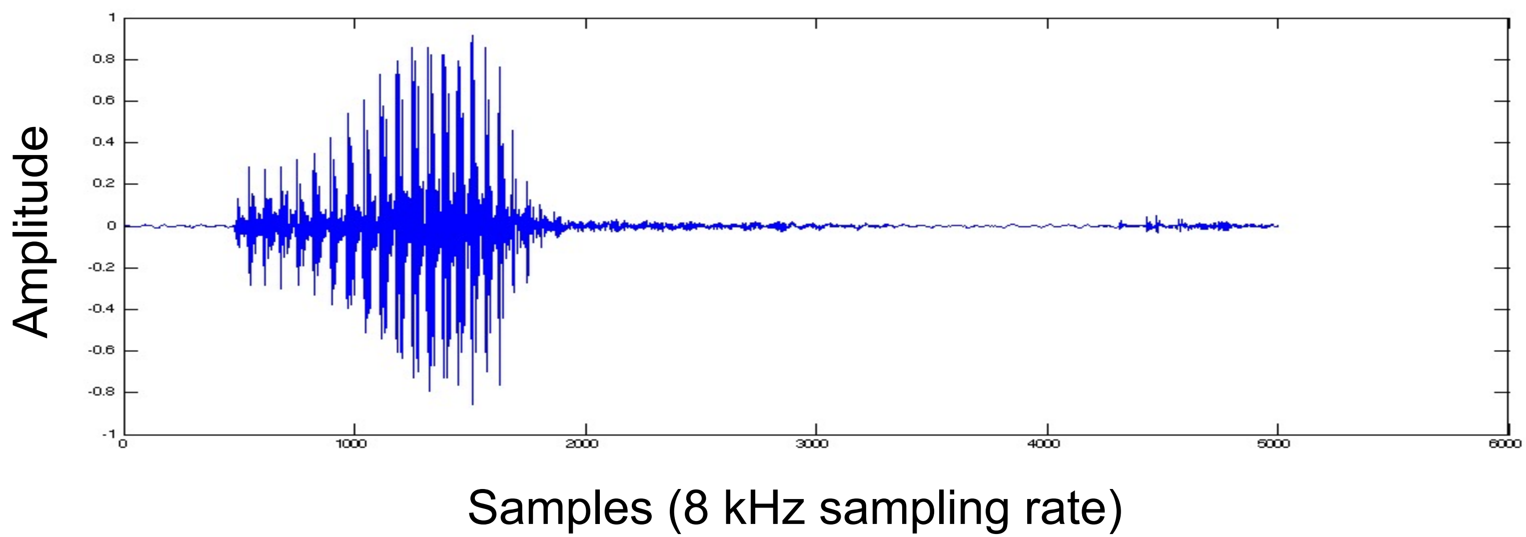
- Diphtongs (German)
  - aI (Bein), aU (Haus), OY (Heu)

Speech Signal
What we know

- Consonants
  - Fricatives: voiced/unvoiced
    - f, v, s, S, z, Z, h
  - Plosives: voiced/unvoiced
    - p, t, k, b, d, g
  - Nasals/laterals
    - m, n
    - l, r, R
  - Others
    - ? (glottal stop)

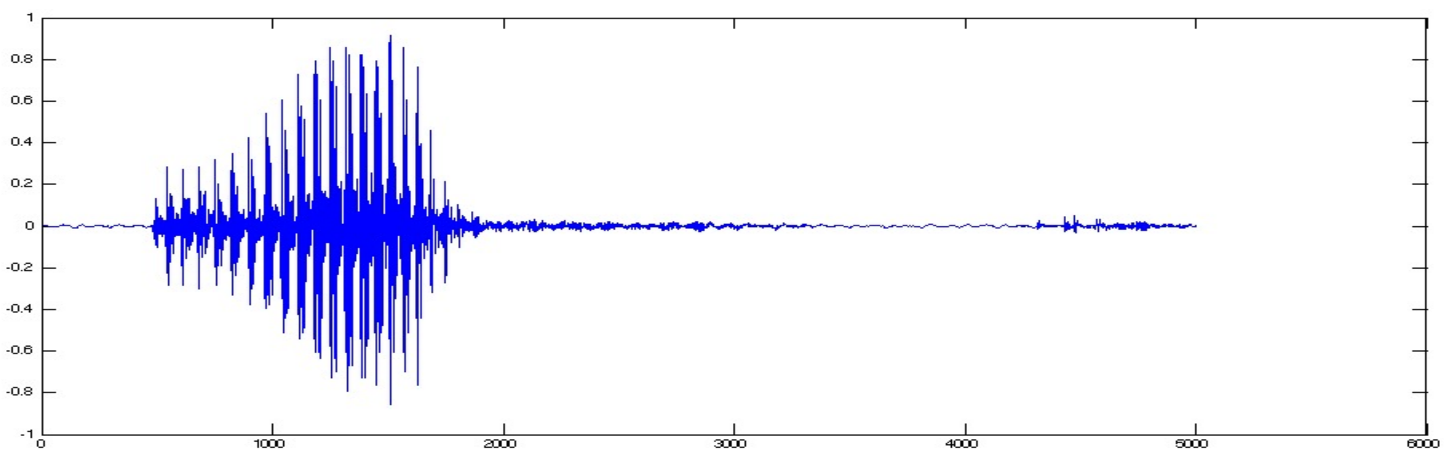
Analysis of Speech Signal
Time Domain

What do you see?



Analysis of Speech Signal Time Domain

- Periodic part:
  - Fundamental frequency ca. 112 Hz -> male
  - Fundamental frequency between 50 Hz (deep man's voice) and 400 Hz (child's voice)

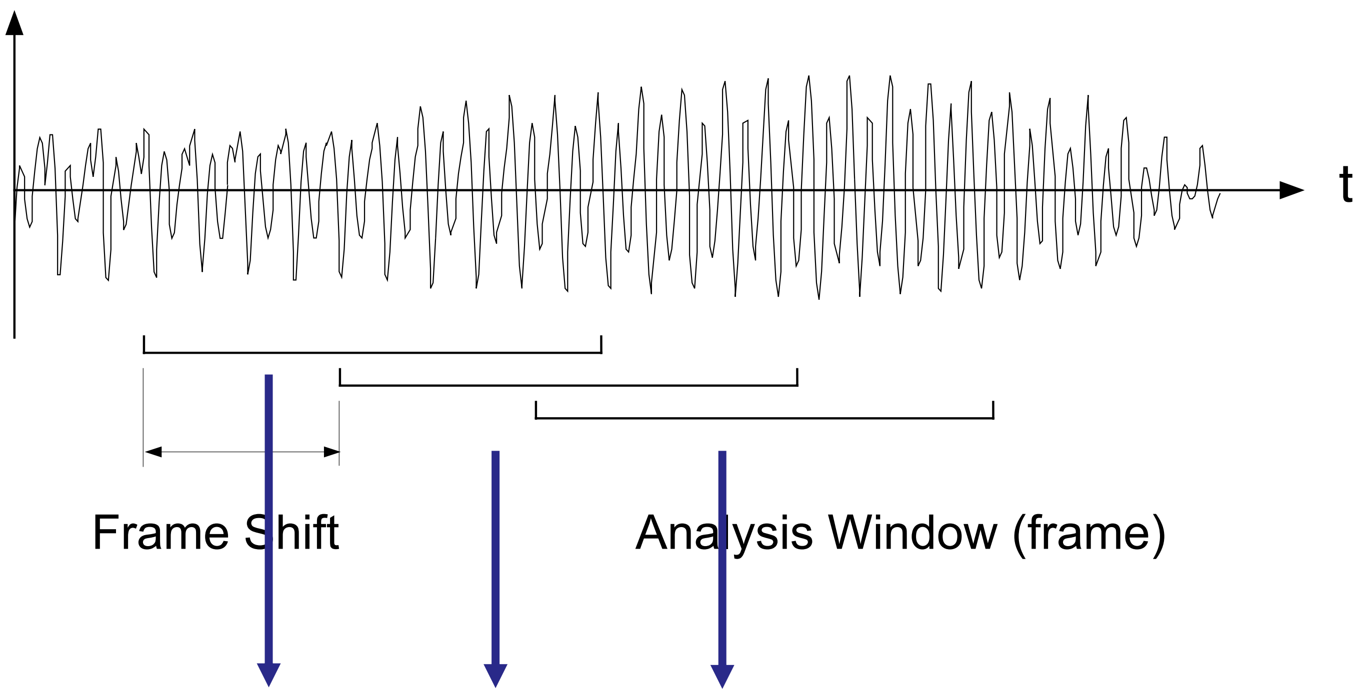


Analysis of Speech Signal Frequency Domain

- Problem
  - If we calculate the spectrum over the whole signal, we only get the averaged spectrum of the whole signal
    - Not very interesting for speech recognition
    - Main information lies in the change of the spectral characteristic
- Solution
  - Short-time spectral analysis
    - We only analyse a short segment of the signal at a time (window)
      - We assume that characteristic of speech signal does not change significantly within this window
    - We shift the analysis window a small amount in time (frame shift) and do the same short-time analysis again

Analysis of Speech Signal
Short-time Analysis

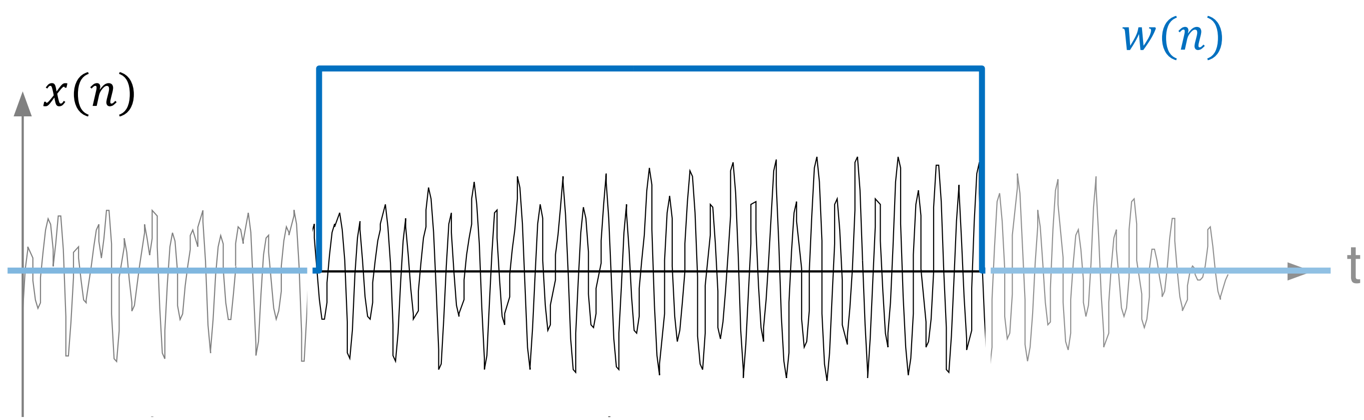
- Short-time Analysis



Feature Vectors $\boldsymbol{x}(1)$ $\boldsymbol{x}(2)$ $\boldsymbol{x}(3)$ ...

Analysis of Speech Signal
Short-time Analysis

- Analyzing only a window of the signal is equivalent to multiplying the signal with a rectangular window function $w(n)$



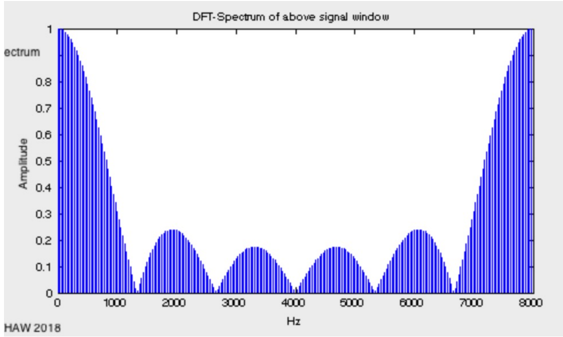
$$\bar{x}(n) = x(n)w(n)$$

Analysis of Speech Signal
Short-time Analysis

- Consequences of the windowing
  - Multiplying the window $w(n)$ with signal $x(n)$ in the time domain means in the spectral domain:

    The resulting spectrum is the convolution of the spectrum of the window and the signal

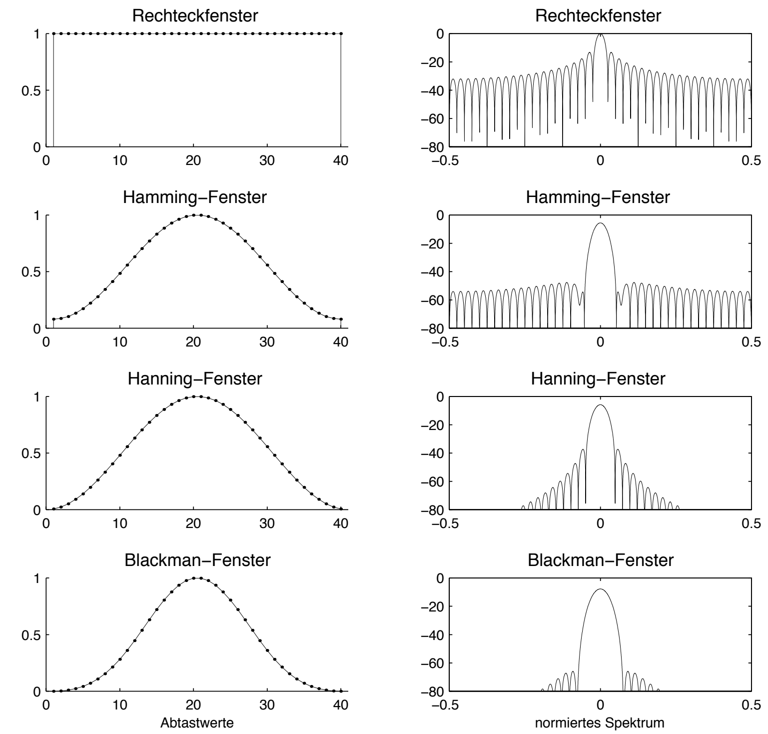
$$\bar{X}(\omega) = X(\omega) * W(\omega)$$



  - Spectrum of rectangle window:
    - $\sin(\omega)/\omega$ shape
    - Zeros at $\mathrm{f} = k\frac{f_S}{N}$, $N = \text{length of window}, k = \pm 1, \pm 2, \pm 3, \ldots$

Analysis of Speech Signal
Short-time Analysis

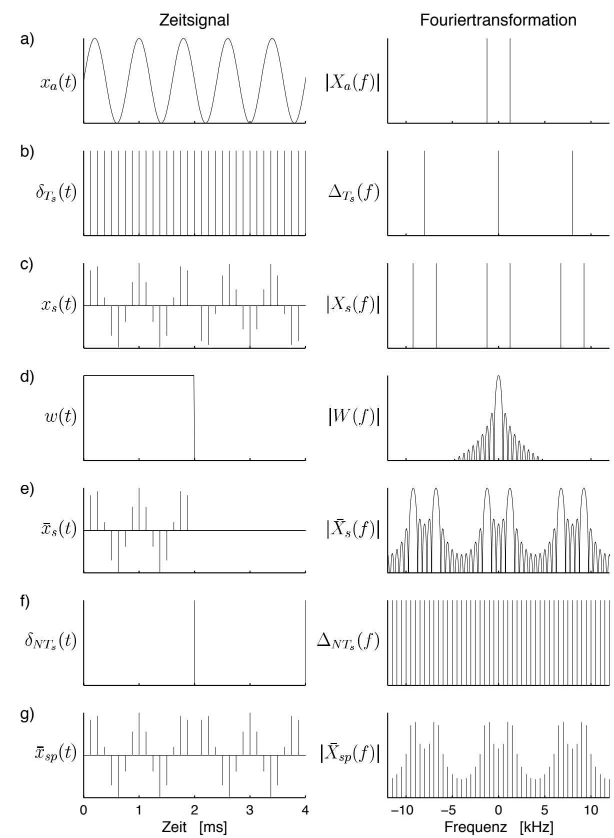
- Influence of the windowing
  - Spectrum is more or less blurred
    - Depending on the length of the window
      - Longer window -> less blurring
    - Depending on the window type
      - Relative hight of side lobes with resp. to main lobe
  - Different window types



Analysis of Speech Signal
Short-time Analysis

- When a analog signal is digitized and analyzed with a short-time window the following happens to the spectrum of the original signal
  a) The analog signal $x_a(t)$ is a sin-wave with 1250 Hz

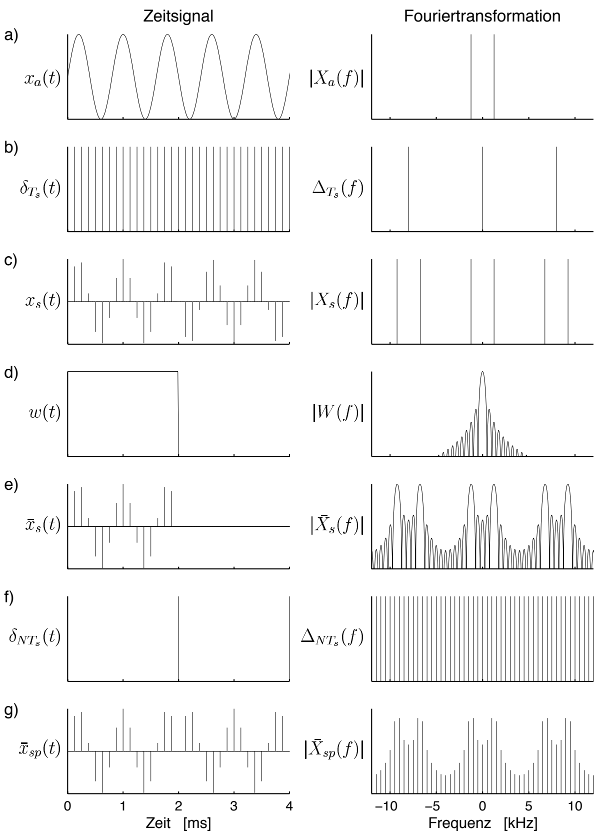
     The spectrum $X_s(f)$ of $x_a(t)$ shows exactly 2 spectral lines at +/- 1250 Hz



Analysis of Speech Signal
Short-time Analysis

b) Signal $x_a(t)$ is sampled with f_s = 8 kHz, i.e. it is multiplied with the pulse train $\delta_{T_s}(t)$ with a pulse period of $T_s = {}^{1}/_{f_s} =$ $125ms$

The spectrum $\Delta_{T_s}(f)$ of $\delta_{T_s}(t)$ is another pulse train who‘s pulse period equals the sampling frequency $f_s = 8kHz$



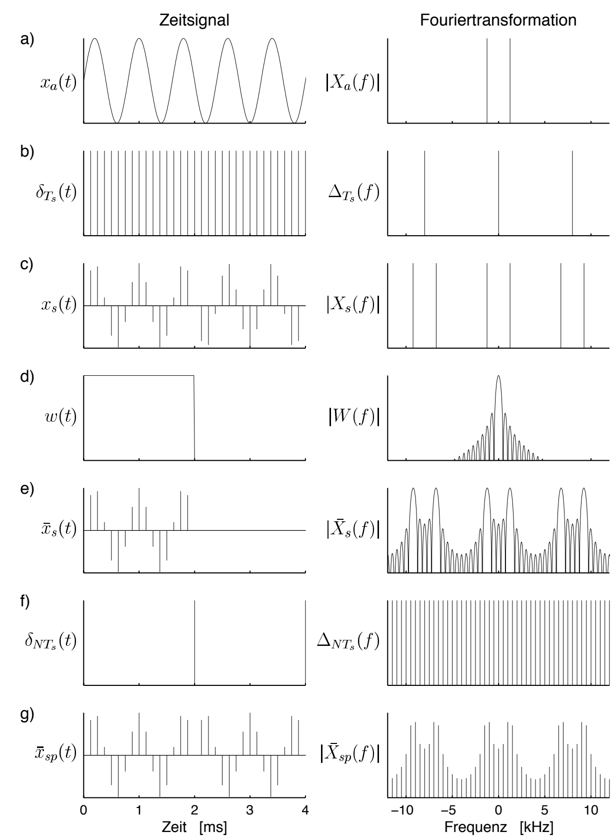
Analysis of Speech Signal
Short-time Analysis

c) The sampling results in a multiplication of the two signals $x_a(t)$ and $\delta_{T_s}(t)$

In the spectral domain this means that the resulting spectrum $X_s(f)$ is the convolution of the two spectra $\Delta_{T_s}(f)$ and $X_a(f)$:

$$X_s(f) = X_a(f) * \Delta_{T_s}(f)$$

$X_s(f)$ is periodic with f_s



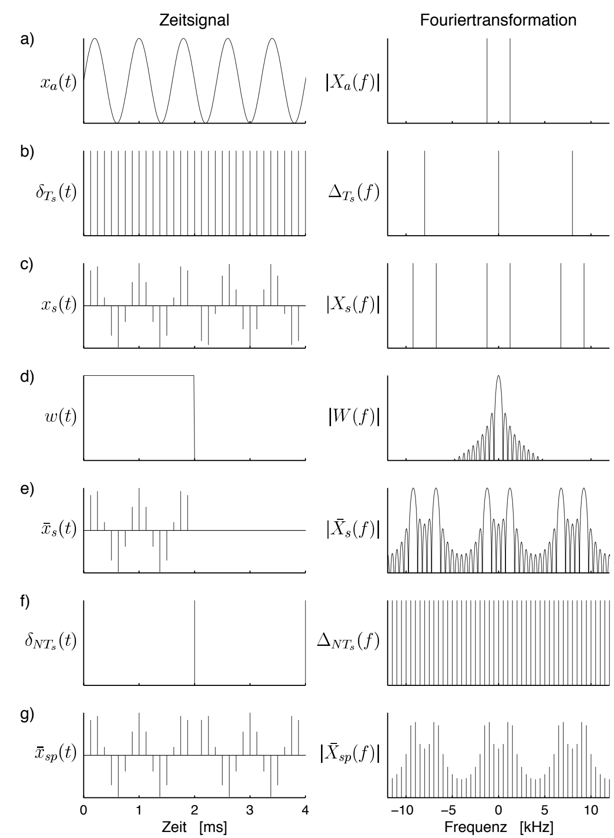
MASTER OF SCIENCE
IN ENGINEERING

Analysis of Speech Signal
Short-time Analysis

d) The rectangular window function $w(t)$ is N samples long.

The corresponding spectrum $W(f)$ is a $\sin(x)/x$ -function with zeros at frequencies

$$f = k\left({}^{f_s}/_{N}\right), k = \pm 1, \pm 2, \ldots$$



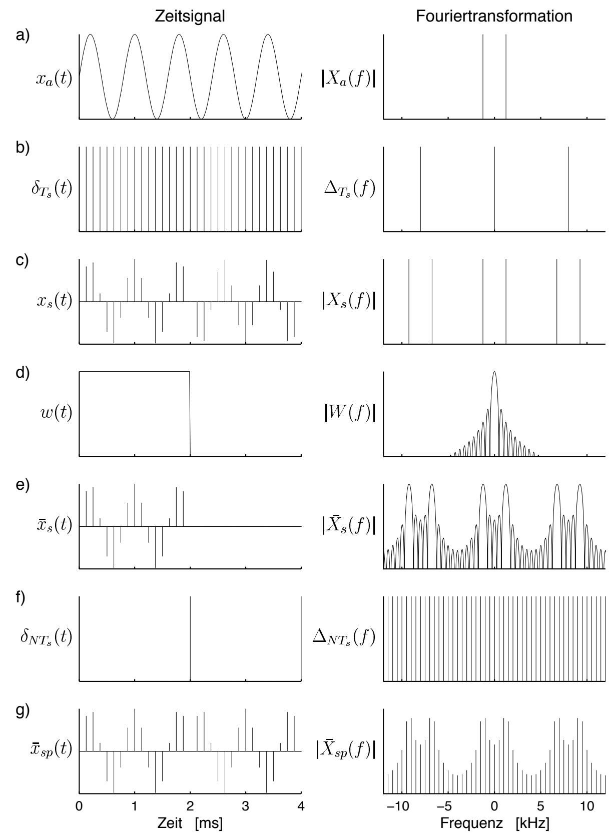
Analysis of Speech Signal Short-time Analysis

e) Windowing results again in a multiplication of the two signals $x_s(t)$ and $w(t)$

In the spectral domain this means that the resulting spectrum $\bar{X}_s(f)$ is the convolution of the two spectra $X_s(f)$and W(f) :

$$\bar{X}_s(f) = X_s(f) * W(f)$$

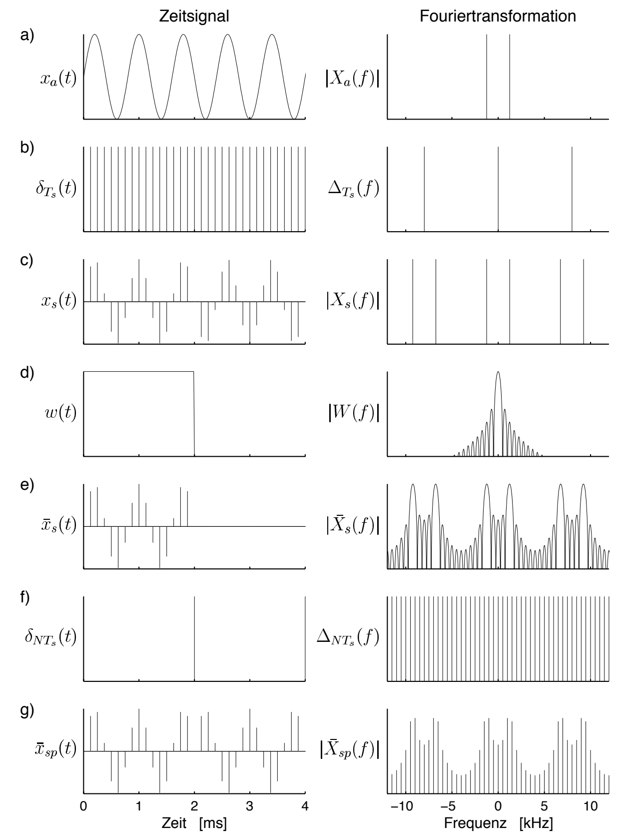
$\bar{X}_s(f)$ is also periodic with f_s



Analysis of Speech Signal Short-time Analysis

f) The spectrum $\bar{X}_s(f)$ is finally also sampled at N equally spaced points in the intervall $0 \leq f \leq f_s$

g) This corresponds to a muliplication of the spectrum $\bar{X}_s(f)$ in the spectral domain with the pulse sequence $\Delta_{NT_s}(f)$ In the time domain this means the the signal $\bar{x}_s(t)$ is periodically repeated with period NT_s



Analysis of Speech Signal
Short-time Analysis

- Conclusions
  - The N-point short-time analysis of a signal $x(n)$ corresponds to multiplying $x(n)$ with a rectangular window of length of N.
  - The N-point DFT assumes that the signal $x(n)$ as well as the spectrum $X(k)$ is periodic with period N.
  - In this case the N-point short-time DFT represents the exact spectrum of the original signal at N discrete points.
  - In all other cases the DFT-spectrum is only a more or less accurate approximation of the spectrum of $x(n)$.
  - The approximation is the more accurate the longer the window is in time.

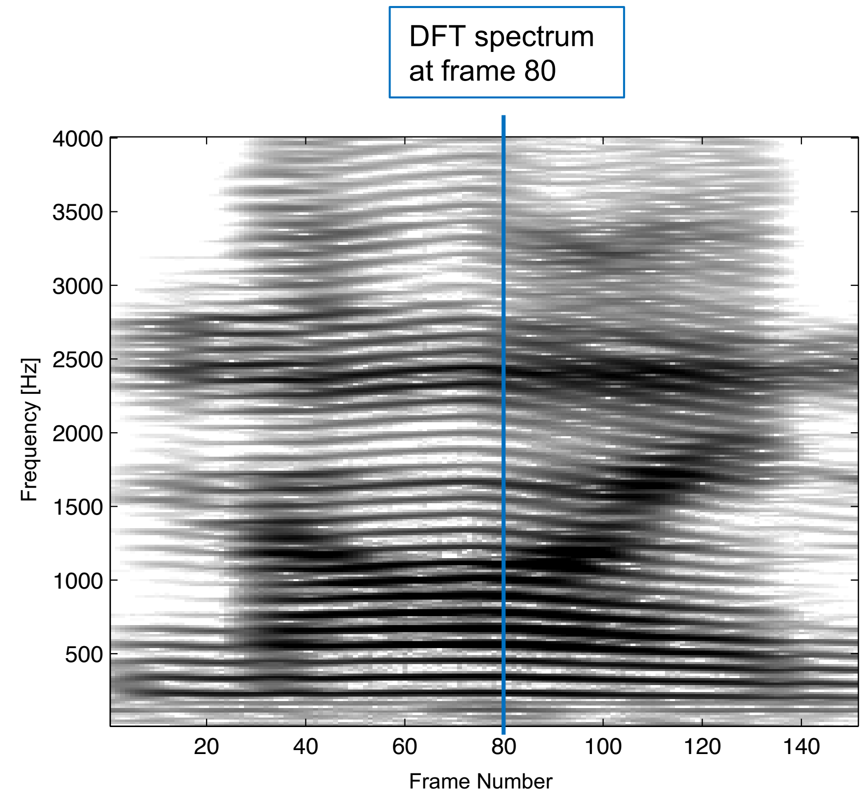
Analysis of Speech Signal
Frequency Domain

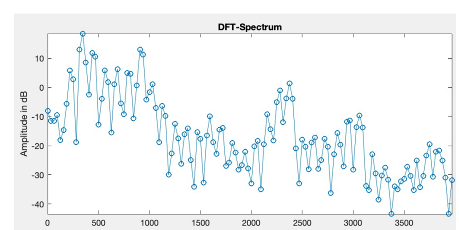
- With short-time analysis
  - As speech signal is quasi-stationary and the information we are interested in lies in the temporal variation
  - Therefore we calculate successive short-term spectra of the signal
  - Problem: How to visualize successive spectra?

- Analysis of spectral charactaristics of speech signal
  - Without a model: **Spectrogram**

Analysis of Speech Signal Spectrogram

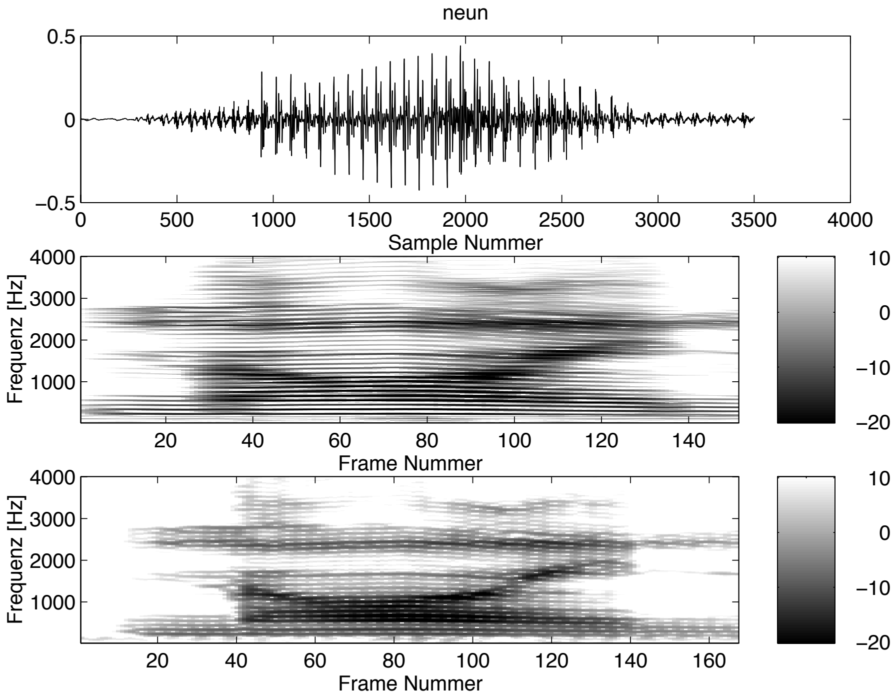
- Shows the temporal changes in the spectrum of a signal
- One vertical line shows the DFT spectrum of the signal at the corresponding time frame
  - Dark means high energy at that frequency
  - Light means low energy at that frequency





Analysis of Speech Signal Spectrogram

- Two types of spectrogram
  - Narrowband spectrogram
    - More spectral resolution
    - Long temporal window used (500 samples = 62.5ms)
  - Broadband spectrogram
    - Less spectral resolution
    - Short temporal window used (180 samples = 22.5ms)

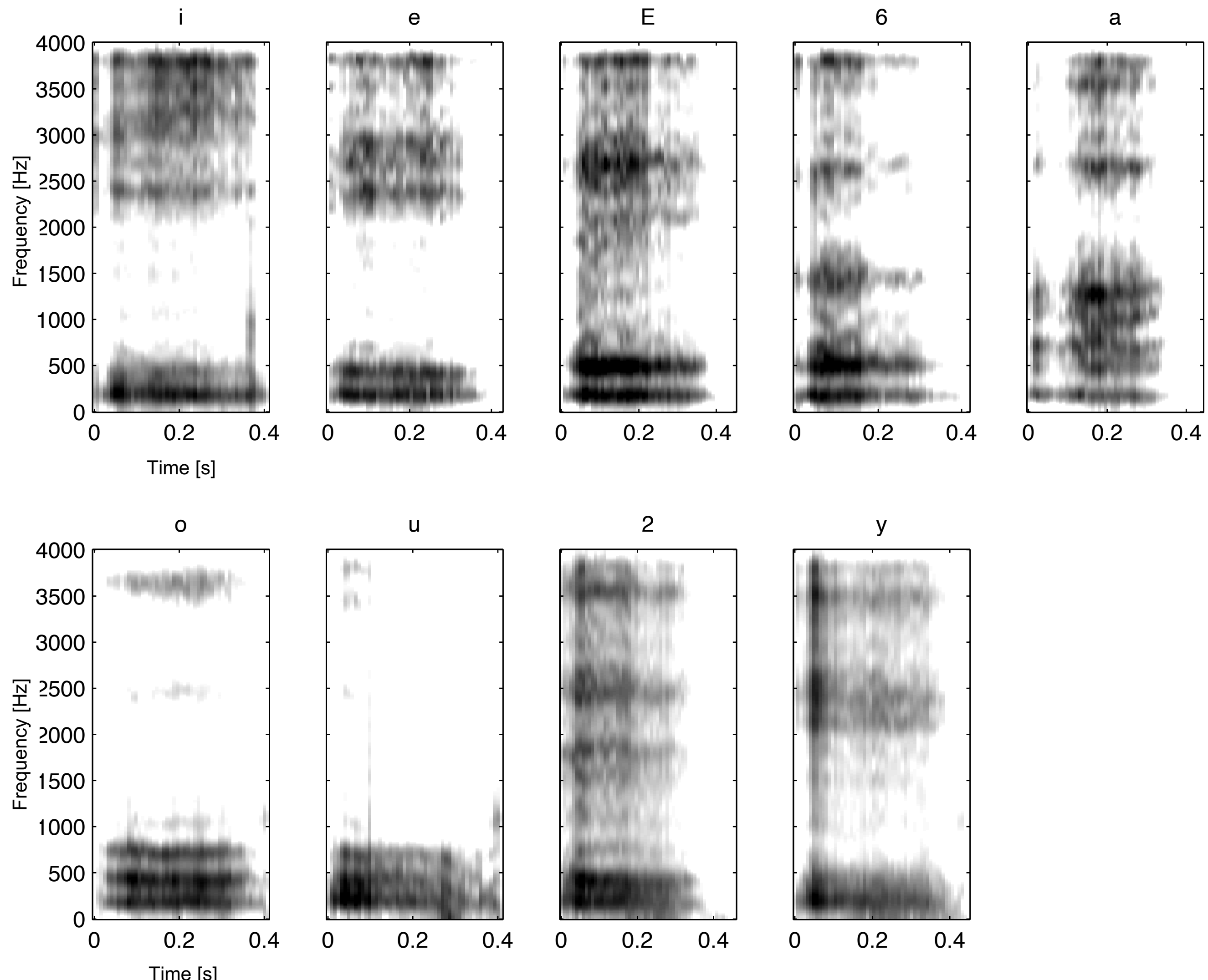


Analysis of Speech Signal
Formants

- Formants
  - Areas in the spectrogram with high energy contributions
    - Correspond to resonances in the vocal tract
    - Speech signal normally has 4-5 formants in frequency range 0-4kHz
    - Formants play an important role in characterising phones

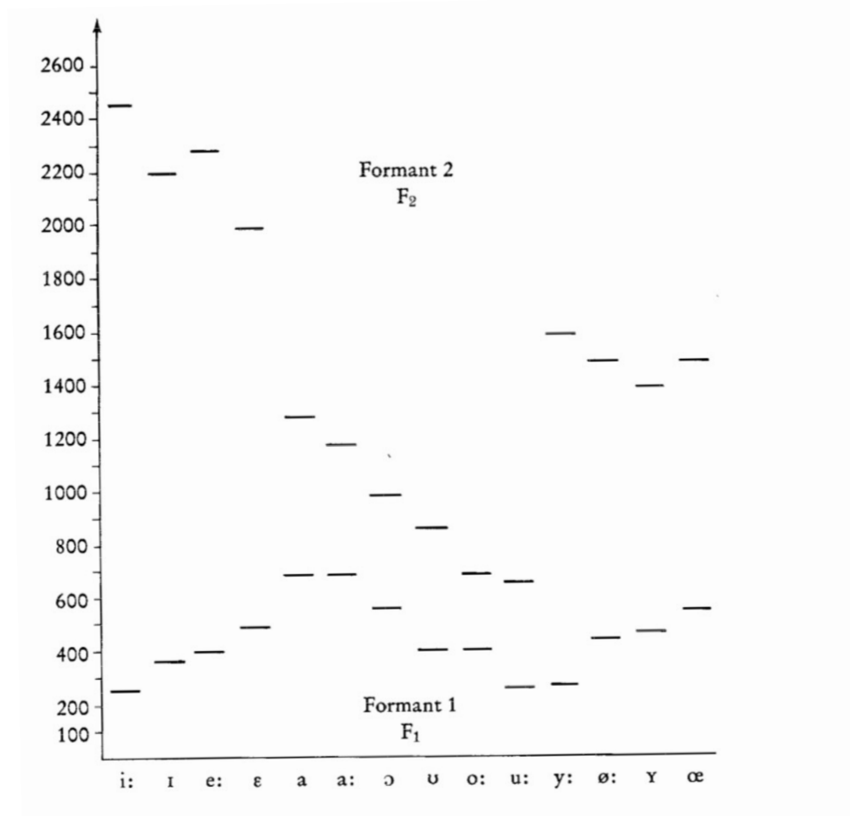
MASTER OF SCIENCE IN ENGINEERING

Analysis of Speech Signal Spectrograms of Vowels



Analysis of Speech Signal
Spectrograms of Vowels

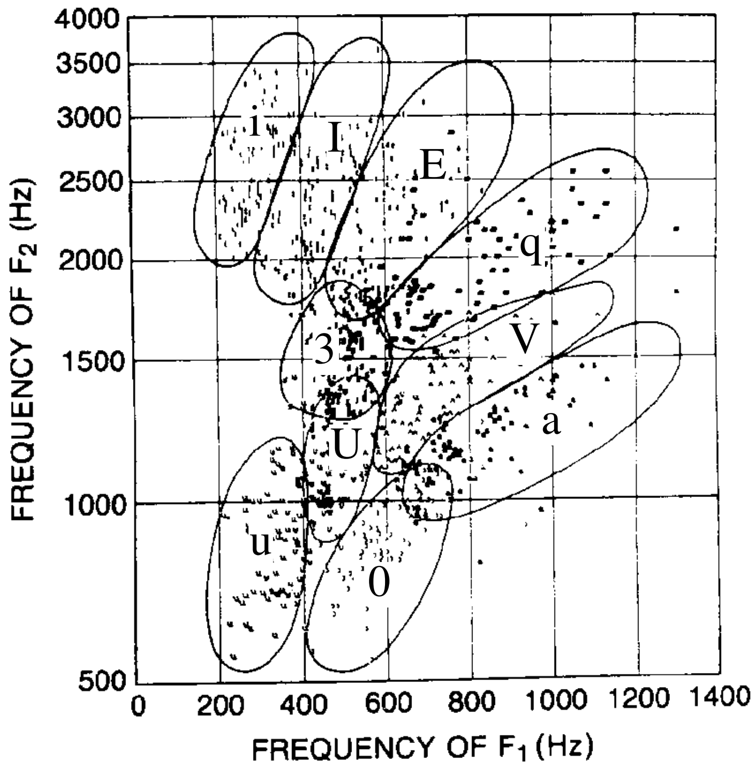
- Isolated vowels can be characterized by the average position of the first 2 formants F_1 and F_2



(Kohler, 1977)

Analysis of Speech Signal
Spectrograms of Vowels

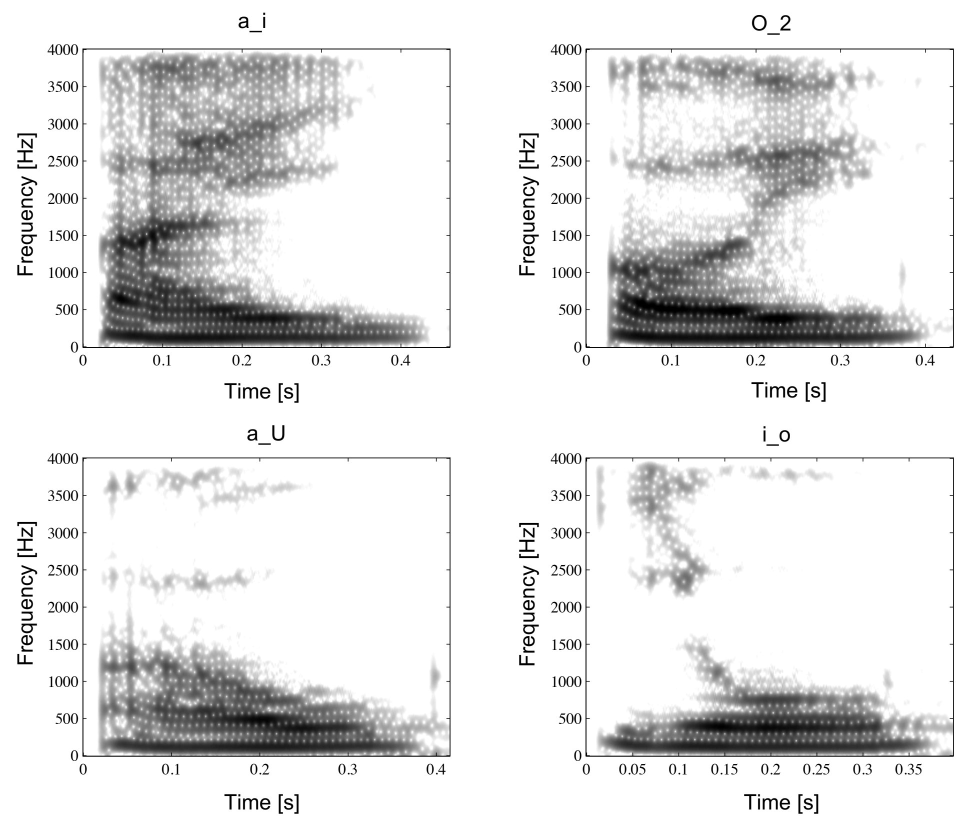
- However, the position of the of the first 2 formants F_1 and F_2 is speaker dependent!
  - E.g.: Formant positions of different English vowels
- Formant positions are not constant when vowels are spoken in context



(Peterson and Barney, 1952)

Analysis of Speech Signal
Diphthongs

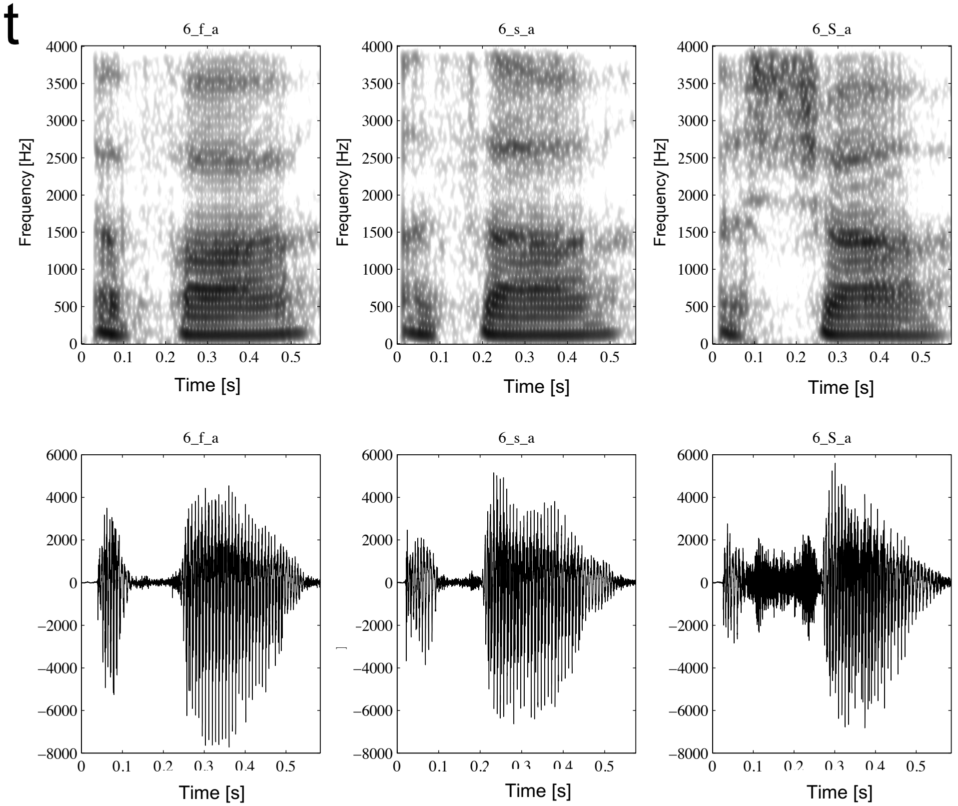
- Spectrograms of Diphthongs
  - Continuous movement of formants of 1st vowel to 2nd vowel



Analysis of Speech Signal Consonants

- The different types of consonants have different characterstics in the spectrogram
  - Fricatives
    - voiced/unvoiced
  - Plosives
    - Voiced/unvoiced
  - Nasals
  - Laterals
  - Others: r, R, |

- Fricatives (unvoiced)

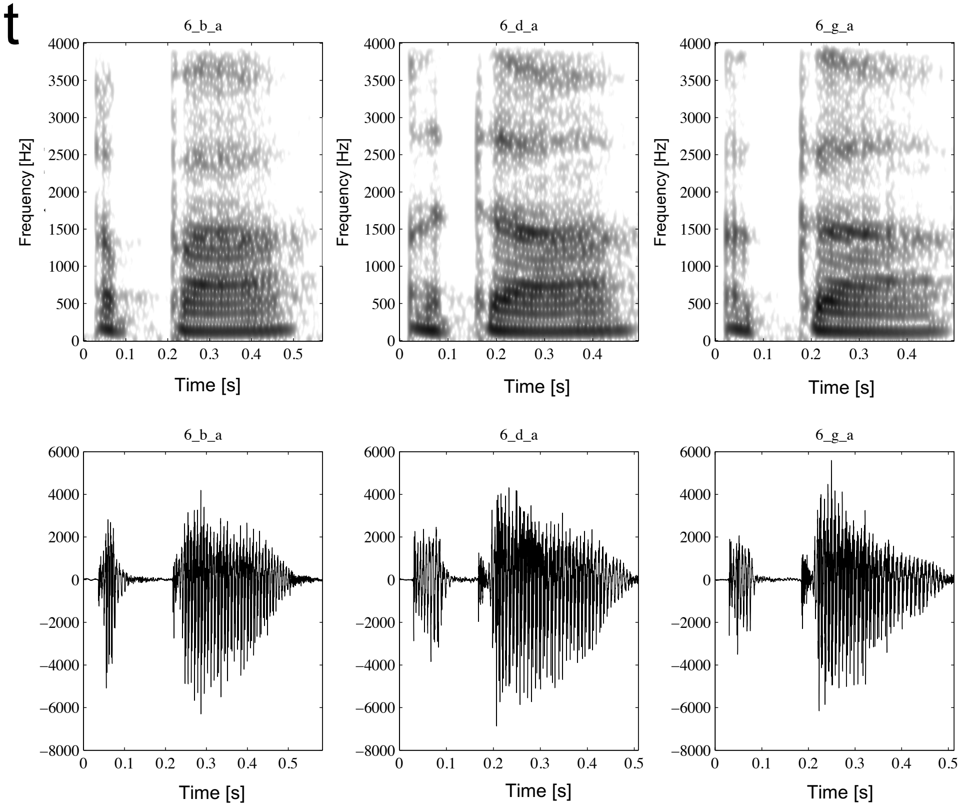


MASTER OF SCIENCE
IN ENGINEERING

Analysis of Speech Signal
Consonants

- The different types of consonants have different characterstics in the spectrogram
  - Fricatives
    - voiced/unvoiced
  - Plosives
    - Voiced/unvoiced
  - Nasals
  - Laterals
  - Others: r, R, |

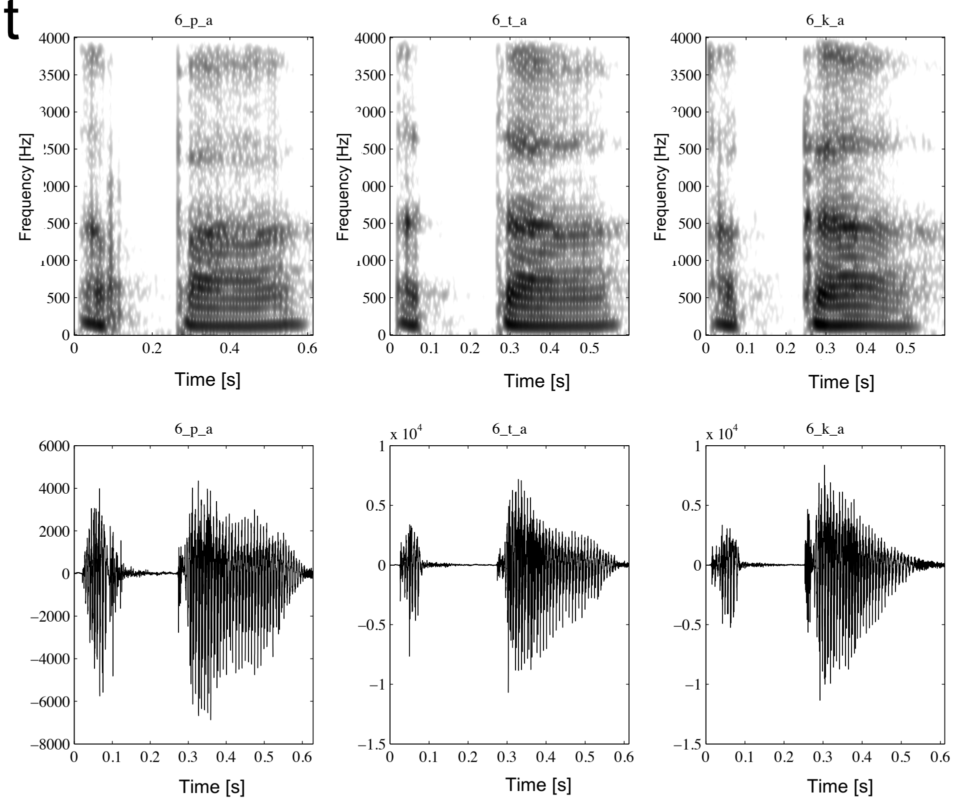
- Plosives (voiced)



Analysis of Speech Signal Consonants

- The different types of consonants have different characterstics in the spectrogram
  - Fricatives
    - voiced/unvoiced
  - Plosives
    - Voiced/unvoiced
  - Nasals
  - Laterals
  - Others: r, R, |

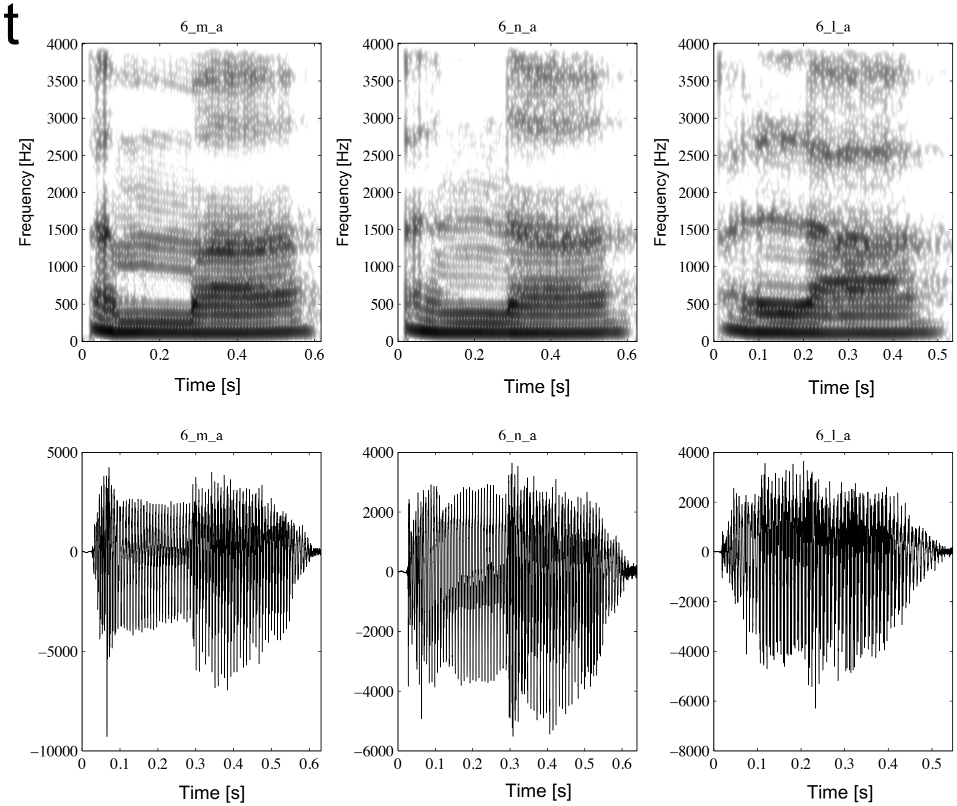
- Plosives (unvoiced)



Analysis of Speech Signal Consonants

- The different types of consonants have different characterstics in the spectrogram
  - Fricatives
    - voiced/unvoiced
  - Plosives
    - Voiced/unvoiced
  - Nasals
  - Laterals
  - Others: r, R, |

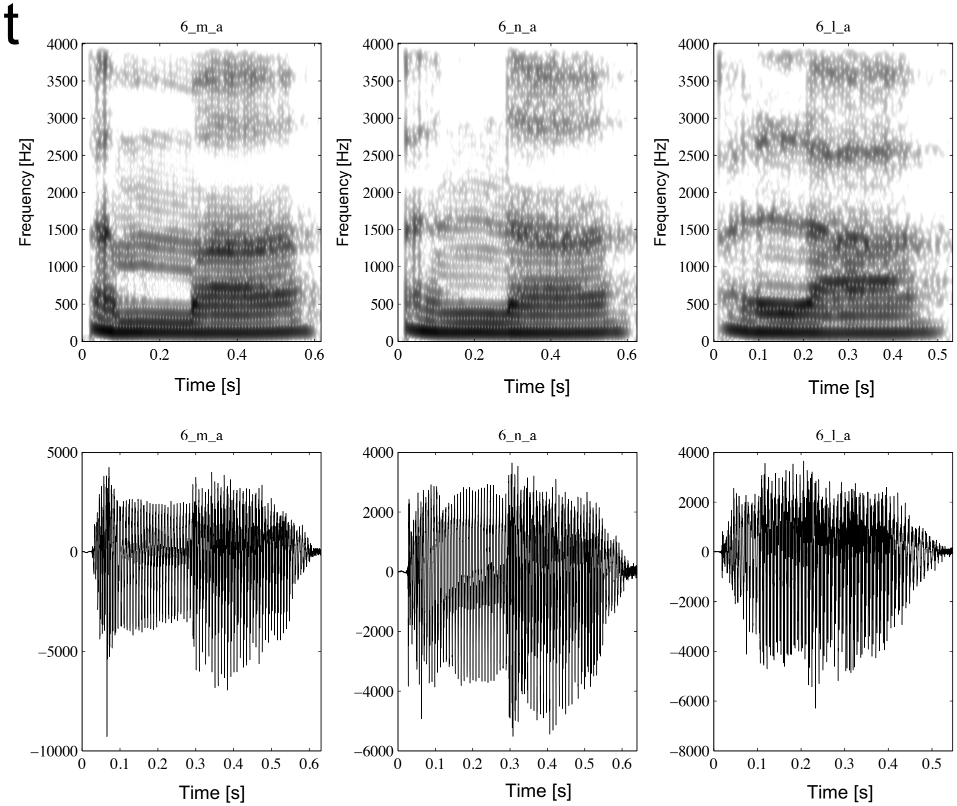
- Nasals/Laterals



Analysis of Speech Signal Consonants

- The different types of consonants have different characterstics in the spectrogram
  - Fricatives
    - voiced/unvoiced
  - Plosives
    - Voiced/unvoiced
  - Nasals
  - Laterals
  - Others: r, R, |

- Nasals/Laterals



Analysis of Speech Signal

- Vowels are easier to distinguish than consonants
- Which of the two is more important to understand what has been said?

- Exercise: Try to find out what the two sentences are:

  Th. k.ds l..rn i. th. f.rst cl.ss h.. t. r..d

  I .e.. .ou a .e..e. .i.. a .i..u.e o. .y .o.

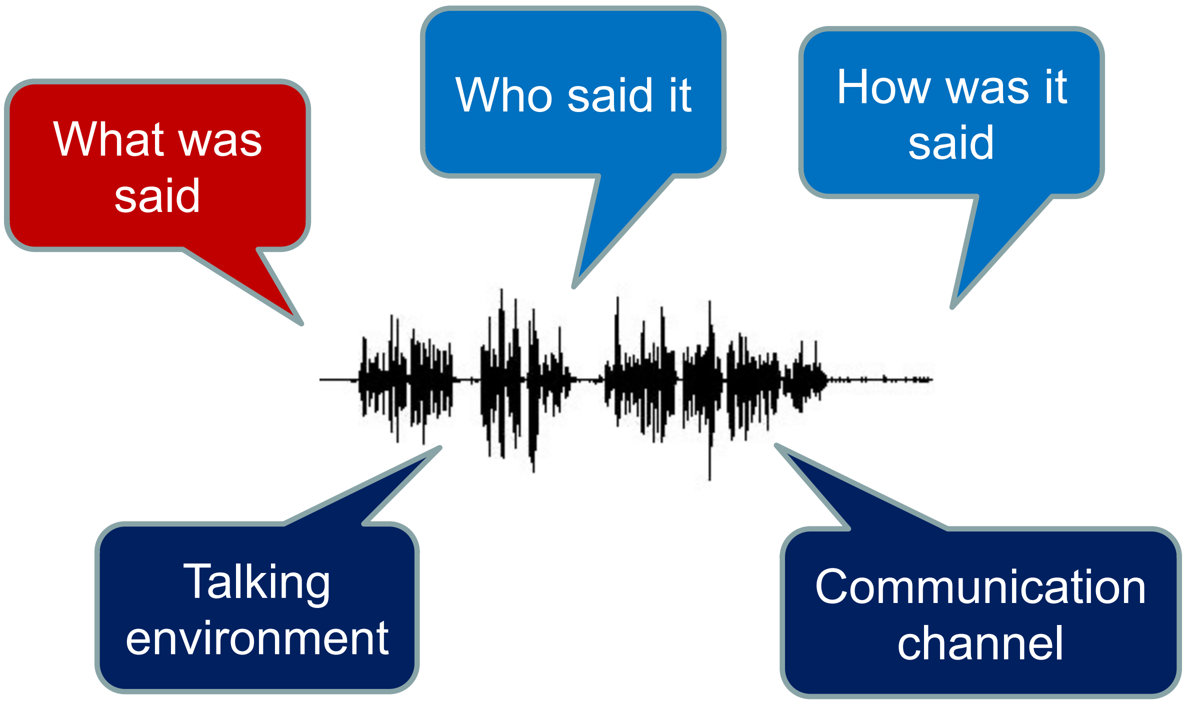
Analysis of Speech Signal Feature Extraction

- Goal of feature extraction
  - Extract all information out of the signal that is important for the given task
  - Discard all information that is irrelevant to the given task

MASTER OF SCIENCE IN ENGINEERING

Analysis of Speech Signal Feature Extraction

- Speech signal conveys a lot of information



MASTER OF SCIENCE IN ENGINEERING

Analysis of Speech Signal Feature Extraction

- Speech recognition: which features are important, which not and why?
  - Phase (time delay)?
    - Not important for SR (important for sound source localization)
  - Signal amplitude?
    - Not important for SR, only determines loudness
  - Spectrum?
    - Important, it distinguishes the phones
    - Temporal change of spectrum (spectrogram)
      -> determines phone sequence

MSE | MASTER OF SCIENCE IN ENGINEERING

Analysis of Speech Signal Feature Extraction

- Speech recognition: which features are important, which not and why?
  - What is important in the spectrogram?
    - Fundamental frequency F_0?
      - Less, conveys mainly information about the speaker less about the content (only about intonation)
    - Formants $F_1 - F_4$?
      - Yes, very important regarding phone sequence
      - What is important from the formants?
        - Number of formants
        - Center frequency positions of the formants
        - Bandwidth

Analysis of Speech Signal
Feature Extraction

- How to separate important information form unimportant information (noise)
  - Filter out unimportant information
    - In the time domain
    - In the spectral domain
  - Apply a suitable model and fit the parameters
    - Speech production model (see slide 5)
    - Speech perception model (ear model)

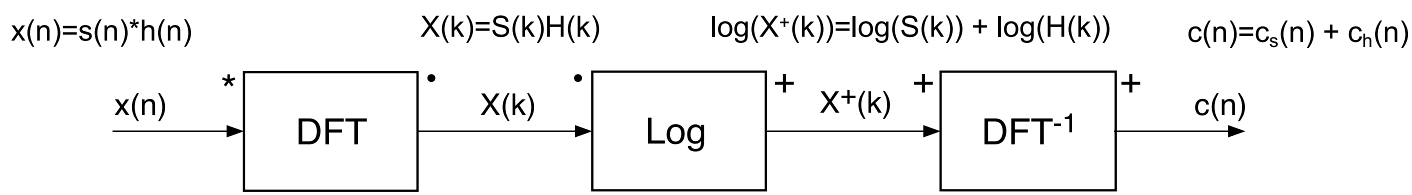
Feature Extraction
Filtering

- Time domain
  - Classic filters: LP, BP, HP, ...
  - LPC (to filter out excitation function e(n))

- Frequency domain:
  - Homomorphic filtering
  - Allows the convolution of two signals in one domain to be transformed into a summation of the two signals in the new domain.
  - Example: DFT-Cepstrum

Feature Extraction DFT-Cepstrum

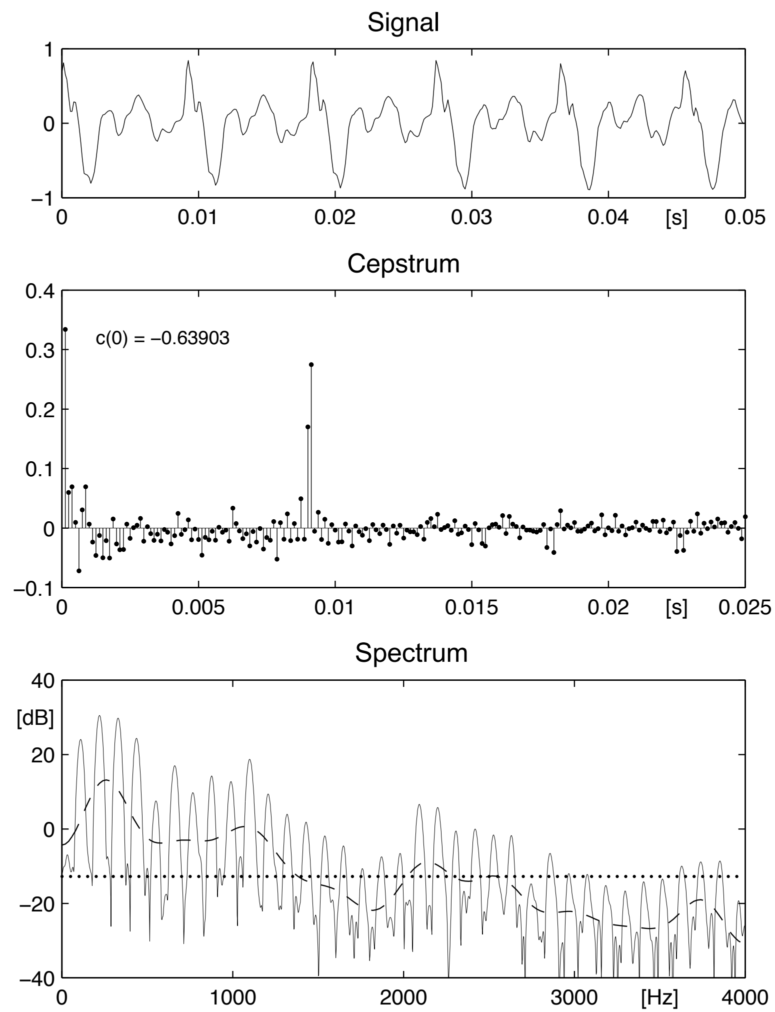
- DFT-Cepstrum c(n) is defined as the inverse DFT of the log-Spectrum of the signal x(n)



- Can be used to separate a source $s(n)$ from the filter $h(n)$
  - By filtering the cepstral coefficients of $s(n)$ or $h(n)$
- Can be used to smooth a spectrum by lowpass filtering the cepstral coefficients
- Cepstrum normally is a complex function
  - If phase of signal is unimportant → real cepstrum sufficient

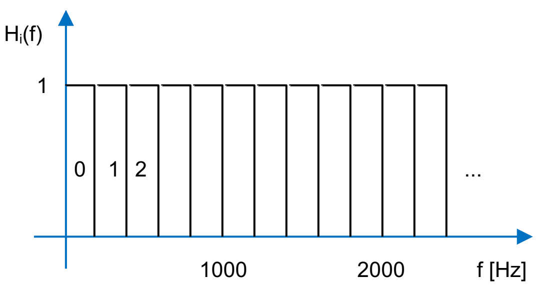
Feature Extraction
Cepstrum

- Cepstral smoothing
  - Low cepstral coefficients -> slow variations in spectrum
  - High cepstral coefficients -> fast variations in spectrum
  - Cepstral smoothing of spectrum
    - Filter out high cepstral coefficients

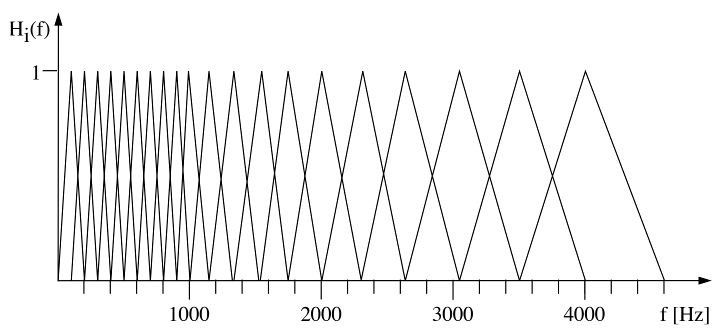


Feature Extraction
Auditory Based Features

- DFT-Spectrum
  - Is equivalent to a uniform filter bank with N filters



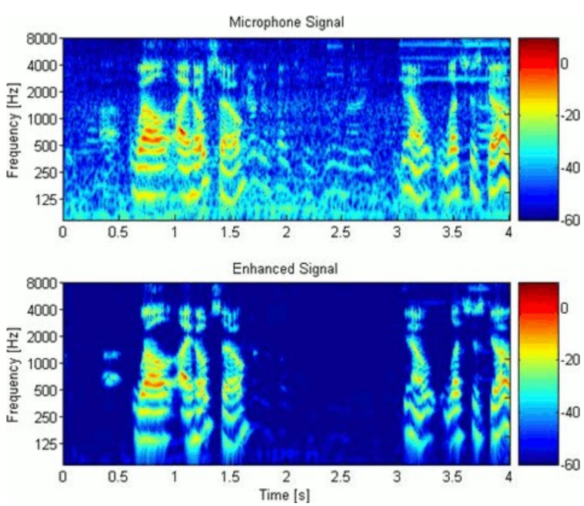
- Mel-Spectrum
  - Models basic characteristics of the ear
    - Log sensitivity of frequencies above 1 kHz
    - Exponentially increasing bandwidth of filters above 1 kHz (critical bands)

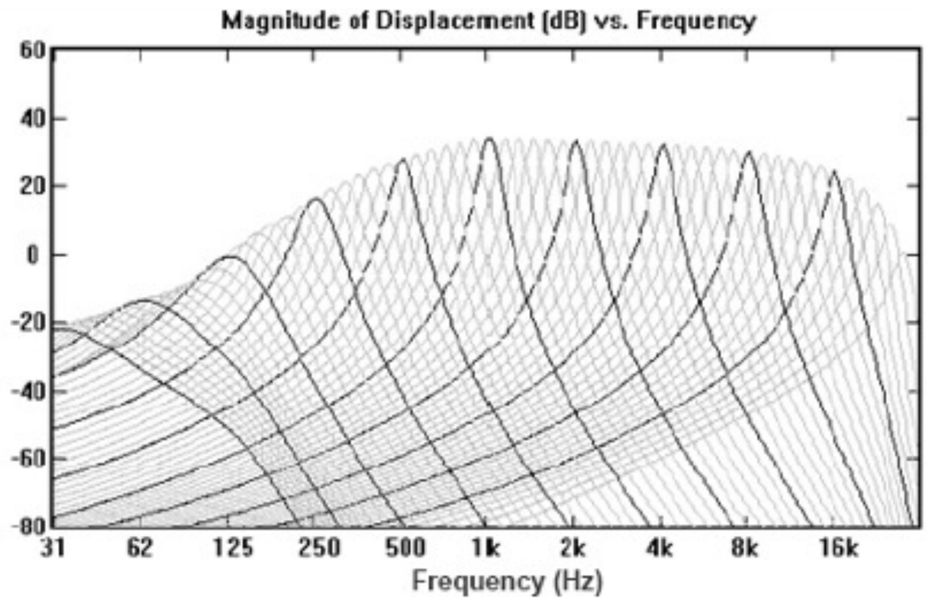


- Mel-Cepstrum
  - IDFT of Log(Mel-Spectrum)

Feature Extraction
Auditory Based Features

- Fast Cochlea Transform (FCT)
  - Audience, Inc., 2008
  - Implemented in noise-cancelling chip in many smartphones





- Proprietary modifications to Lyon's digital IIR biquad filter cascade
- Logarithmic Frequency Scale (unlike FFT)
- Optimal frequency-dependent time-frequency trade-off (unlike FFT)
- Better spectral resolution at low frequencies, better temporal resolution at high frequencies
- Critical bandwidths of human hearing built directly into transform
- Proprietary Inverse transform, low latency <20ms