9. Feature Extraction from Speech

Overview

- Learn about the most established feature extraction from speech
- Mel Frequency Cepstral Coefficients: MFCC

Quantization

- Uniform quantization:
  - 10-12 bit are sufficient to code speech
- Improvement:
  - Use distribution of amplitude values
  - μ-law:

$$f_n^{(\mu)} = f_{\max}\,\mathrm{sgn}(f_n)\frac{\log(1+\mu\frac{|f_n|}{f_{\max}})}{\log(1+\mu)} \qquad \mu \approx 200$$

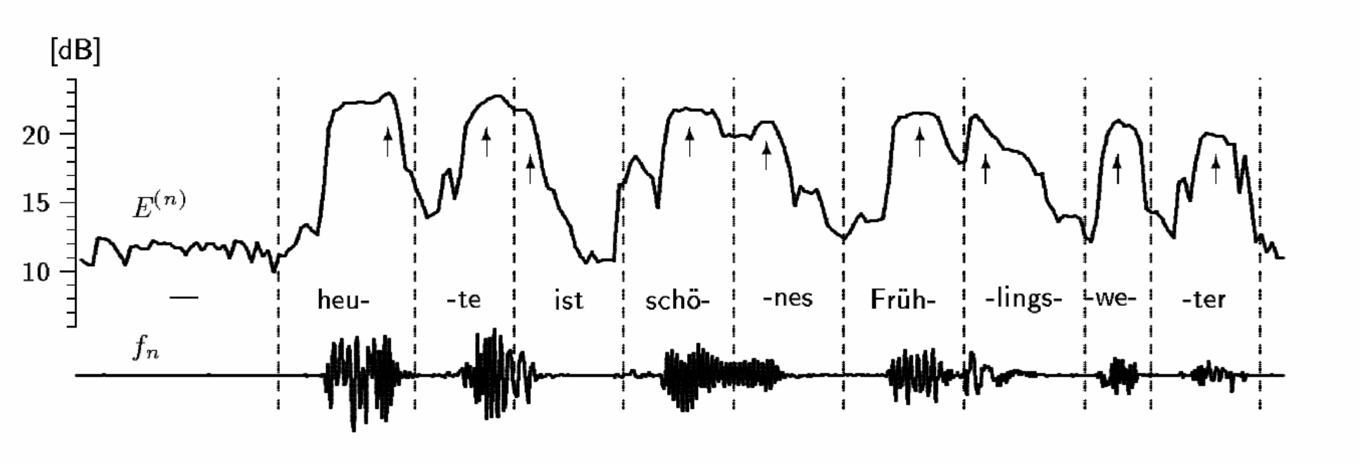
$$\propto \log(1+\mu'|f_n|)$$

Features in the Time Domain: Short-time Energy

Definition:

$$E^{(n)} = \sum_{m=0}^{M-1} | f_{m+n} |^2$$

Example:



From: Schukat-Talamazzini

Pre-emphasis

- Correct for filtering of the lips
- Iterative scheme:

$$f_n' = f_n - \alpha\, f_{n-1}$$

- Typical values: α=0.95

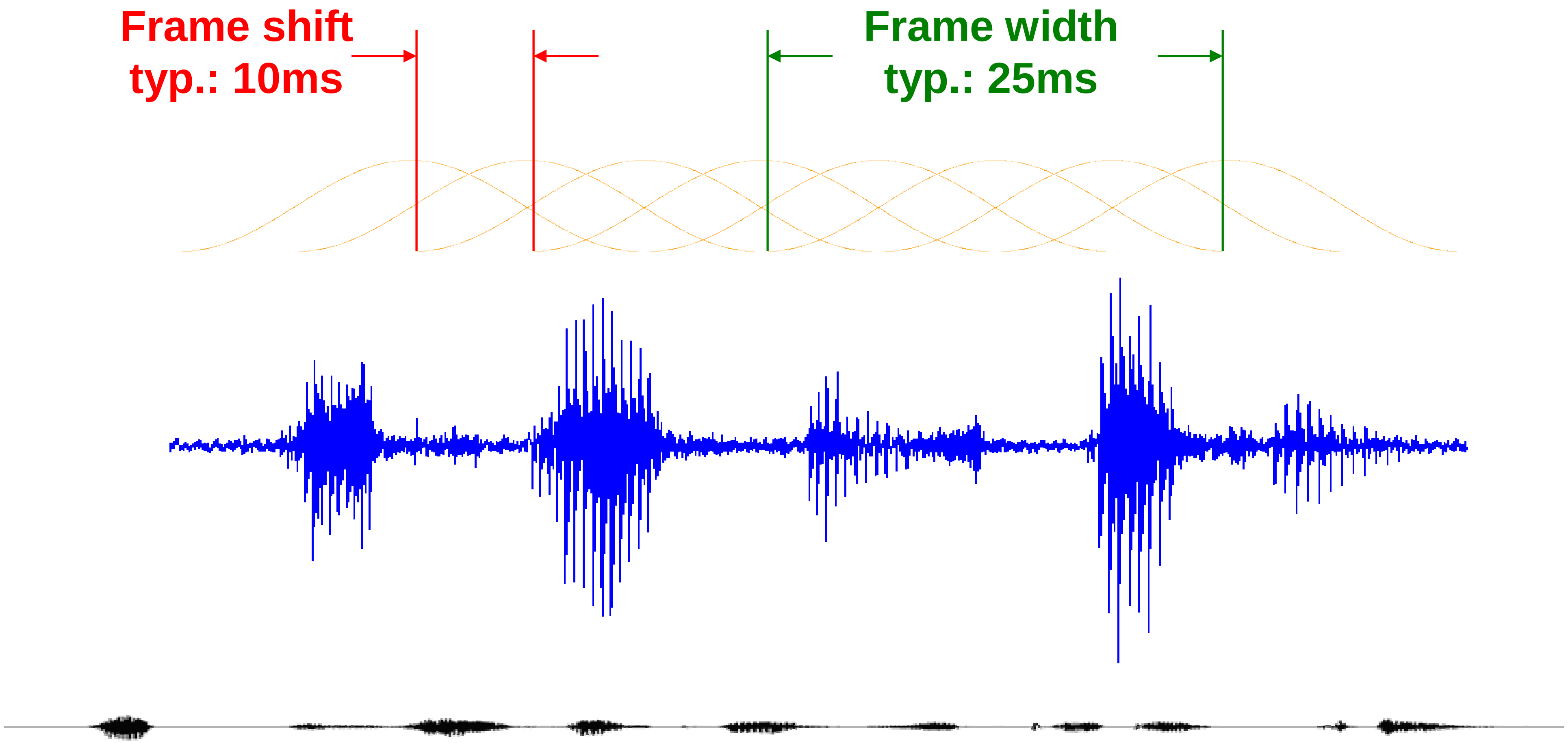
From Signal to Spectrum: Fourier Transform

- Definition

$$F^{(m)}(e^{i\omega}) = \sum_{n=-\infty}^{\infty} f_n w_{m-n} e^{-i\omega n}$$

w_n : window function
ω: frequency times 2π

Example: putting a rectangular on a speech signal



A Simple Example for Fourier Transform

$\mapsto$ Maple script "DFT.mw"

Fourier Transform in Practice

- Use "Fast Fourier Transform" (FFT)
- Requires number of samples N to be power of 2 (e.g. N=256)
- Code available
- Complexity $N \log(N)$

Established Window Functions

- Use to get sharper peaks
- Rectangular window: $w_n^R = 1$
- Generalized Hamming Window:

$$w_n^H = (1-\alpha) - \alpha \cos\left(\frac{2\pi n}{N-1}\right)$$

(α=0.54 : standard
Hamming window)

- Gauss window: $w_n^G = e^{-0.5\left(\frac{n-N/2}{3N/2}\right)^2}$
- Parabola window: $w_n^P = 4\frac{n}{N}\left(1-\frac{n}{N}\right)$

n=0...N-1

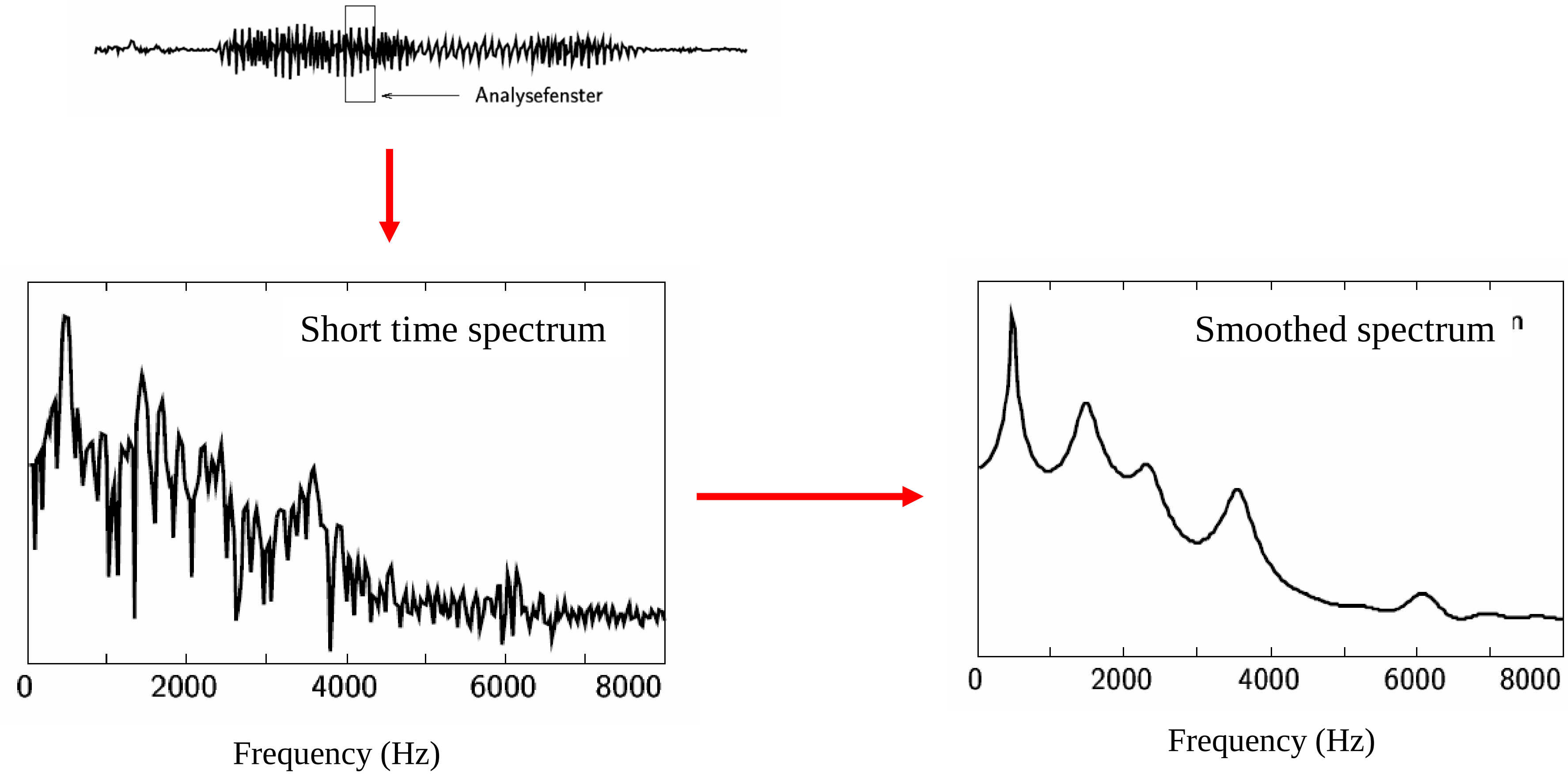
•Window functions vanish outside this interval

Rewrite of Fourier Transform

- Definition: $$F^{(m)}(e^{i\omega}) = \sum_{n=-\infty}^{\infty} f_n w_{m-n} e^{-i\omega n}$$
- Window functions vanish outside the interval n=0...N-1
- Define $\omega = 2\pi\nu \frac{1}{N}$

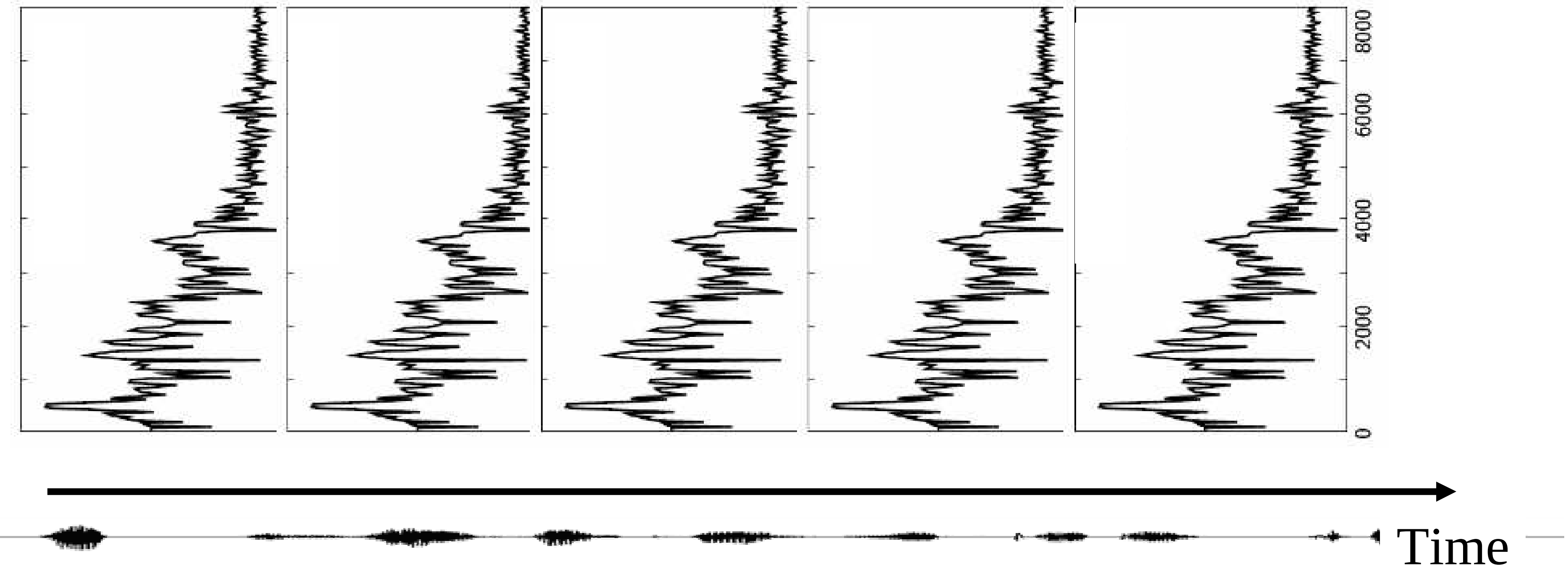
$$F_\nu^{(m)} = \sum_{n=0}^{N-1} f_{m-n} w_n e^{-i2\pi\nu\frac{n}{N}}$$

Example for ö



Spectrogram

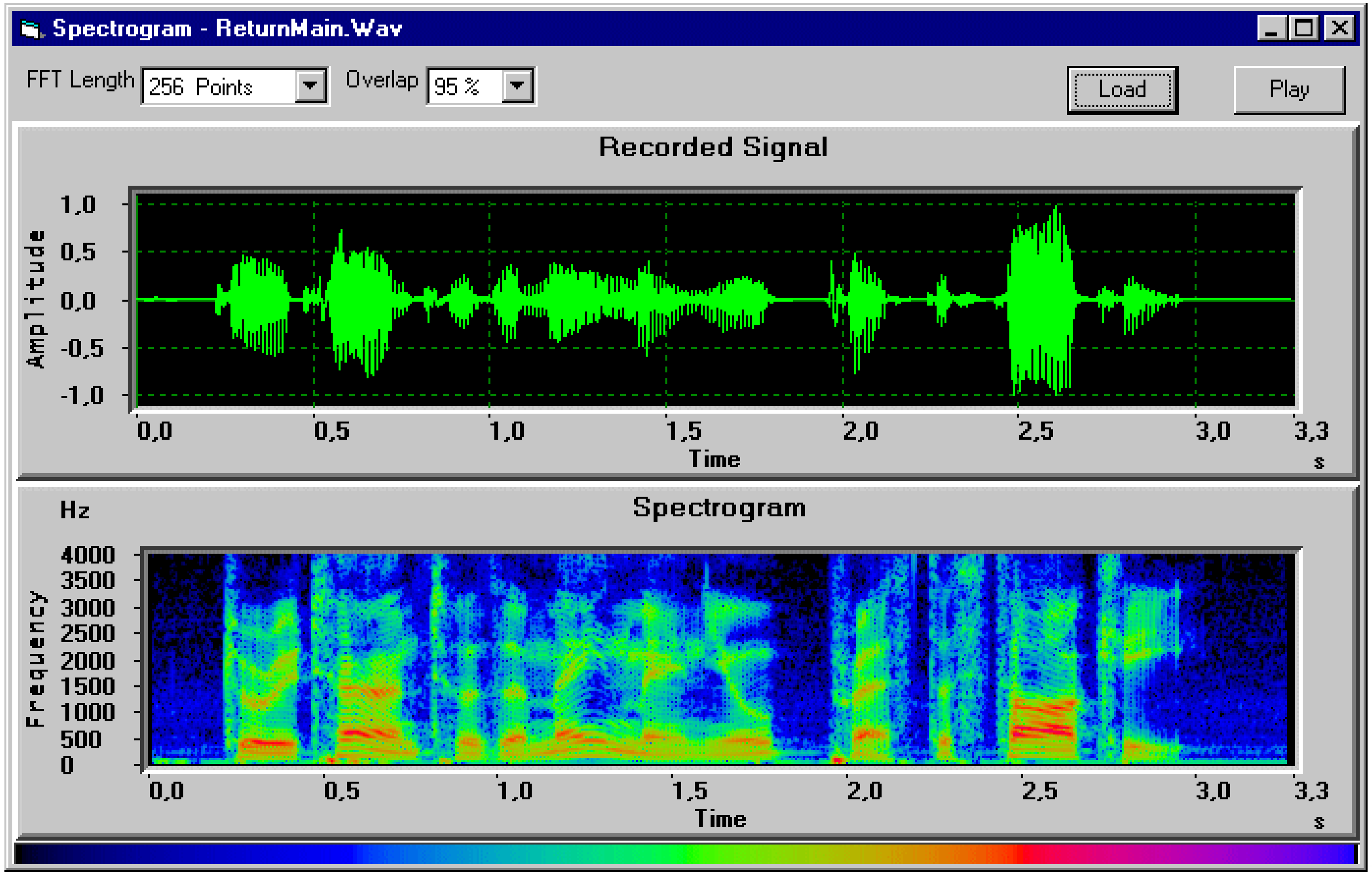
- Calculate a spectrum for any point in time
- Code the local intensity: color/grey scale



Lehrstuhl Sprachsignal Verarbeitung

Spectrogram

http://www.wilhelm-kurz-software.de/dynaplot/applicationnotes/spectrogram.htm

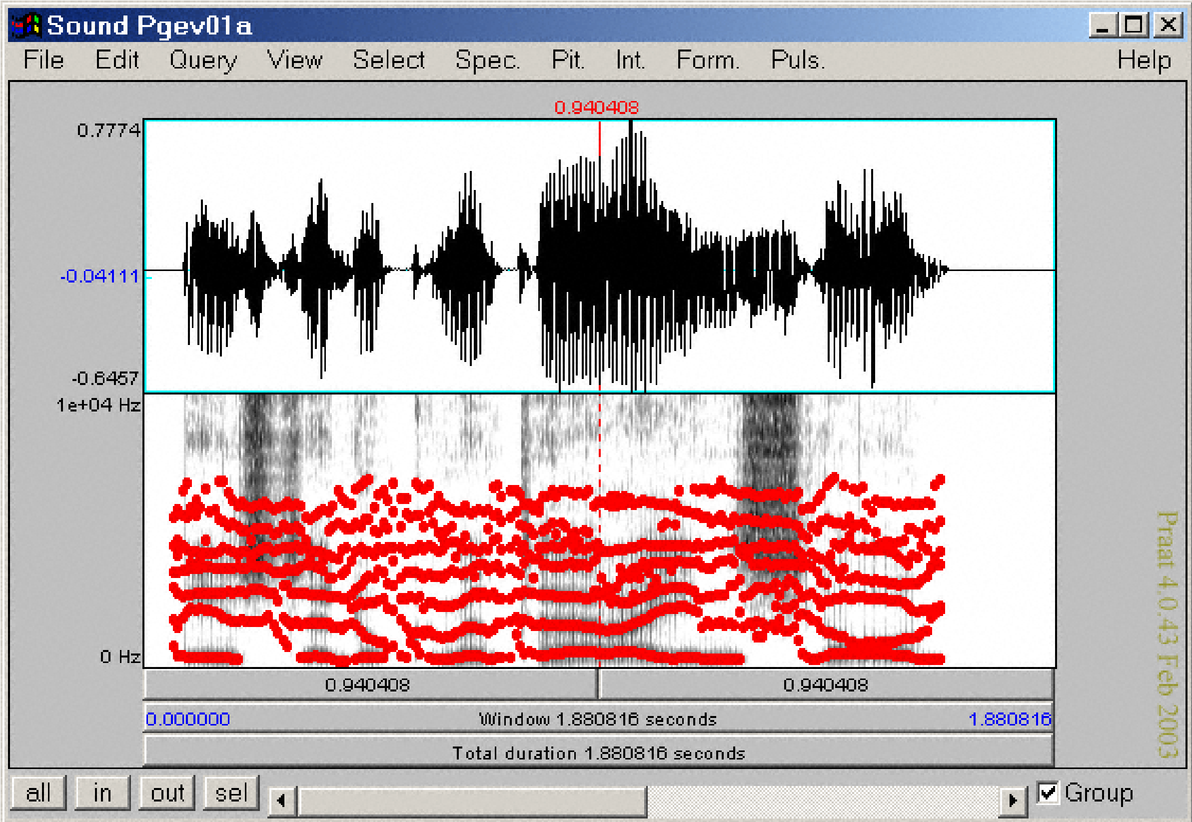


 "To return to the main menu, press the star key".

Use praat to generate a Spectrogram

- Praat: software for doing phonetics by computer
- Written by: Paul Boersma and David Weenink
- quite powerful: spectrograms, formants, pitch, …
- Download: http://www.fon.hum.uva.nl/praat/

Use praat to generate a Spectrogram



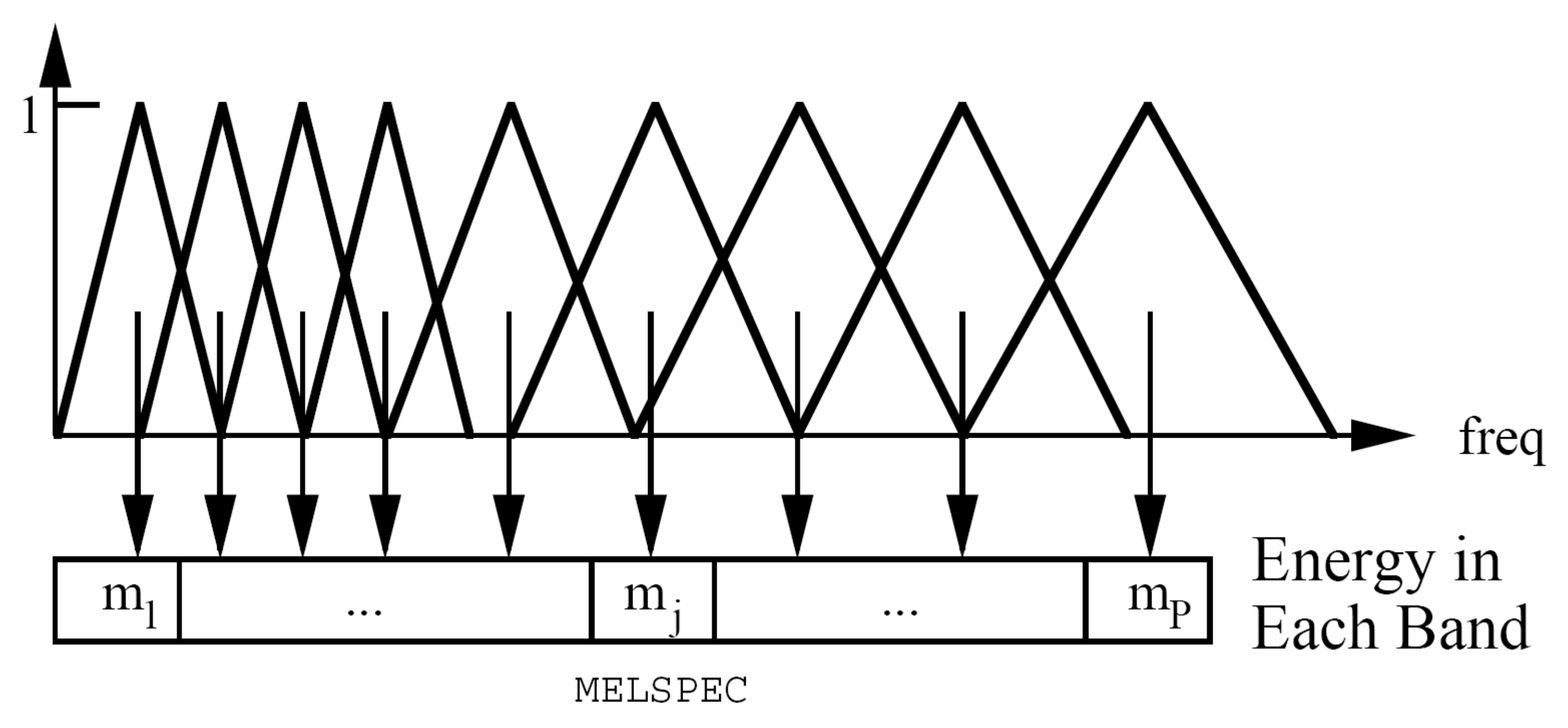
↦ demo

Smoothing the Spectrogram: Filterbank

- Idea: imitate ear
  - Do an average over neighboring frequencies
  - Scale the frequencies according to the mel or the Bark scale

  $\mapsto$ Reduction from 256 Fourier coefficients to 24 outputs of a filterbank

Example of a Filterbank



Filterbank

- Spacing of center frequency:
  - According to mel scale:

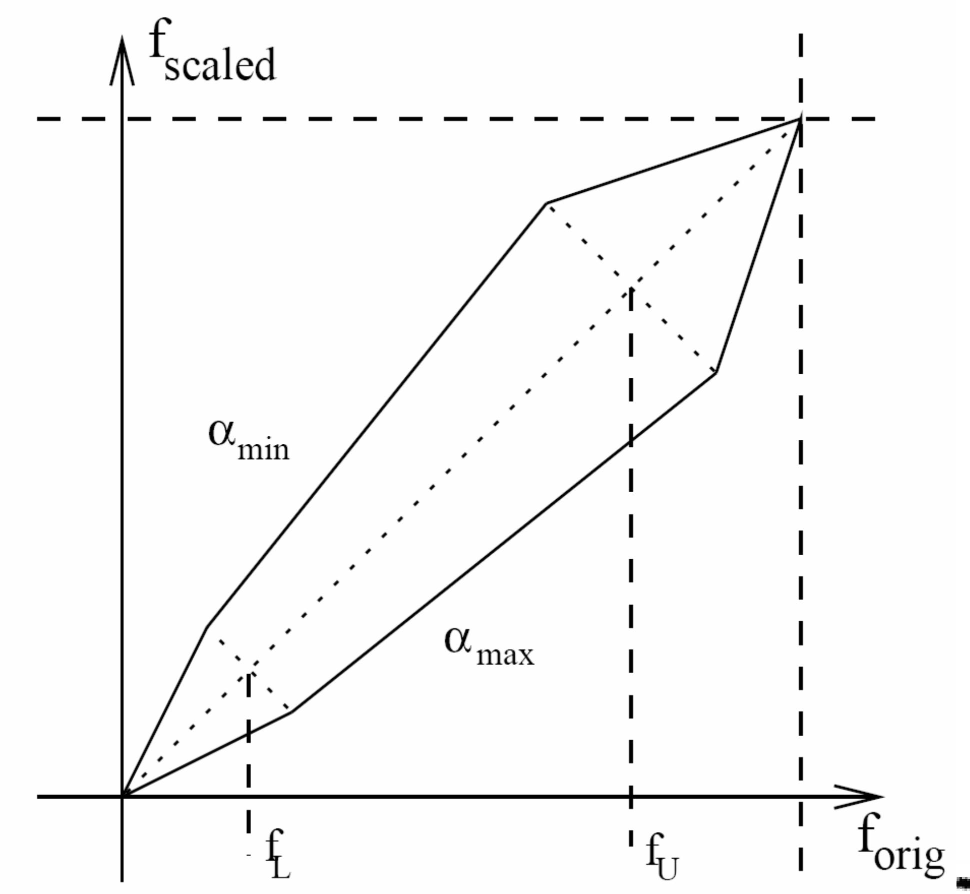
$$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700})$$

- Low frequency cut off:
  - E.g. 300 Hz (for telephone speech)
- High frequency cut off:
  - E.g. 3400 Hz (for telephone speech)
- Different settings for e.g. head set connected PC

Vocal Tract Length Normalization

- Idea:
  - Average position of formants depends on length of vocal tract
  - $\mapsto$ varying position of frequencies of filter bank
- A kind of speaker adaptation

Vocal Tract Length Normalization: Frequency Warping



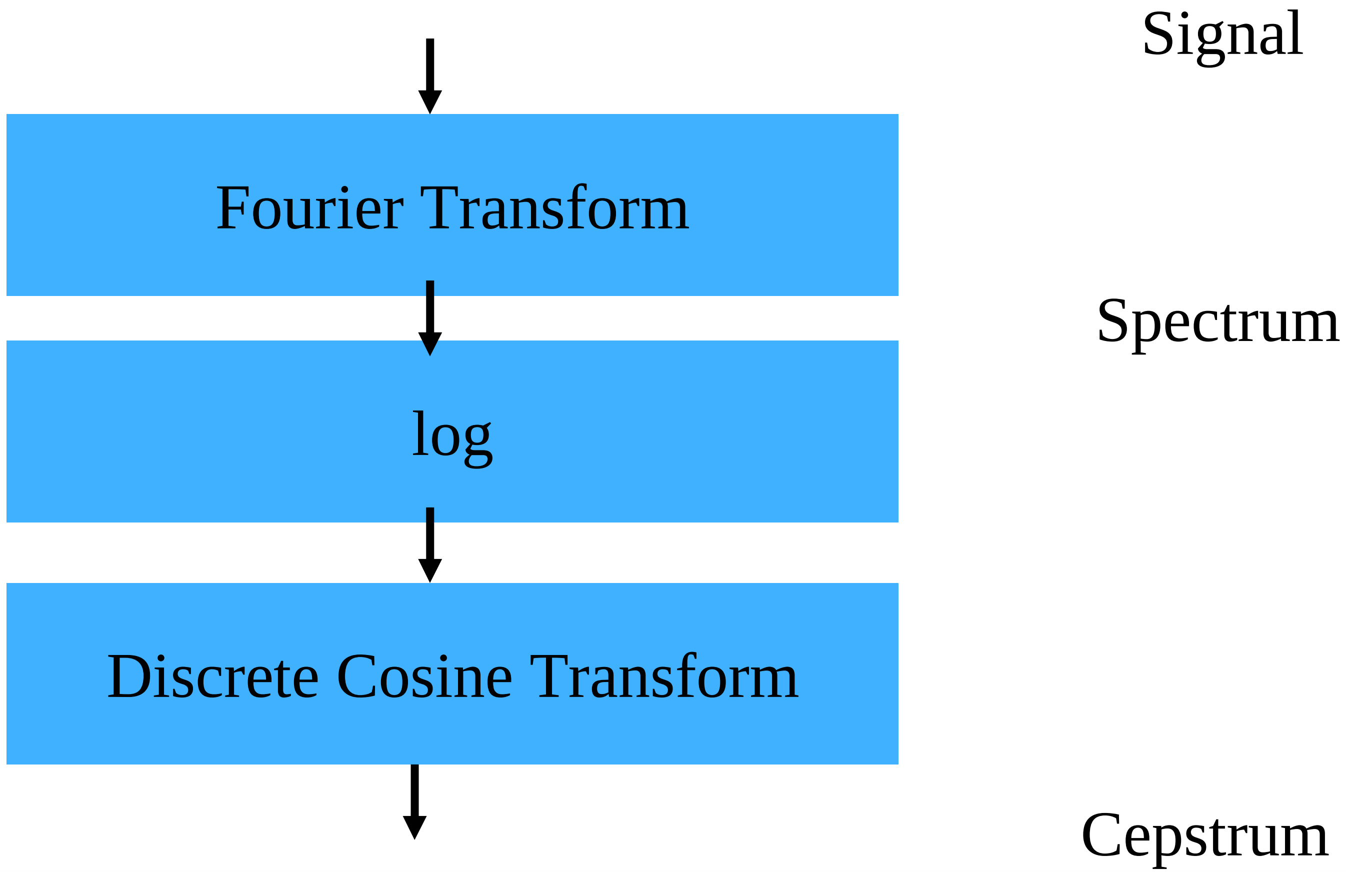
Training the Warping Factor

- Issue: how to scale for a specific speaker
- Slow version:
  - Use 11 different warping factors
  - Do speech recognition with all of them
  - Pick the best one
- Oldest approach
- Not very efficient
- Improvement: 10% less recognition errors

From Spectrum to Cepstrum

- Name: swapping of letters
- Idea: separate out the convolutional contribution
- Useful as a preparation to remove channel distortions (e.g. telephone)
- Cepstral mean subtraction (CMS)

Definition “Cepstrum”



Math for Cepstrum

- e_n: original signal (e.g. excitation from glotis)
- f_n: measured signal
- h_n: impulse response of channel (e.g. vocal tract)

$$f_n = \sum_{n=-\infty}^{\infty} h_{m-n} e_n$$

Math for Cepstrum

- Apply Fourier transform $\mathcal{F}$

$$\mathcal{F}\{f_n\} = \mathcal{F}\{\sum_{n=-\infty}^{\infty} h_{m-n} e_n\}$$

- Use convolution theorem

$$\mathcal{F}\{f_n\} = \mathcal{F}\{h_n\}\mathcal{F}\{e_n\}$$

Math for Cepstrum

- Apply logarithm

$$\log(\mathcal{F}\{f_n\}) = \log(\mathcal{F}\{h_n\}) + \log(\mathcal{F}\{e_n\})$$

- Impulse response and excitation now separated

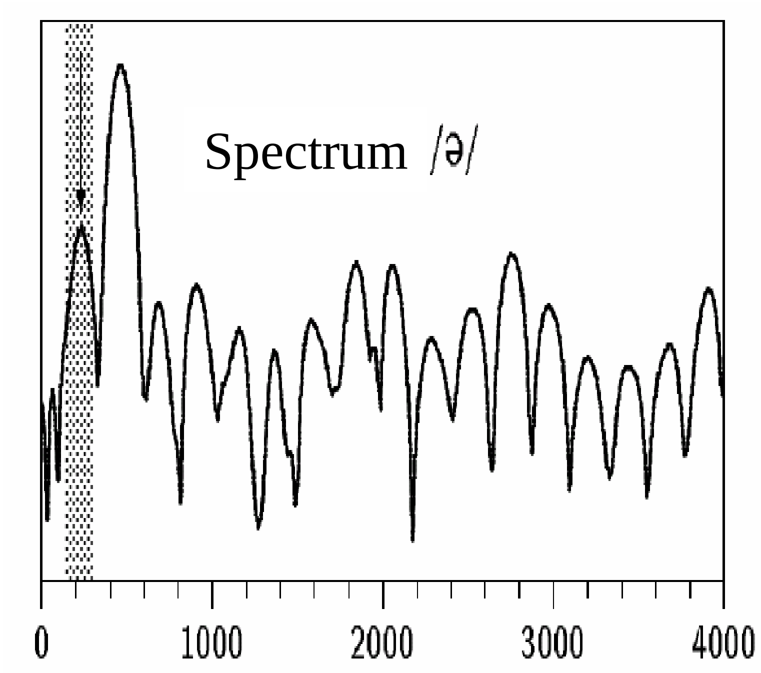
Cepstrum: do discrete cosine transform after log

- Discrete cosine transform:

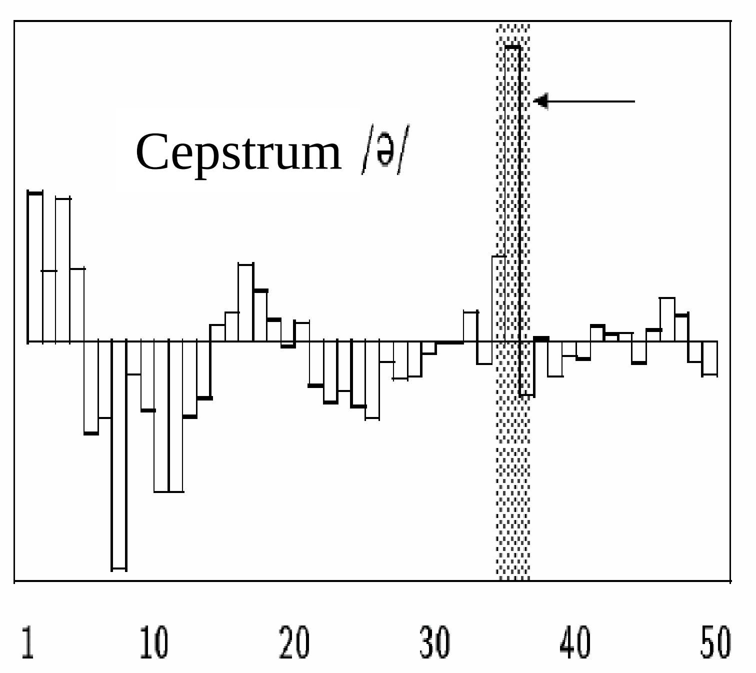
$$c_0^{(m)} = \sqrt{2/N} \sum_{\nu=0}^{N/2-1} \log(F_\nu^{(m)})$$

$$c_q^{(m)} = \sqrt{4/N} \sum_{\nu=0}^{N/2-1} \log(F_\nu^{(m)}) \cos\left(\frac{\pi q(2\nu+1)}{N}\right)$$

Use of Cepstrum I:
Identify Excitation Frequency of Glotis



Frequency (1/s)

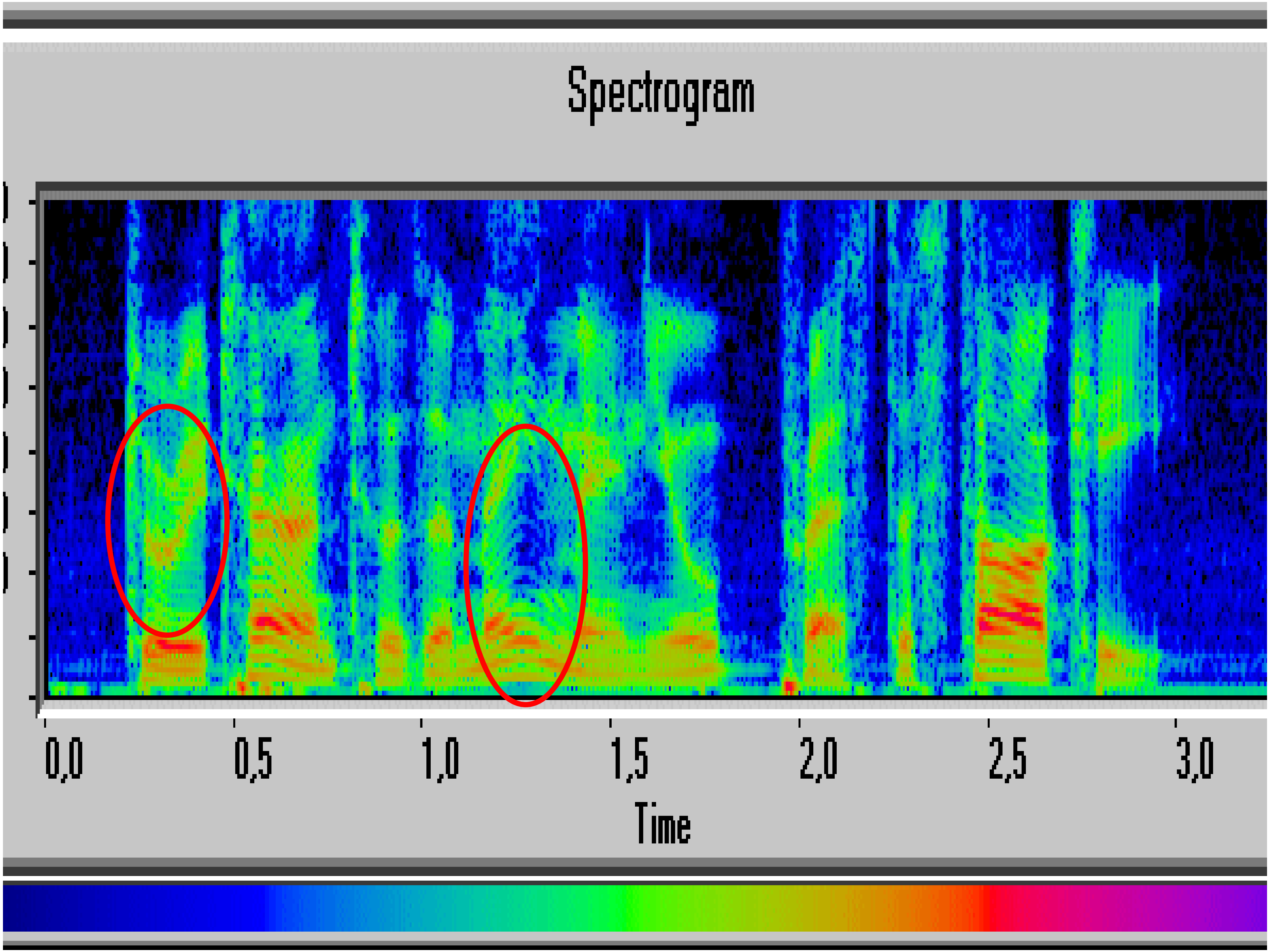


Quefrency (1/8 ms)

From: Schuckat-Talamazzini

Dynamic Features

- Spectrum captures local aspects of speech
- Window size 25 ms
- Capture slow changes in spectrum
- Other name: delta features


Spectrogram
0,0
0,5
1,0
1,5
2,0
2,5
3,0
Time

Dynamic Features

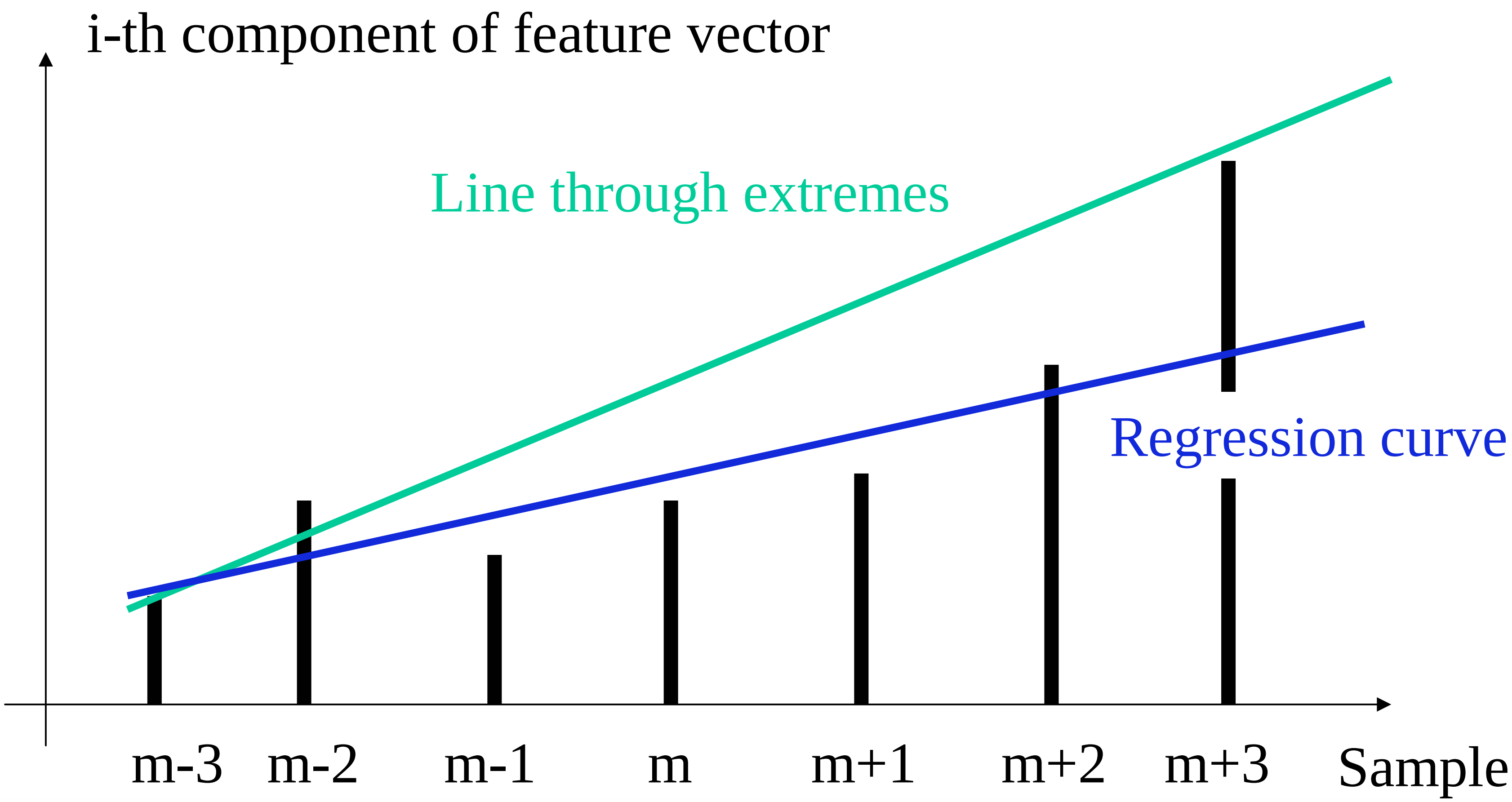
- Calculate first and second derivatives
- Naïve approach to first derivative
  - Continuous function

$$\frac{df(t)}{dt} \approx \frac{f(t+\Delta t) - f(t-\Delta t)}{2\Delta t}$$

  - Time discrete sampling

$$\frac{df(t_m)}{dt} \approx \frac{f(t_{m+\Delta}) - f(t_{m-\Delta})}{2\Delta + 1}$$

Difference/Regression



Regression Formula

$$\frac{df(t)}{dt} = \frac{\sum_{i=1}^{M} i\left(f(t_{m+i}) - f(t_{m-i})\right)}{\sum_{i=1}^{M} i^2}$$

- Check M=1

Dynamic Features

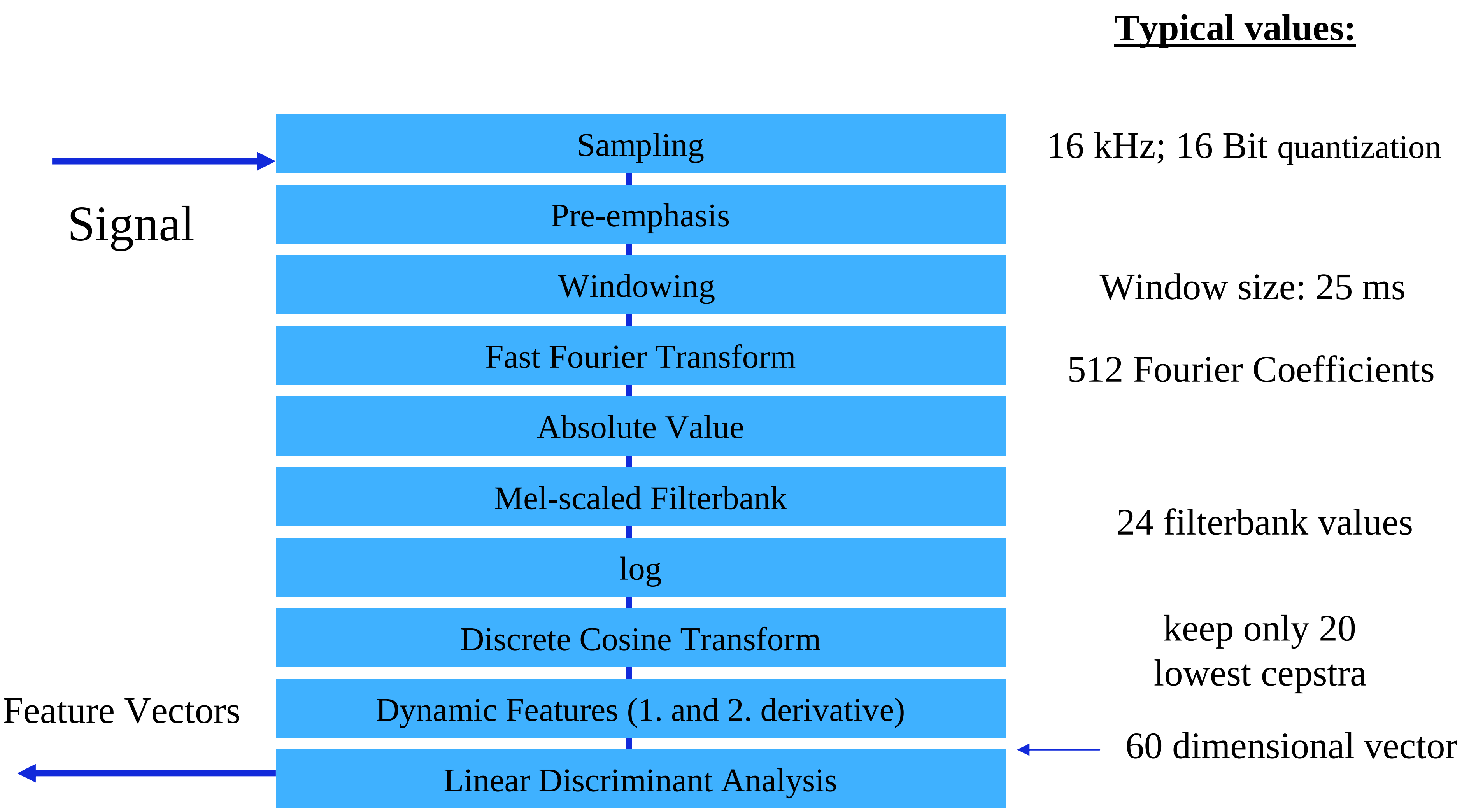
- Invented by Furui 1981
- Standard in any modern ASR system

- Alternative:
  - Linear mapping of neighboring feature vectors
- Issue:
  - Dimension of feature vectors

Linear Discriminant Analysis

- Method to decrease size of feature vector
- Maximize severability of class regions
- Linear transform of feature vectors
- More: later in the lecture

Complete Pipeline for
Mel-Frequency Cepstral Coefficients (MFCC)



Alternative Feature Extraction Methods

- LP-Cepstrum (LP=linear prediction)
  - Derived from speech coding
  - No longer much in use
- PLP (=Perceptual linear prediction)
  - For certain applications popular
  - Claim: mode noise robust than MFCCs
  - Main change: us $|.|^{1/3}$ instead of log in MFCC

Summary

- Classical “plain vanilla” feature extraction:
  Mel-Frequency Cepstral Coefficients
- Main deficiency: not very noise robust
- Used in
  - Speech Recognition
  - Speaker Recognition
  - Music genre classification