Advanced Signal Processing Seminar SS 2010

Feature Extraction

Lukas Pfeifenberger

Signal Processing and Speech Communication Laboratory

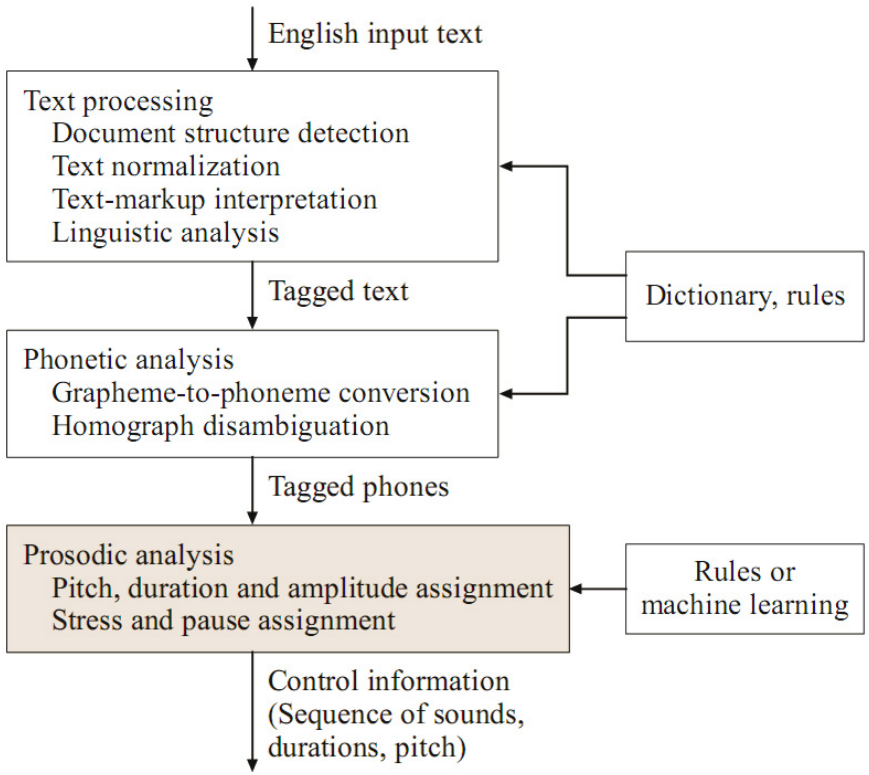
20.06.2010

Outline

- **Feature Extraction:**
  Why and What
- **Speech signal analysis:**
  TD: Short-time energy, Zero-Crossings
  FD: STFT, Cepstrum, LPC, LSF
- **Estimation of Pitch and Vocal tract information:**
  MFCC, PLP
- **Cepstral Smoothing:**
  RASTA
- **Phone Clustering:**
  classification methods

Why and What

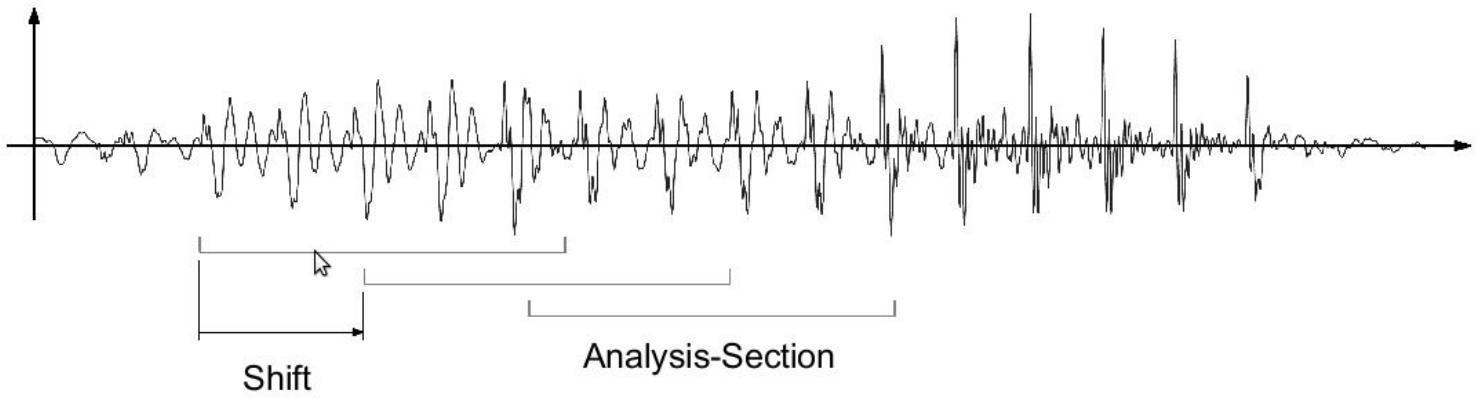
- **Feature Extraction for a TTS system:**



Analyse prosody of speech for synthesizing phones from a given input text.

Speech signal analysis

- **Analysis window:**

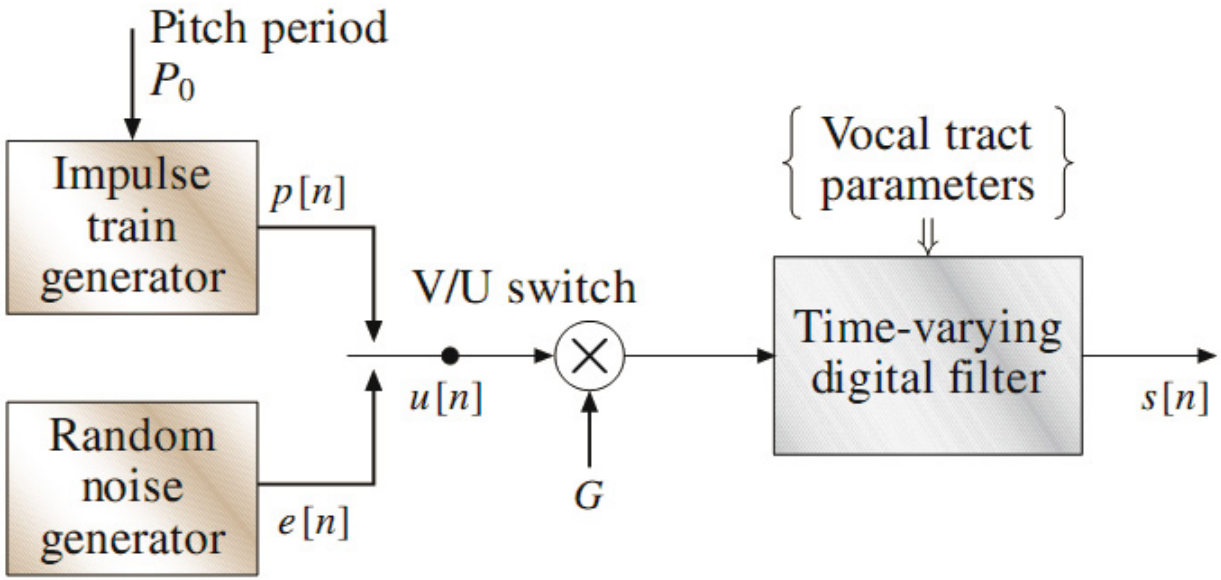


Compromise between non-stationarity and frequency resolution. Typically 25ms frame length and 10ms overlap.

Window functions used: *Hanning* or *Hamming*

Speech signal analysis

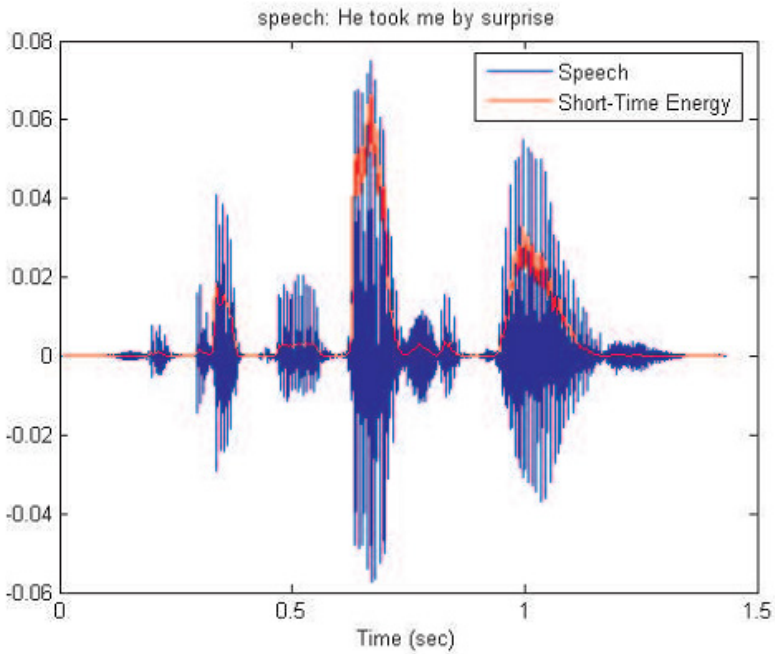
- **Speech production model:**

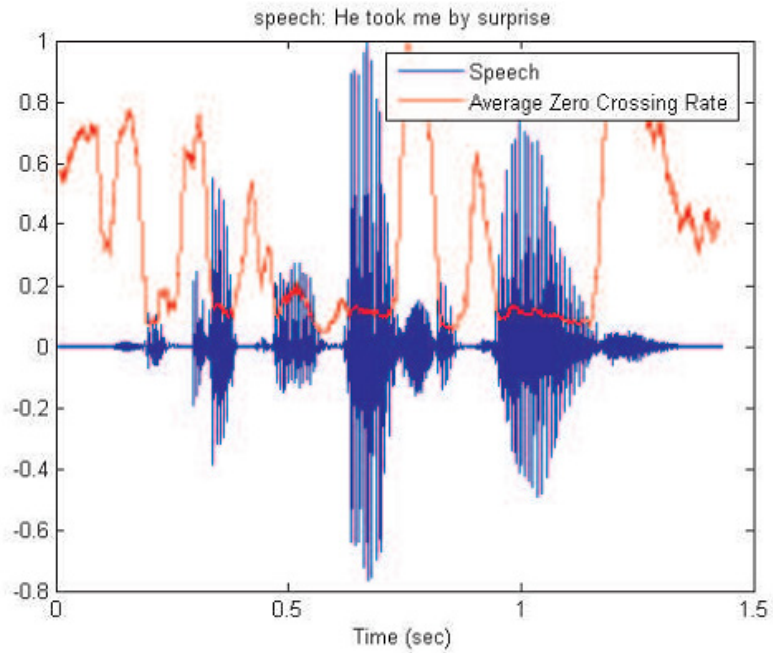


Pulses are used for voiced sounds, white noise for fricatives and a post-filter models the vocal tract: $s[n] = u[n] * h[n]$

Speech signal analysis

- **TD methods:**





$$E_n = \sum_{-M \leq m \leq M} [x(m) * w(n-m)]^2$$

$$Z_n = \sum_{-M \leq m \leq M} |sgn[x(m)] - sgn[x(m-1)]| * w(n-m)$$

Could be used to separate voiced from unvoiced sounds , or as very simple VAD, but extremely noise sensitive.

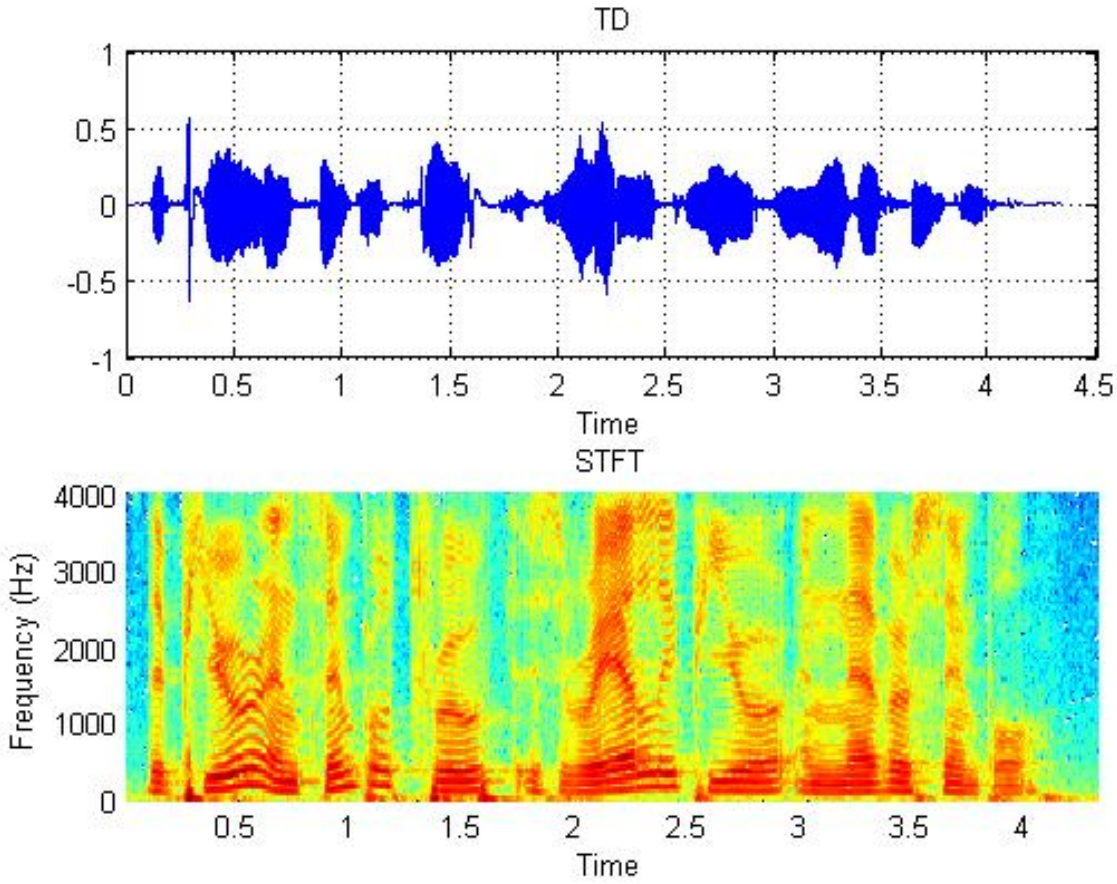
Speech signal analysis

- **FD methods:**

STFT:

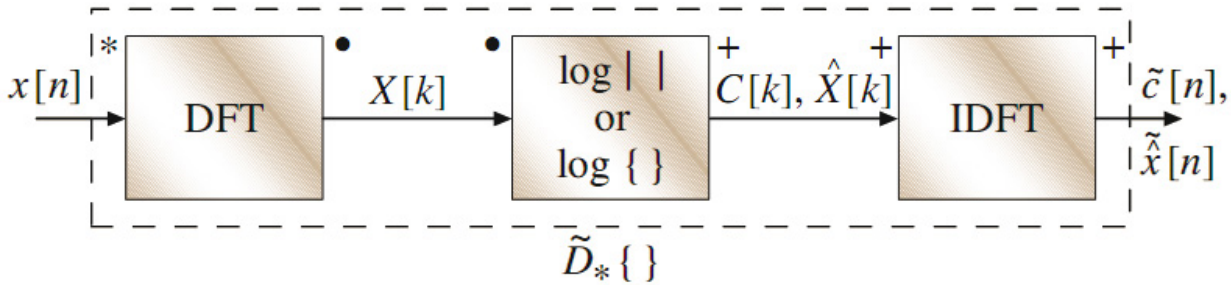
$$X[k,l] = \sum_{n=0}^{N-1} x_l[n]\, e^{-i2\pi nk/K}$$
$$= \sum_{n=0}^{N-1} w[n]x[n+lL]\, e^{-i2\pi nk/K}$$

L ... 10ms
N ... 25ms



Speech signal analysis

- **Real cepstrum:**



$$c[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |X(e^{i\omega})| \, e^{i\omega n} \, d\omega$$

Speech production model:

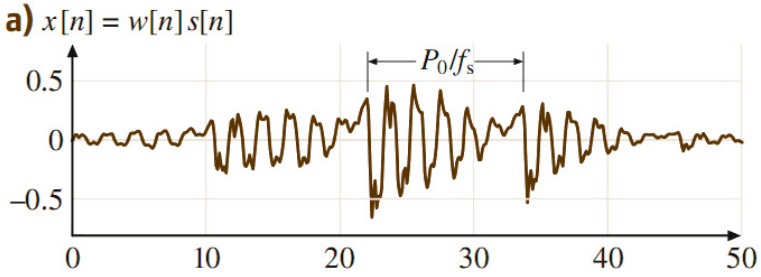
$$x[n] = u[n] * h[n]$$
$$X(e^{j\omega}) = U(e^{j\omega})H(e^{j\omega})$$
$$\hat{X}(e^{j\omega}) = log\{U(e^{j\omega})\} + log\{H(e^{j\omega})\}$$

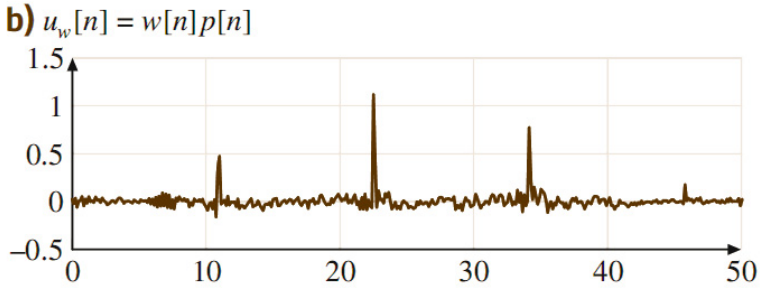
Speech signal analysis

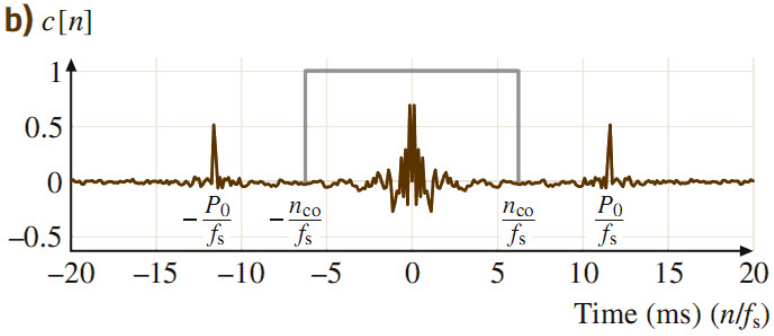
- **liftering:**

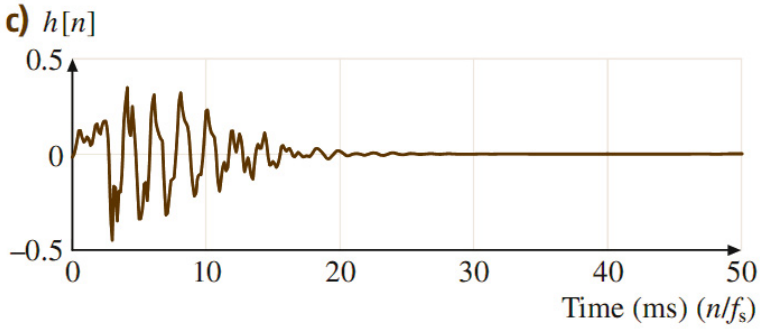
We can separate the pitch from the vocal tract filter by liftering:

$$\hat{Y}(e^{i\omega}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{X}(e^{i\theta})\hat{L}(e^{i(\omega-\theta)})\,d\theta$$





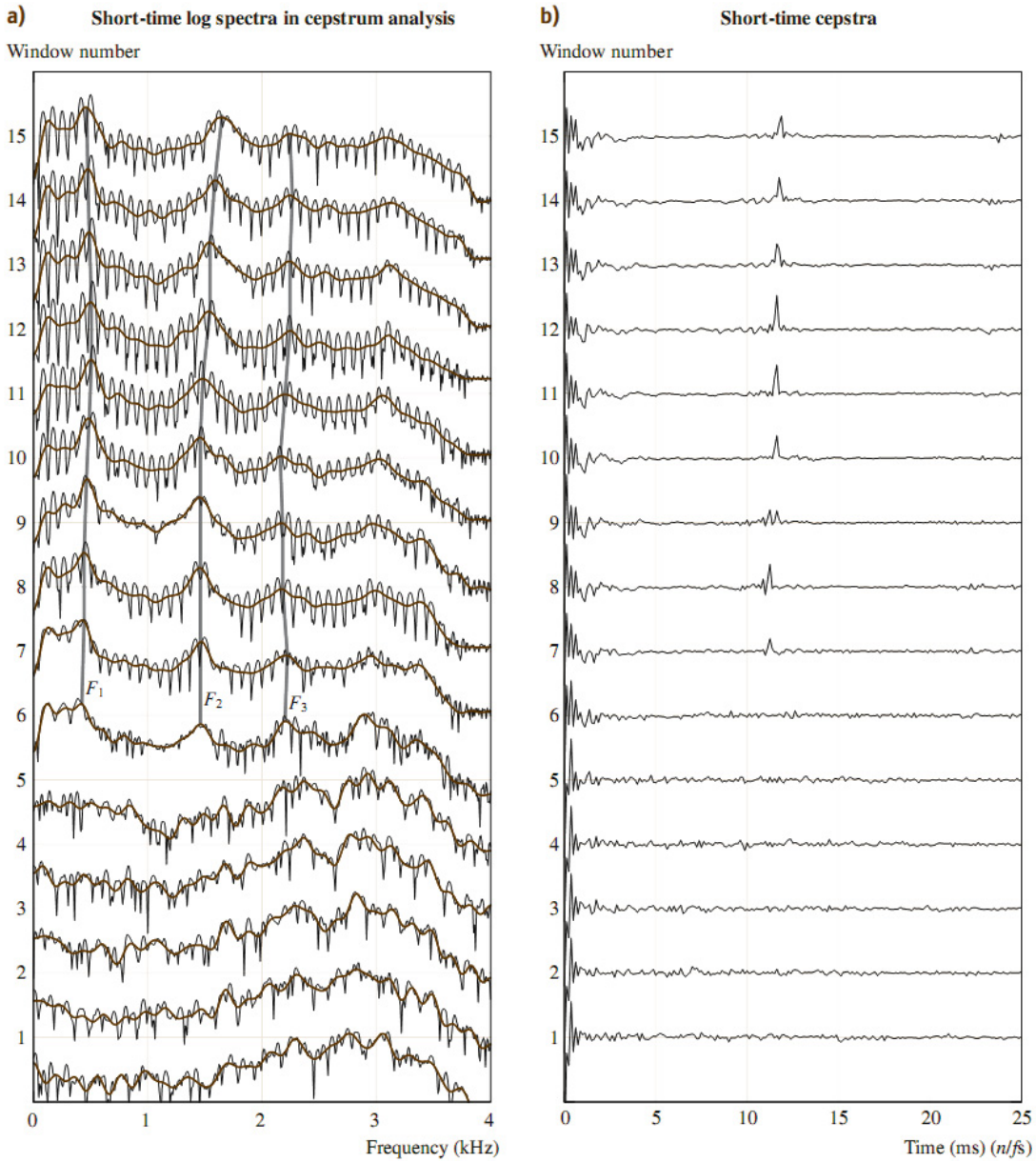




Short-time-cepstra

evolution of pitch and formants:

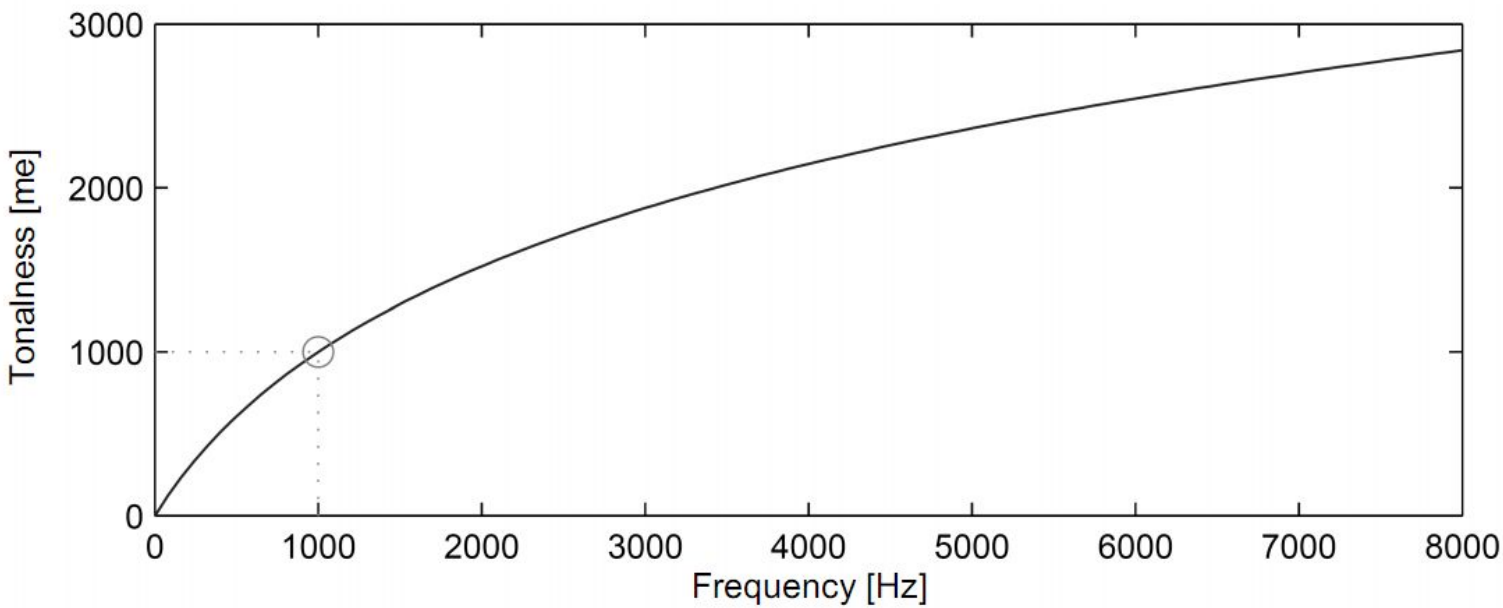
→ Matlab demo



Cepstral smoothing

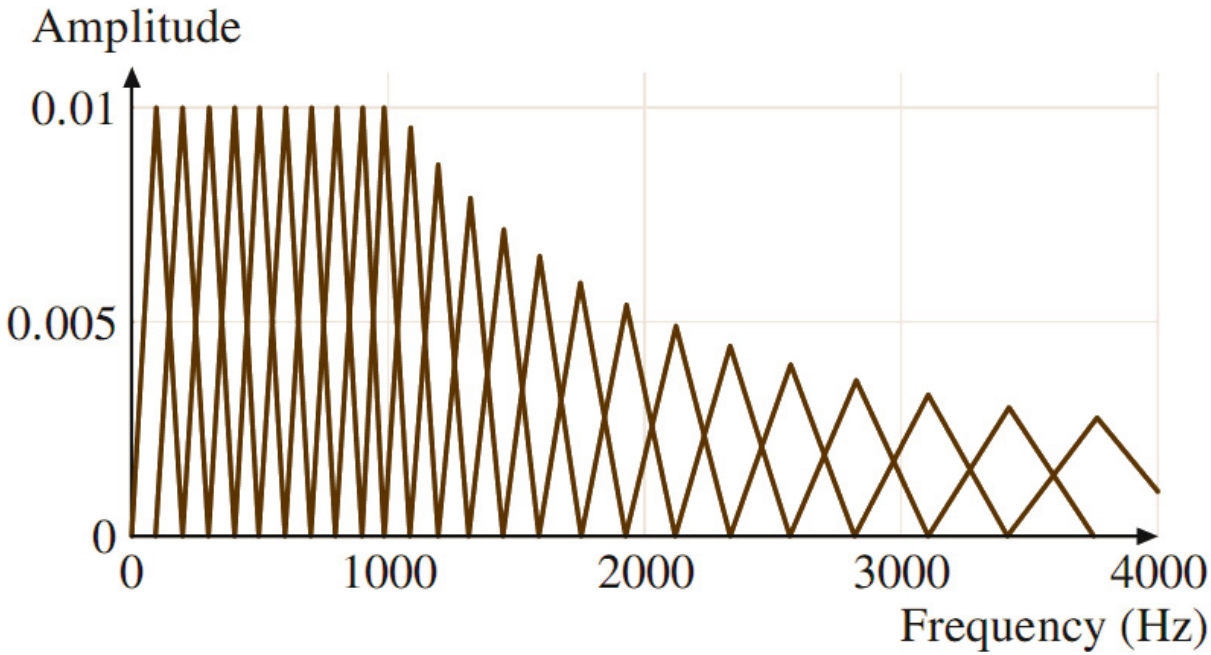
- **Problem:**
  - Cepstral coefs are sensitive to noise sensitive to noise
  - High computational load
- **Idea:**
  - Scale frequencies according to the Mel or Bark scale.
  - Average over neighboring frequencies.
  - Reduction of e.g. 256 fourier coefs to 24 outputs of a mel-scaled filter bank.
  - Calculation of 24 MFCCs.

Mel scale



$$f[mel] = 2595 * log_{10}(1 + \frac{f[Hz]}{700})$$

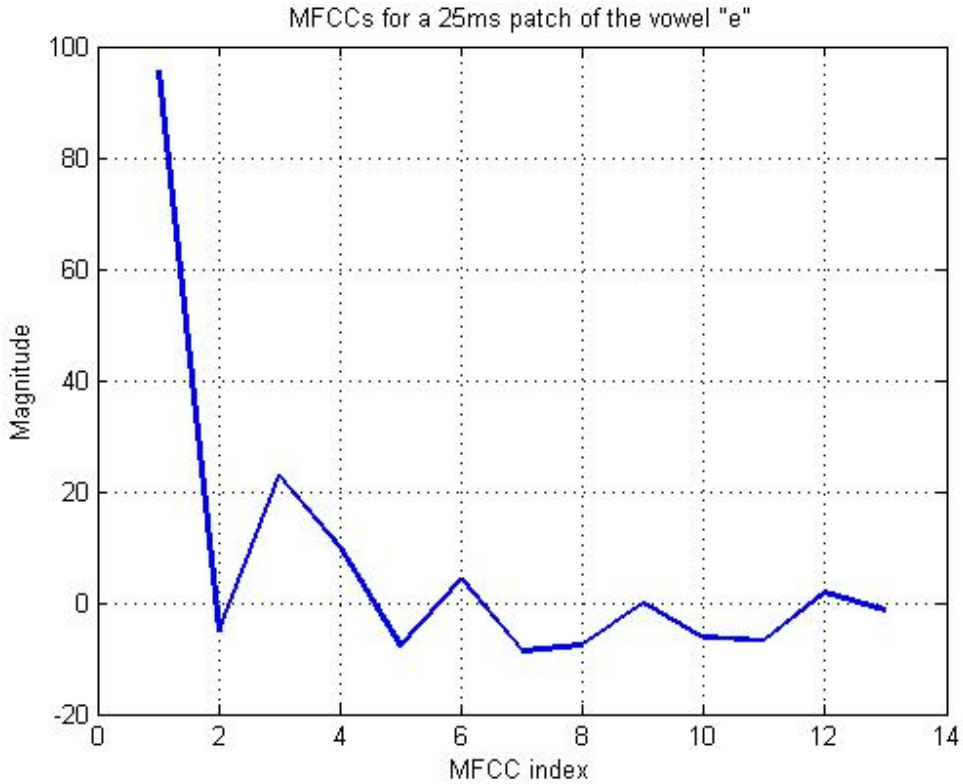
Mel scaled filterbank



$$\mathrm{MF}_m[r] = \frac{1}{A_r} \sum_{k=L_r}^{U_r} |V_r[k] X_m[k]|^2$$

Normalized weighing functions produce a flat mel-spectrum from a flat fourier-spectrum.

Mel-Frequency Cepstrum Coefcients

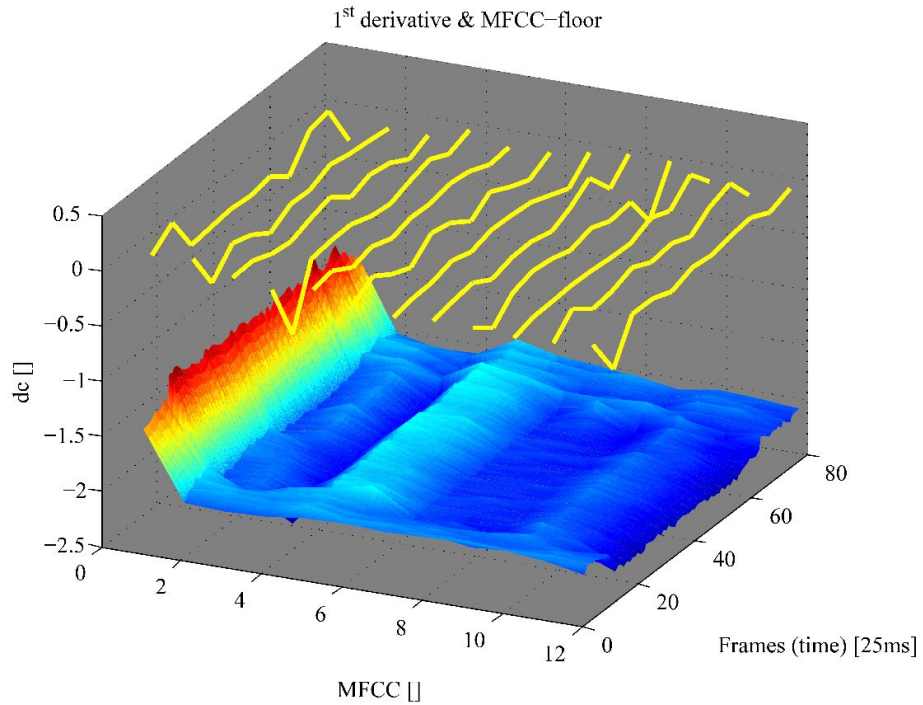


$$\text{mfcc}[n] = \frac{1}{R}\sum_{r=1}^{R}\log\left(\text{MF}_m[r]\right)\cos\left[\frac{2\pi}{R}\left(r+\frac{1}{2}\right)n\right]$$

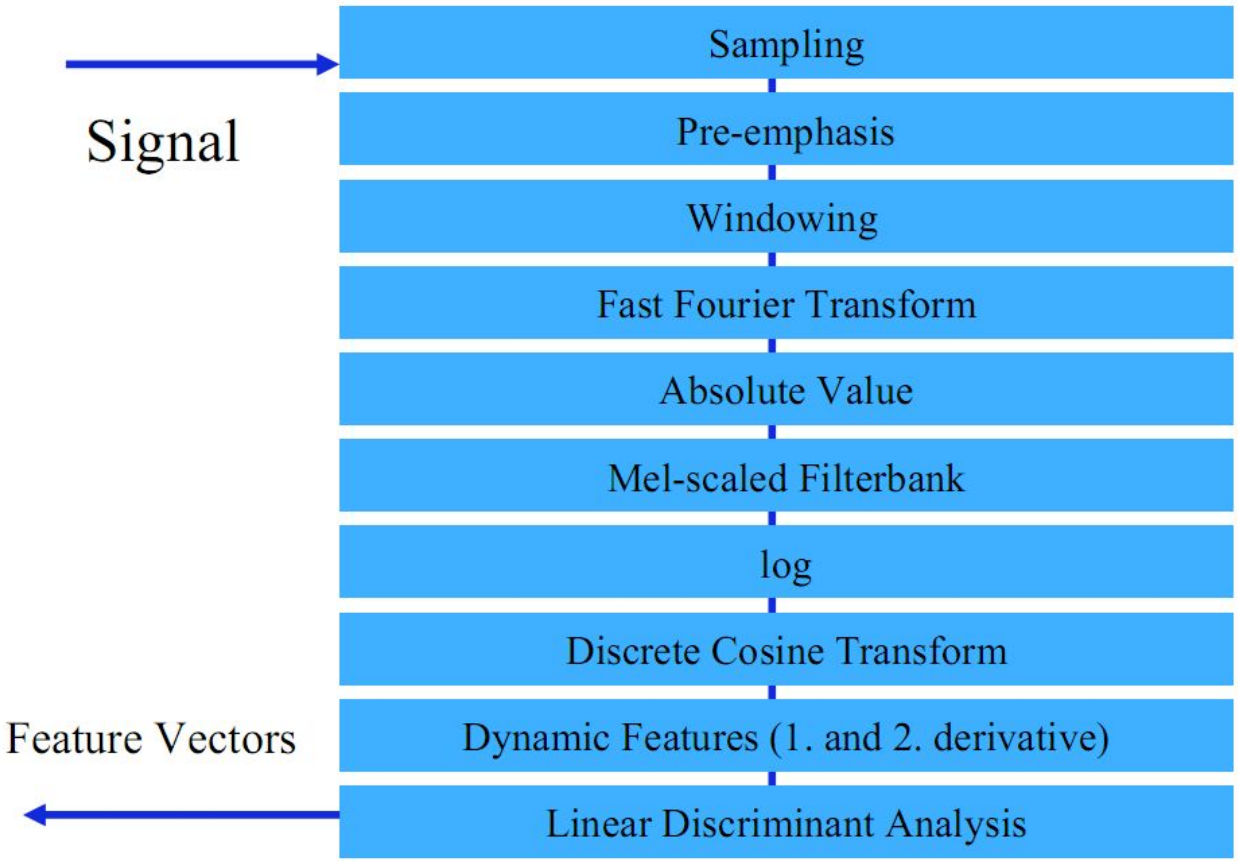
delta-MFCCs

1st and 2nd order derivatives are used as additional features.
Usually they are obtained by linear regression:

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



MFCC summary



Perceptual linear prediction

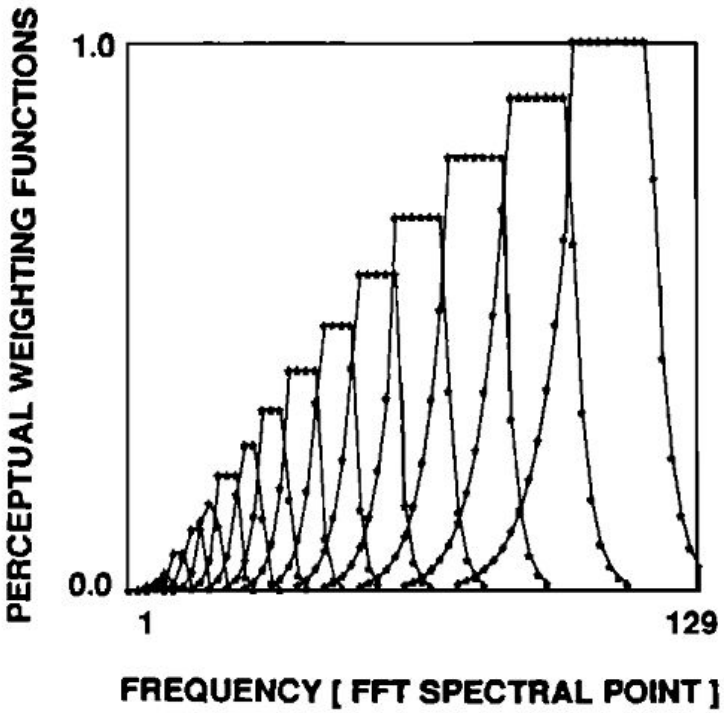
- **more robust against noise than MFCC:**
- do the STFT $\rightarrow |\cdot|^2$
- spectrum-warping to bark-scale
- loudness-preemphasis based upon human hearing
- loudess-to-density conversion by taking the cube-root
- do the IDFT, this yields the autocorrelation sequence
- solve the Yule-Walker equation to get the LP-coefs a_l:

$$x(k) = \sum_{l=1}^{L} a_l x(k-l) + Gu(k)$$

- LP-coefs can be quantized into LSFs for speech coding

Perceptual linear prediction

- very similar to cepstrum-based methods
- main differences:
  - $\sqrt[3]{\cdot}$ instead of $log()$
  - perceptual weighting functions:



Relative Spectra (RASTA)-PLP

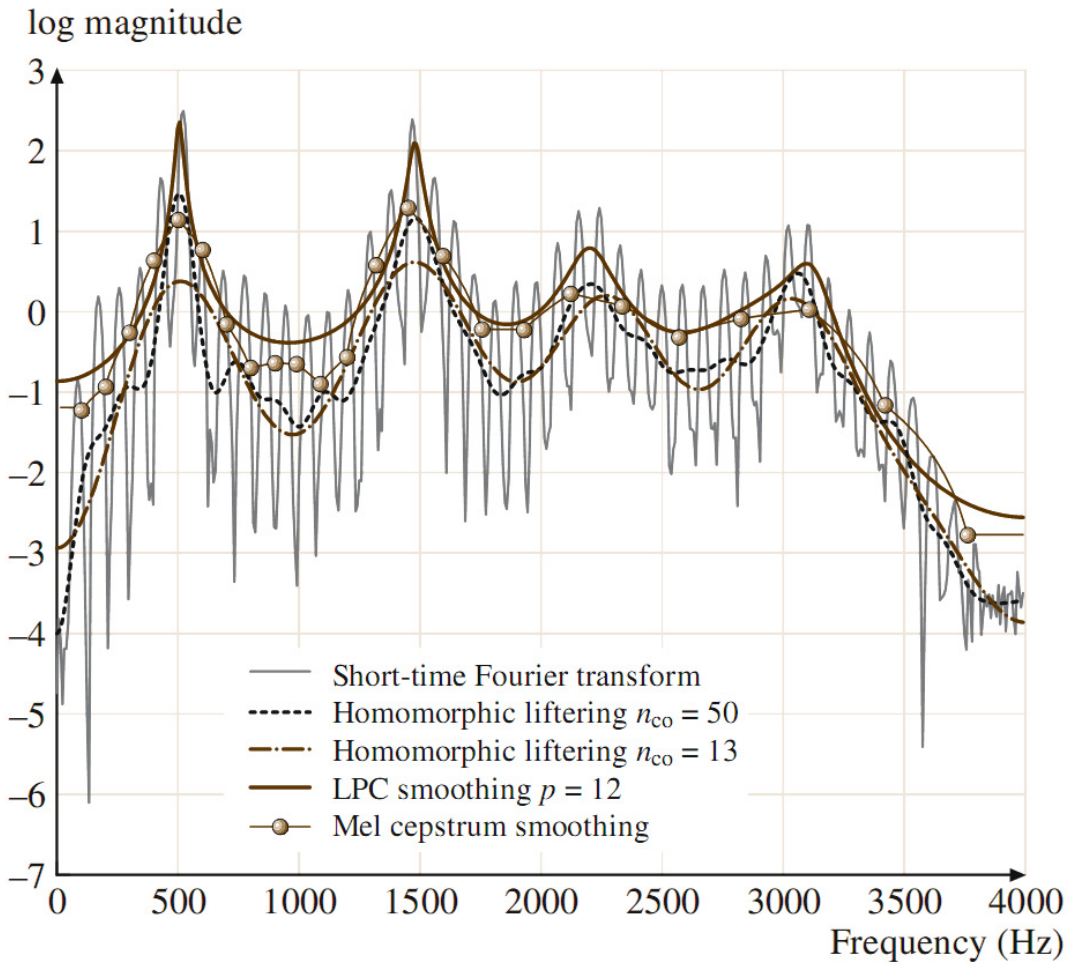
- **Problem:**
  Unknown convolutional effects: microphone transfer function, echoes.
  Features are degraded by noise.

- **Idea:**
  Filter out any cepstral components varying slower or faster than the typical range of change of speech.

- **IIR-bandpass applied in the cepstral domain:**

$$H(z) = 0.1(z^4)\frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}}$$

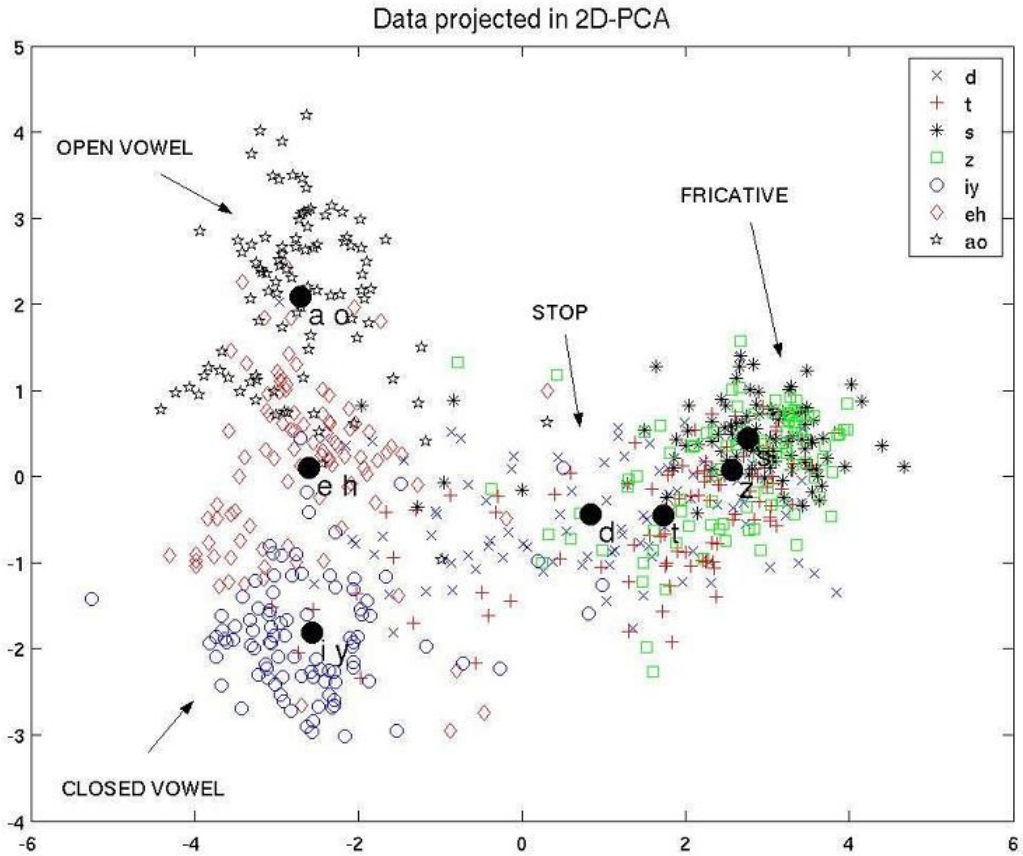
Comparison of spectral smoothing methods



Phone Clustering

- **Problem:**
  Speech, and hence phones are strongly speaker dependent.
  E.g. TIMIT contains 10 sentences spoken by 630 speakers from 8 major dialect regions of the US.
- **Aim:**
  phonetic classification in featurespace
  group phonetic features automatically
- **methods:**
  - PCA
  - LDA
  - K-means
  - EM

Phoneme Clustering



First 2 principal components of 100 feature vectors per phoneme, segmented in the phonetic classes {fd,t,s,z,iy,eh,aog}

$\rightarrow$ Matlab demo

Bibliography

Julien Neel

Cluster analysis methods for speech recognition

Master Thesis in Speech Technology, Department of Speech, Music and Hearing Royal Institute of Technology, S-100 44 Stockholm

Fang Zheng, Zhanjiang Song, Ling Li, Wenjian Yu, Fengzhou Zheng, and Wenhu Wu

THE DISTANCE MEASURE FOR LINE SPECTRUM PAIRS APPLIED TO SPEECH RECOGNITION

Speech Laboratory, Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, P. R. China

Raimund Eigner

Feature Extraction for Speech Recognition

University of Applied Sciences Telematics/Network Engineering, FH Krnten

Brian Mak Etienne Barnard

PHONE CLUSTERING USING THE BHATTACHARYYA DISTANCE

Center for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology 20000 N.W. Walker Road, Portland, OR 97006

Hynek Hermansky

Perceptual linear predictive (PLP) analysis of speech

Speech Technology laboratory, Division of Panasonic Technologies, Inc., 3888 State Street, Santa Barbara, California 93105, 1989

Bibliography

Hynek Hermansky

RASTA processing of speech

IEEE transactions on speech and audio processing, vol.2, No.4, October 1994.

Jacob Benesty, M. Mohan Sondhi, Yiteng Huang

Springer Handbook of Speech Processing, 2008

ISBN: 978-3-540-49125-5

Visala Namburu

Speech Coder using Line Spectral Frequencies of Cascaded Second Order Predictors

Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, November 2001, Blacksburg, VA