Lecture

Speech and Audio Signal Processing

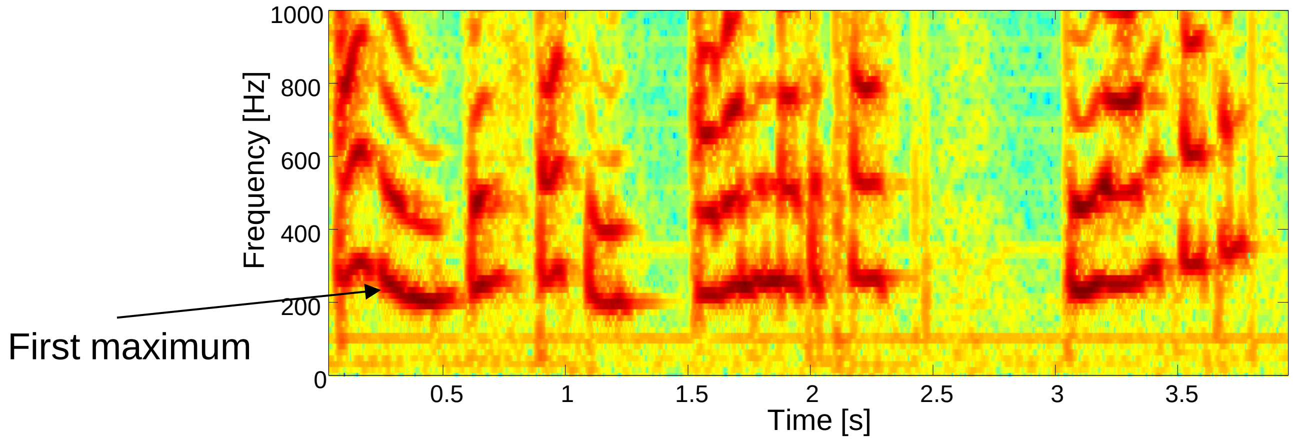
Lecture 8: Fundamental Frequency Estimation and Cepstral Processing, MFCCs

Content

- Methods for the estimation of the fundamental (or pitch) frequency
- Voiced / unvoiced classification
- Post processing for an enhanced fundamental frequency estimation
- Cepstrum calculation and processing
- Applications of
  - Fundamental frequency processing
  - Cepstral processing

Applications based on estimated fundamental frequency

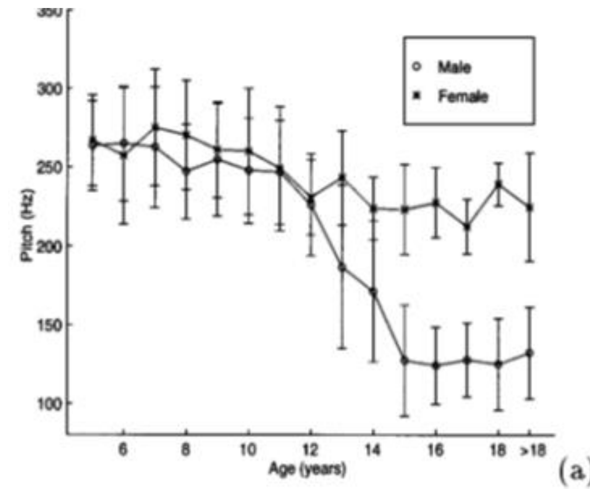
- Fundamental frequency:



- **Applications based on the detected fundamental frequency:**
  - Audio coding
  - Pitch adaptive post filter for noise reduction
  - Noise reduction => additional attenuation between multiples of the fundamental frequency.

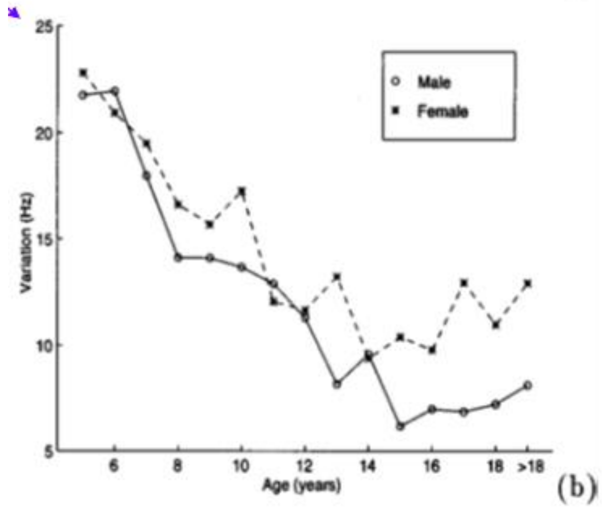
Fundamental frequency values: sex / age

- Absolute value



(a)

- Variation



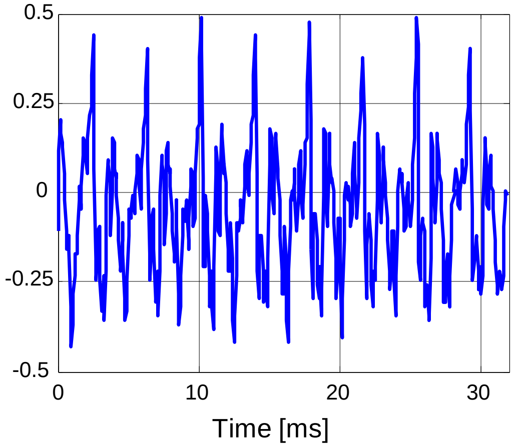
(b)

Fundamental frequency estimation methods

- **Following estimation methods will be analyzed:**
  - Autocorrelation based method
  - YIN procedure

1) Autocorrelation based method

- Periodic voiced input signal frame



- Biased ACF estimate:

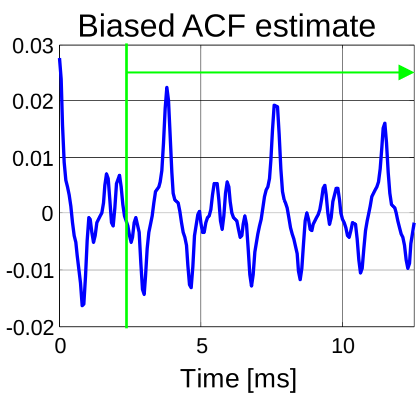
$$\hat{r}_{\mathrm{b},xx}(l,n) \;=\; \frac{1}{L}\sum_{n_0=n}^{n+L-1-l} x(n_0)\,x(n_0+l), \quad \text{for } l \geq 0,$$

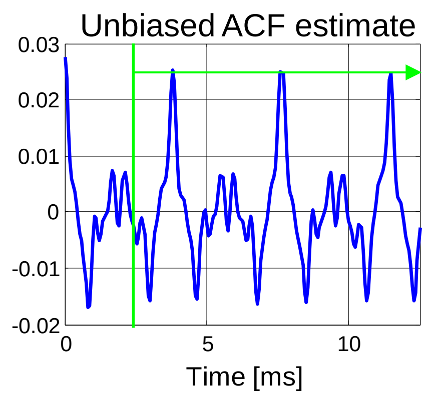
=> decaying estimated over time-lag index *l*

- Unbiased ACF estimate:

$$\hat{r}_{\mathrm{ub},xx}(l,n) \;=\; \frac{1}{L-l}\sum_{n_0=n}^{n+L-1-l} x(n_0)\,x(n_0+l), \quad \text{for } l \geq 0,$$

- Search for maxima of the ACF in a frame of typical fundamental periods:



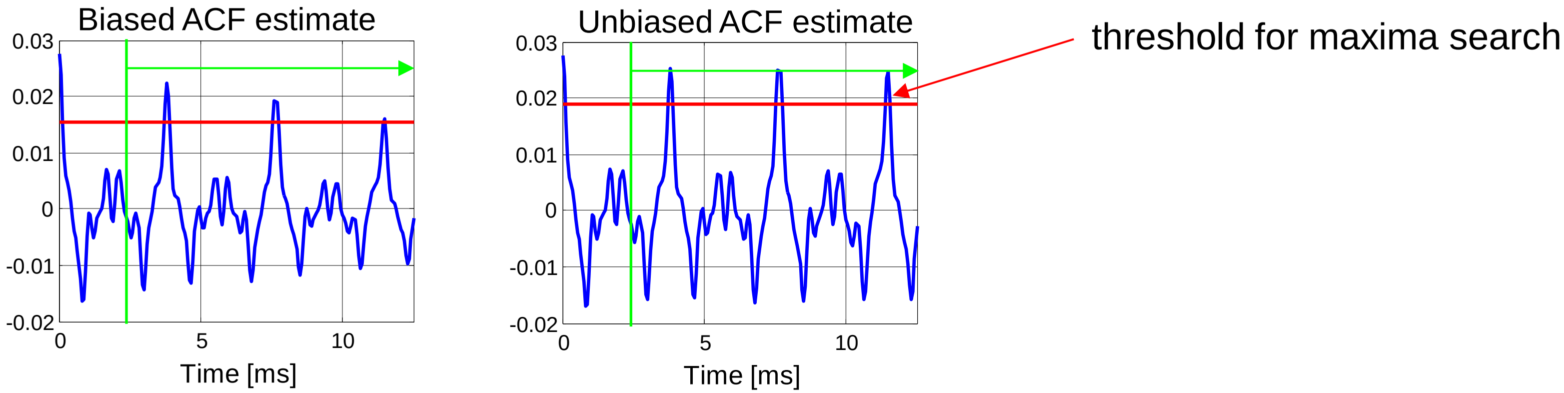


Range of human fundamental frequencies:
=> 50 Hz – 400 Hz

=> fundamental period length:
=> 2.5 – 20 ms

1) Autocorrelation based method

- Search for maxima of the ACF in a frame of typical fundamental periods:



- In order to avoid wrong estimates => search only above a threshold

- Threshold based on the mean of the positive ACF values above the minimum fundamental period:

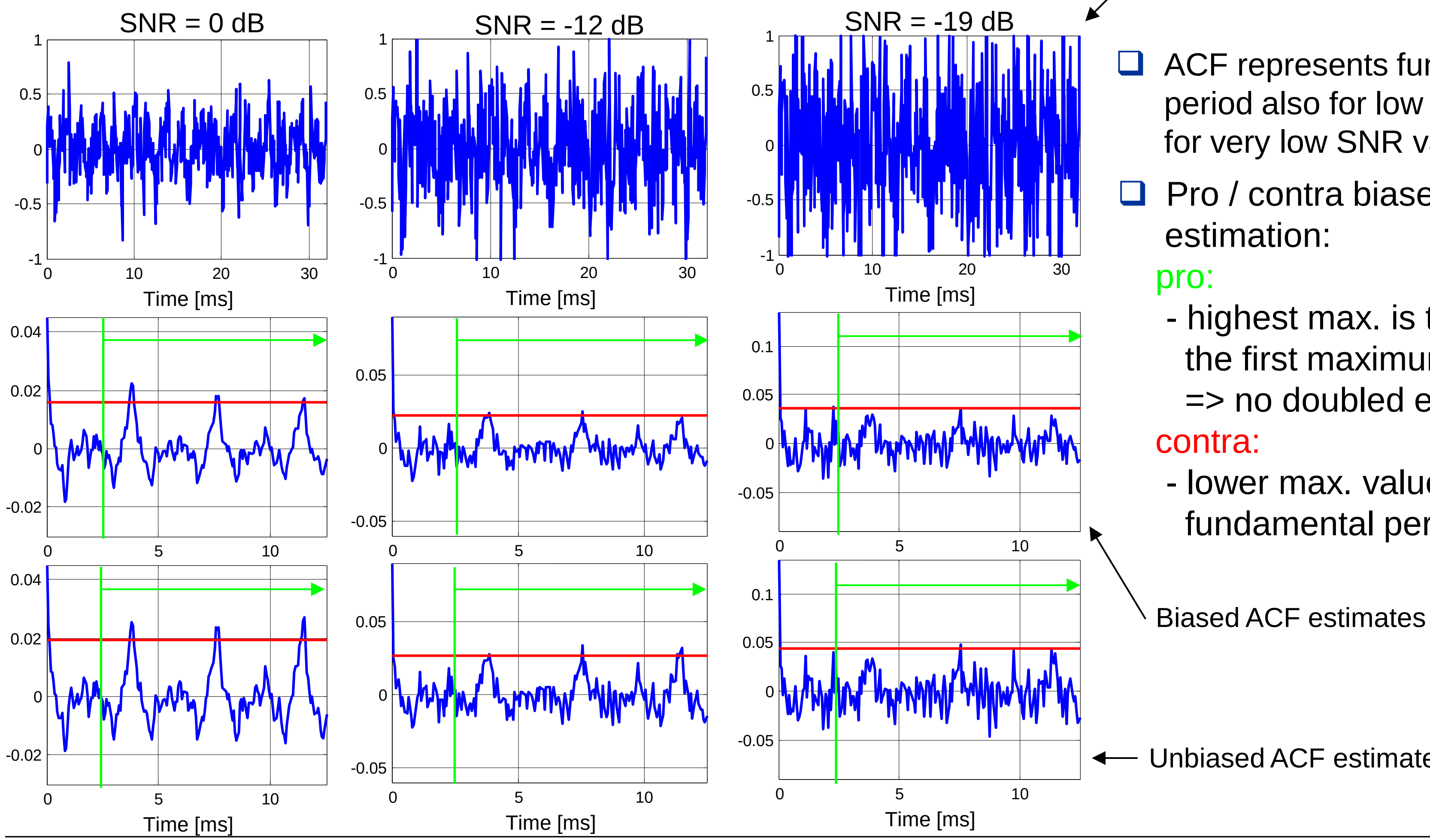
$$\overline{\hat{r}_{xx}}(n) \;=\; \frac{1}{L_{\max} - L_0 + 1} \sum_{l=L_0}^{L_{\max}} \max\left\{\hat{r}_{xx}(l,n), 0\right\}$$

- Threshold: a multiple m of the mean. Value applied here: $m = 6$

$$tr(n) \;=\; m\,\overline{\hat{r}_{xx}}(n)$$

1) Autocorrelation based method

- Noisy input signals (white noise superposition):



- ACF represents fundamental period also for low but not for very low SNR value.
- Pro / contra biased estimation:

  pro:
  - highest max. is typically the first maximum
    => no doubled estimation

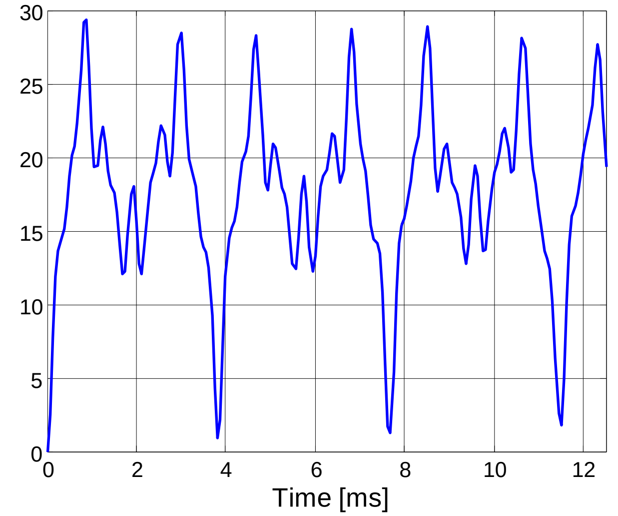
  contra:
  - lower max. value for higher fundamental periods

2) YIN approach [3]

❑ Concept of the approach:

1) Use the difference function and search for minima of this difference function:

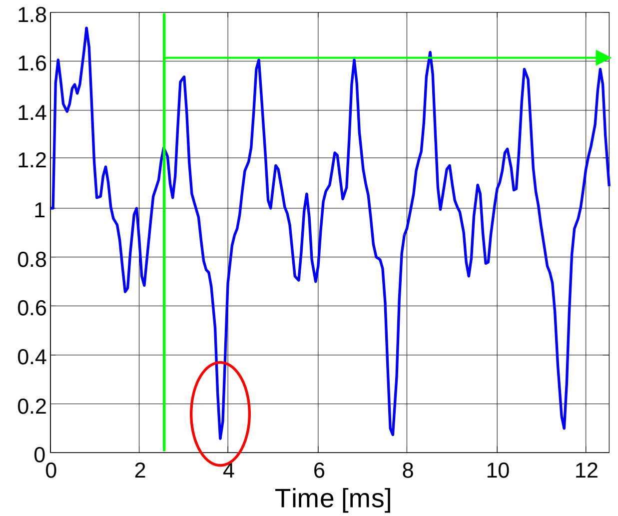
$$d_{\text{ind}}(l,n) \;=\; \sum_{n_0=n}^{n+L-1} \left(x(n_0) - x(n_0+l)\right)^2$$



2) Normalization:

$$d_{\text{ind,norm}}(l,n) = \begin{cases} 1 & \text{for } l = 0, \\ d_{\text{ind}}(l,n) \Big/ \left[\frac{1}{l}\sum_{j=0}^{l} d_{\text{ind}}(j,n)\right] & \text{else.} \end{cases}$$

3) Search for the first minimum of the normalized difference function (in the search window)
=> fundamental period

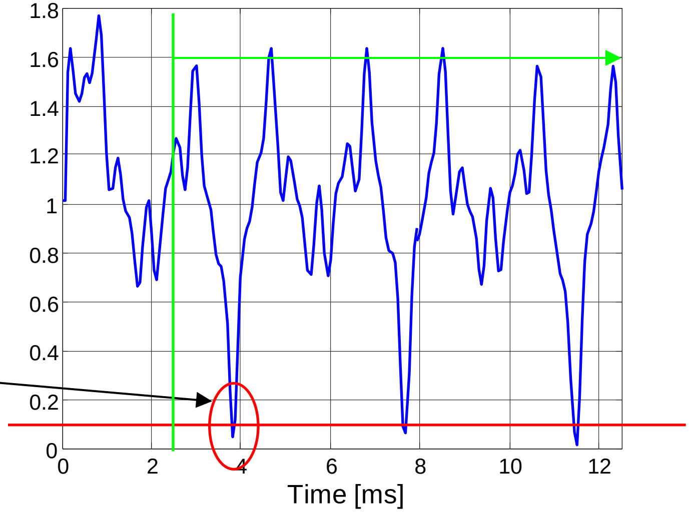


2) YIN approach

4) Method for minimum search:

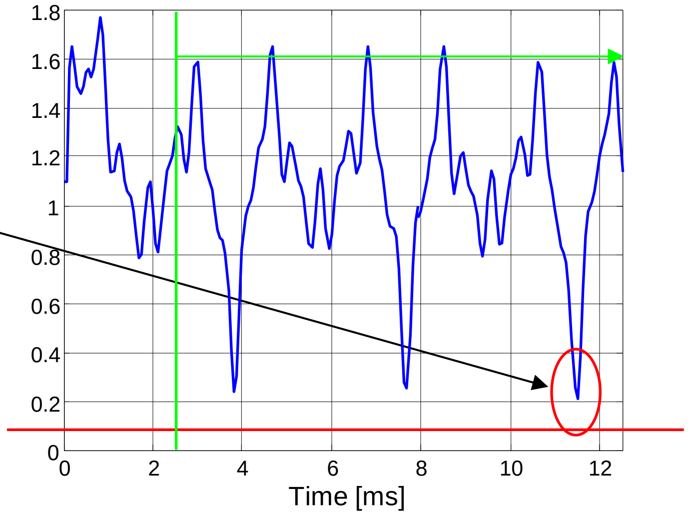
- Compare normalized indicator function with an absolute threshold: 0.1

1) In case several values are below the threshold
=> take the minimum with the lowest fundamental period



2) In case no value is below the threshold
=> take the lowest minimum

=> fundamental period: corresponding time value.

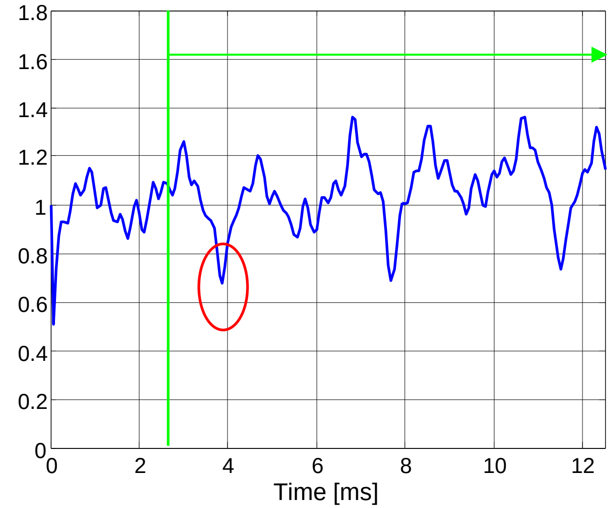


2) YIN approach

TECHNISCHE UNIVERSITÄT DARMSTADT

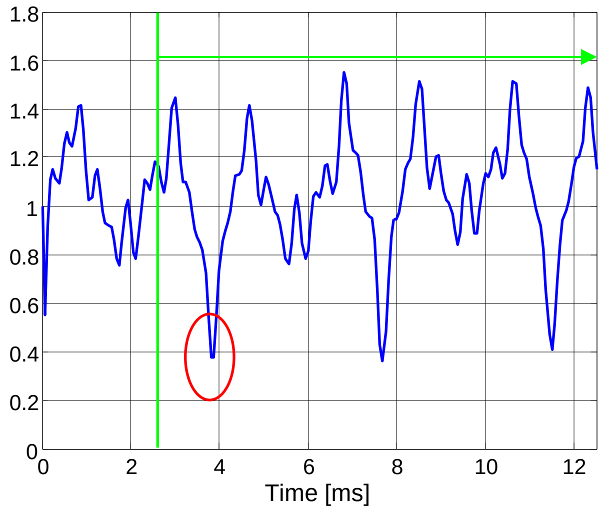
- **Sensitivity to white noise of the indicator function:** $d_{\text{ind,norm}}(l, n)$

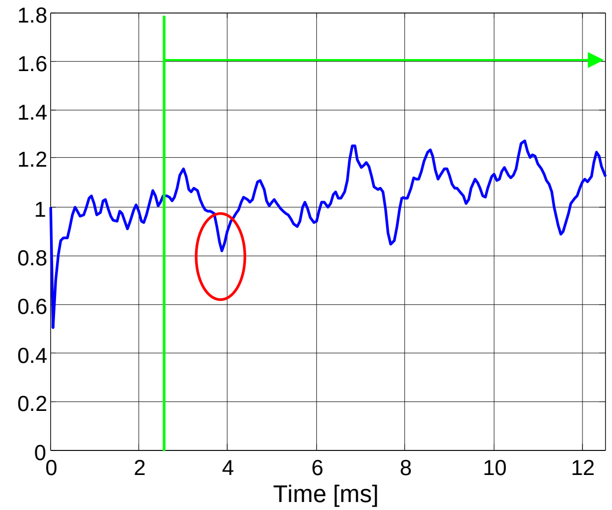
  Correct fundamental frequency is determined for all SNR values of this example.



SNR = -12 dB

SNR = 0 dB





SNR = -19 dB

Error analysis

- **Gross error rate (GER):**

  When is the determined value of the fundamental frequency an error?

  The distance to the true value is determined. In case there is more than x % difference => the estimated value is considered as an error.

  Typically for x one takes a value of 10 %

$$error(n) = \begin{cases} 1 & : \quad \text{if } \frac{|\hat{p}(n) - p(n)|}{p(n)} > 0.1 \\ 0 & : \quad \text{else} \end{cases}$$

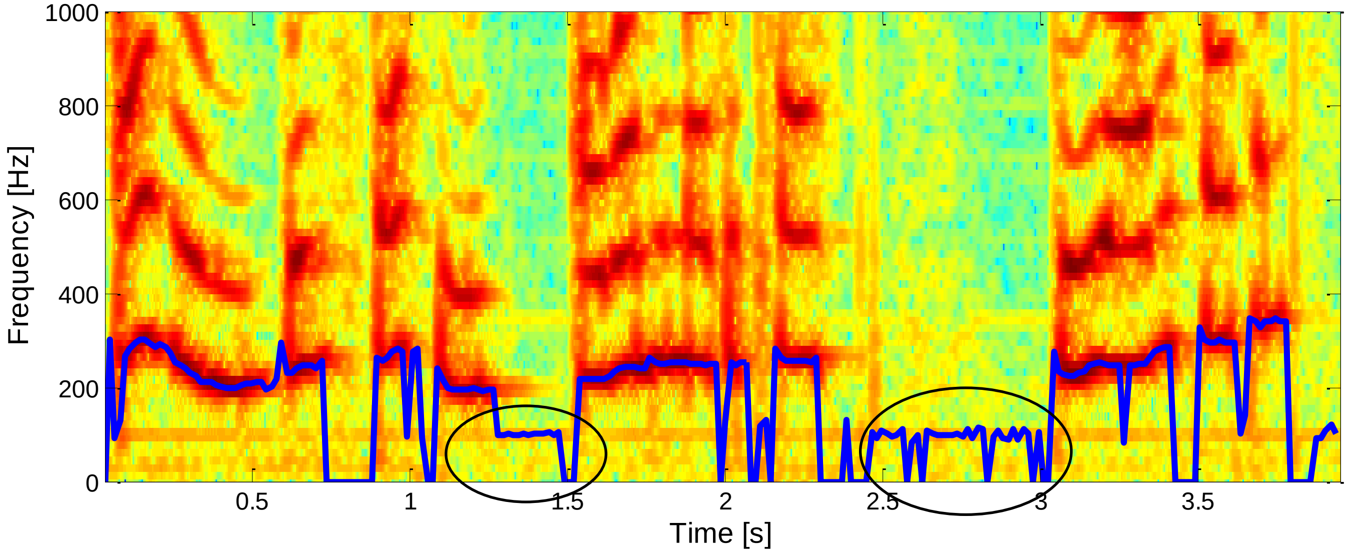
$$\text{GER}[\%] = 100\% \, \frac{1}{N} \sum_{n=0}^{N-1} error(n)$$

- Errors which often occur are double or half fundamental frequency estimates.

Intermediate result

- **Gross error rate:**

  Also for non-voiced time frames, often an estimate is determined.
  Here, based on some soft tonal noise



=> Needs to be avoided by determining voiced speech frames.

Detection of voiced speech frames

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Different possibilities:**

1) **Total frame energy**:

$$pow(n) = \sum_{n_0=n}^{n+L-1} x^2(n_0)$$

Voiced speech frames are louder than unvoiced speech frames.

2) **Maximum of the normalized ACF** value in the search window of possible fundamental periods.

$$r_{\max}(n) = \max\left\{\frac{\hat{r}_{xx}(l,n)}{\hat{r}_{xx}(0,n)},\ \ l \in [l_{\min}, l_{\max}]\right\}$$

Only high correlation values in the search frame for periodic signals, i.e., voiced speech frames.

Detection of voiced speech frames

- **Different possibilities:**

3) **Normalized high-pass energy**:

$$p_{\mathrm{HP}}(n) \;=\; \frac{\sum\limits_{n_0=n}^{n+L-1} \left(x(n_0) - x(n_0-1)\right)^2}{\sum\limits_{n_0=n}^{n+L-1} x^2(n_0)}$$

Strongly more energy for low frequency components
than high frequency components for voiced speech frames.

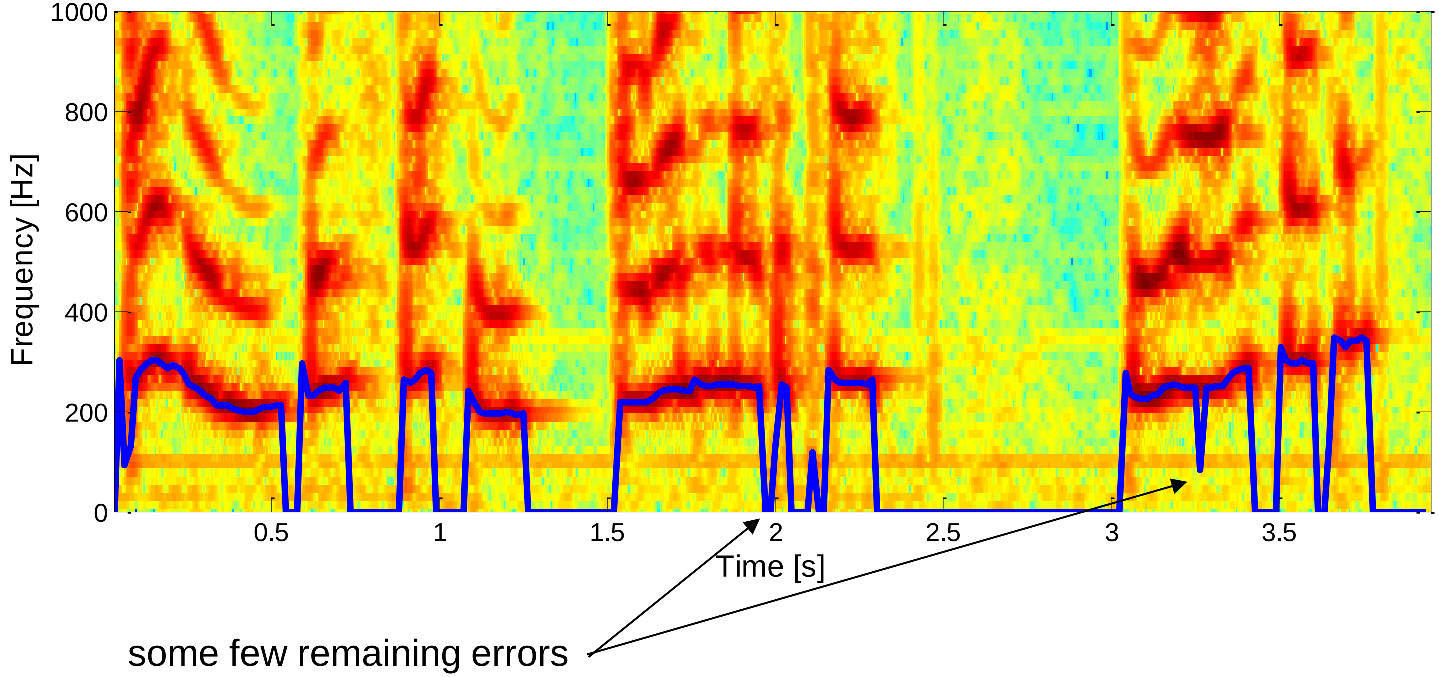
4) **First prediction filter value:**

$$a_1(n) = \frac{\hat{r}_{xx}(1,n)}{\hat{r}_{xx}(0,n)}$$

High correlation for voiced speech frames.

Results after detection voiced periods

- **Fundamental frequency calculation only for voiced frames:**

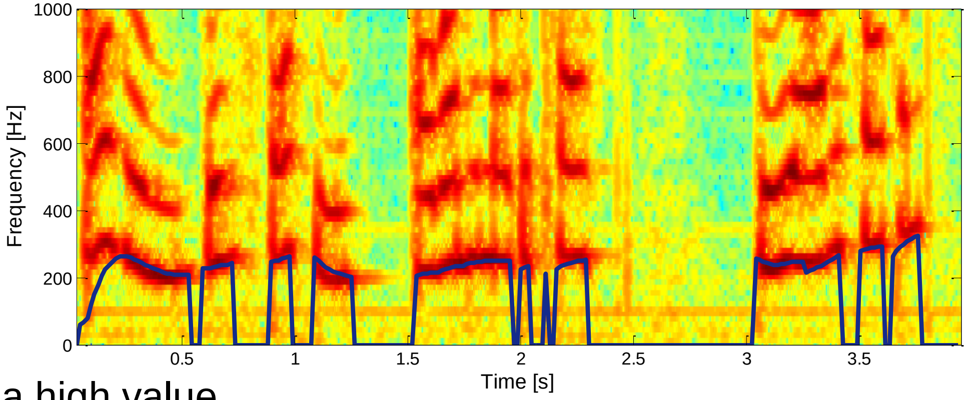


Post processing

- **Smoothing:** => large estimation errors lead to a wrong estimation

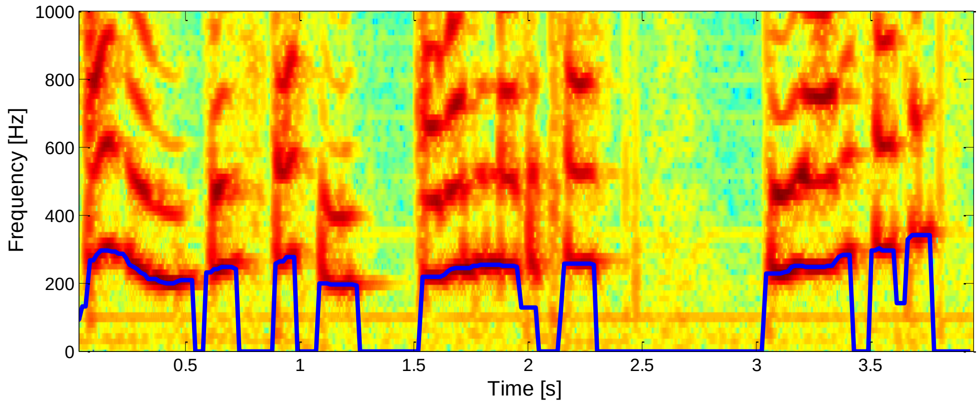
$$\overline{\hat{f}_0(n)} = \alpha\, \overline{\hat{f}_0(n-1)} + (1-\alpha)\, \hat{f}_0(n)$$

Caution: Smoothing only for voiced speech frames!
Skip non-voiced frames.



- **Median filtering:** => fill a dip or remove a high value

$$\hat{f}_{0,\mathrm{median}}(n) = \mathrm{median}(\hat{f}_0(n-2), \ldots, \hat{f}_0(n+2))$$



Cepstrum

- **Calculation of the Cepstrum of the signal periodogram:**

Continuous frequency:

$S(e^{j\Omega}, n)$:short-term spectrum

Cepstral coefficients:

$$c_i(n) = \frac{1}{2\pi} \int_{\Omega=-\pi}^{\pi} \ln\left\{|S(e^{j\Omega}, n)|^2\right\} e^{j\Omega\, i}\, d\Omega$$

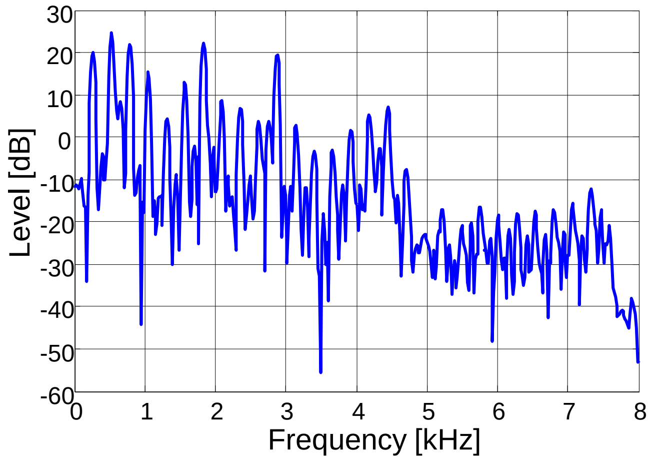
Discrete frequency:

$S(e^{j\Omega_\mu}, n)$:short-term spectrum

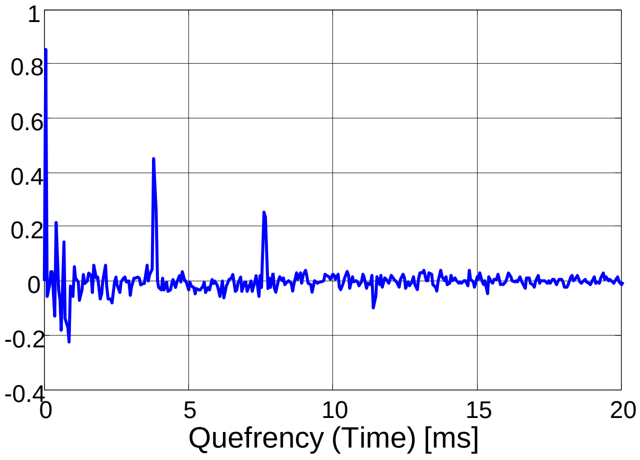
Cepstral coefficients:

$$c_i(n) = \frac{1}{M} \sum_{\mu=0}^{M-1} \ln\left\{|S(e^{j\Omega_\mu}, n)|^2\right\} e^{j\frac{2\pi}{M}\mu i}$$

- Spectrum of voiced speech :



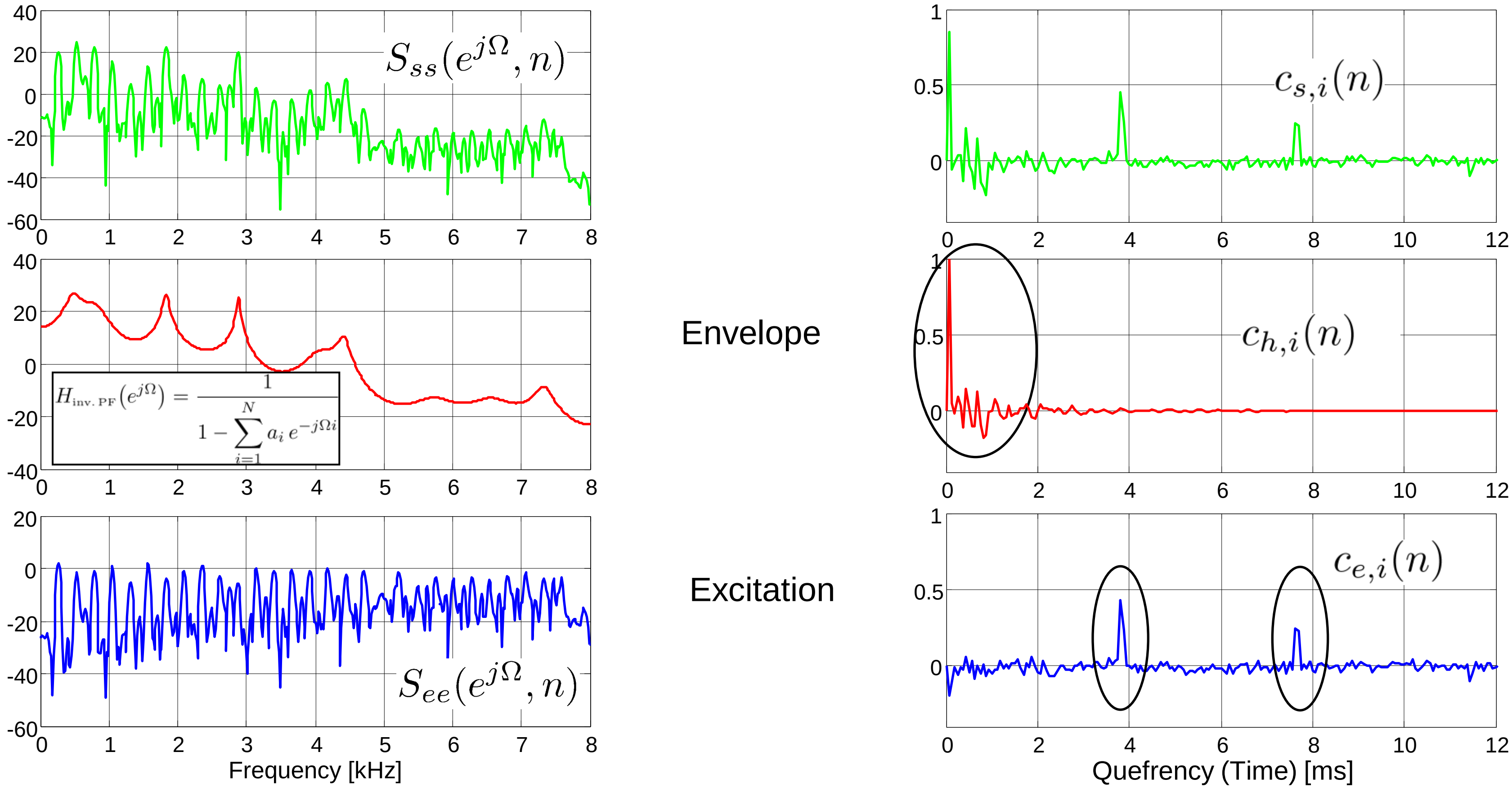
- Cepstrum of voiced speech:



Cepstrum

- **Additive** relation of the cepstra of the excitation signal and the envelope:

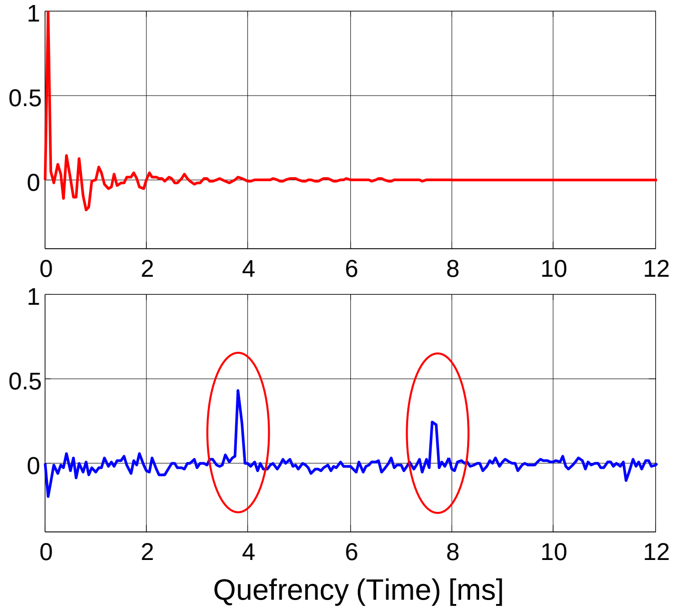
$$S_{ss}(e^{j\Omega},n) = S_{ee}(e^{j\Omega},n)\,|H_{\text{inv. PF}}(e^{j\Omega},n)|^2 \quad \Rightarrow \quad c_{s,i}(n) \; = \; c_{e,i}(n) + c_{h,i}(n)$$



Cepstrum

- Additive relation of the cepstra of the excitation signal and the envelope:

$$c_{s,i}(n) \ = \ c_{e,i}(n) + c_{h,i}(n)$$

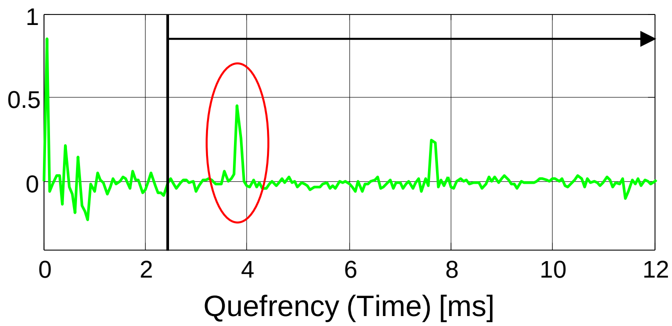


Envelope: Components only at low „quefrencies“

Excitation: Peaks at multiples of the fundamental period

Fundamental frequency estimation based on cepstrum

- First strong peak of the cepstrum in the typical range for speech fundamental frequencies.



- Problem: In noisy speech frames, the spectral peak is strongly less prominent.
  => Fundamental frequency estimation by using the cepstrum is typically not applied.

Applications based on estimated fundamental frequency

TECHNISCHE UNIVERSITÄT DARMSTADT

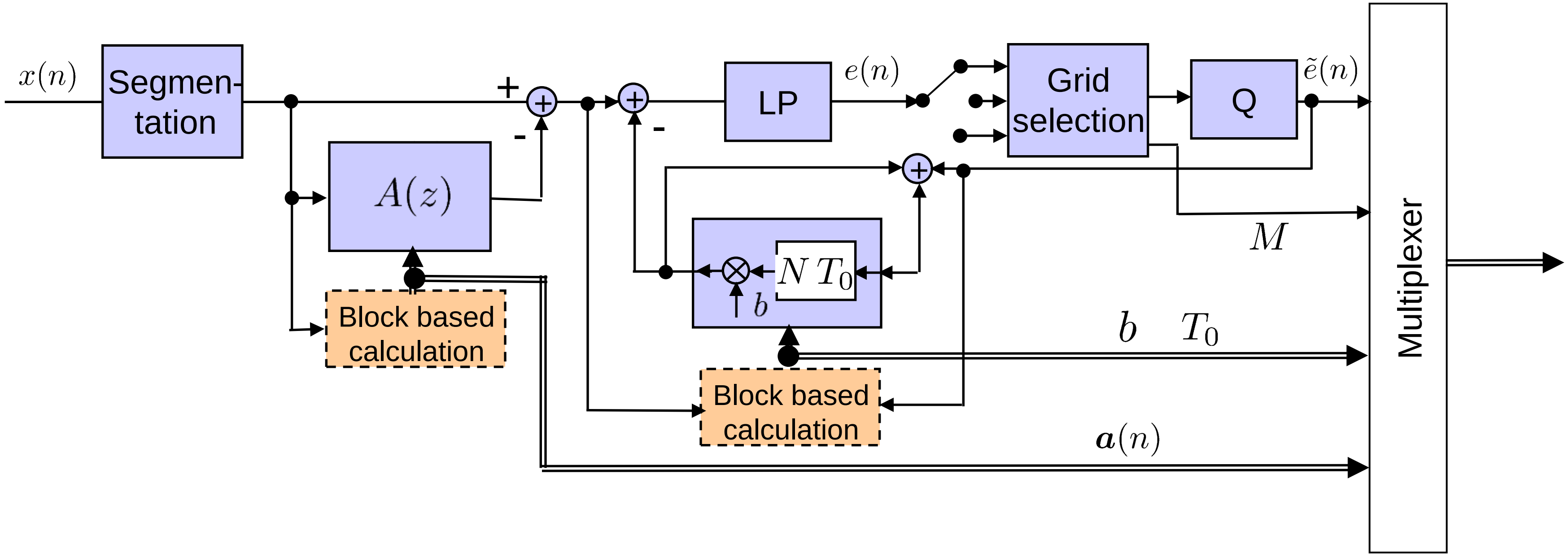
- **Applications based on the detected fundamental frequency:**
  - Audio coding
  - Pitch adaptive post filter for noise reduction

Applications (I): Audio coding

- **Fundamental period used in long-term prediction units:**

  T_0 : fundamental period

Reminder: Residual Excited Linear Prediction:



Applications (II): Pitch adaptive post filter [4]

- **Concept:**
  - Apply additional attenuation (after typical Wiener filter based noise reduction) between the components of the fundamental frequency after the calculation of the noise reduction

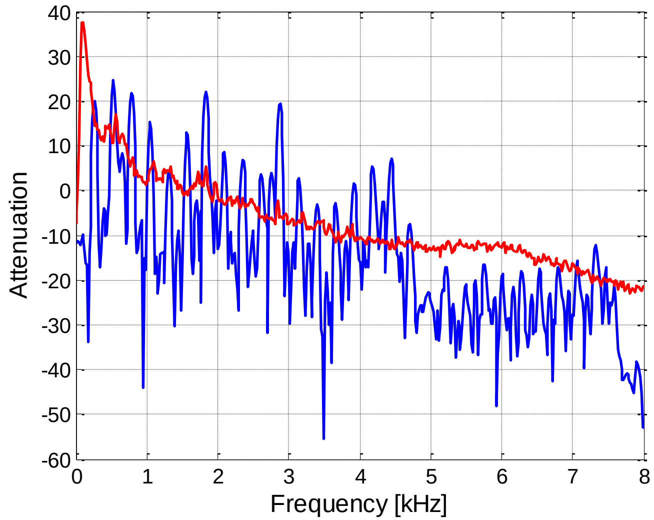
O calculo da freq fundamental tem uma certa tolerancia, entao o calculo das harmonicas múltiplas pode ser um pouco ruim

Section of a voiced speech frame:

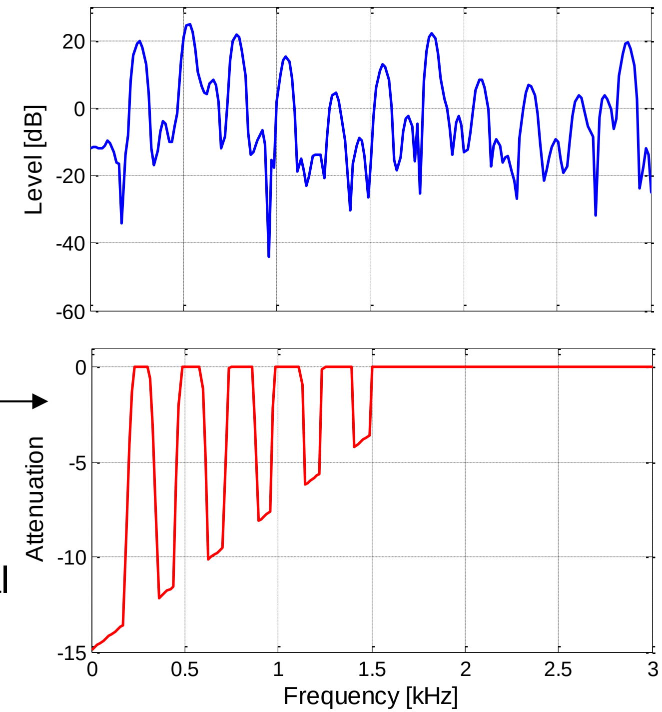
Example for one time frame:

Speech spectrum (blue)

Noise spectrum (red)



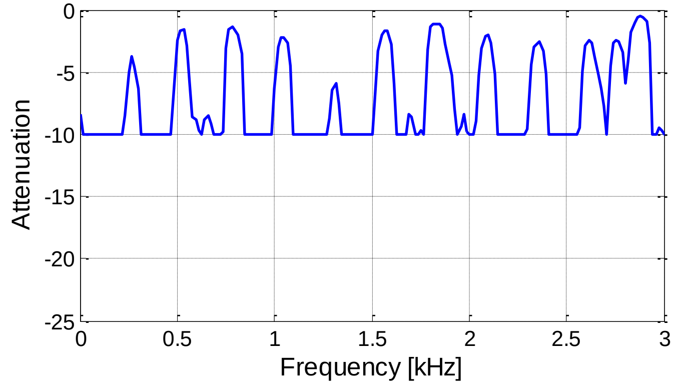
Attenuation between periodic components for voiced frames designed based on the known fundamental frequency



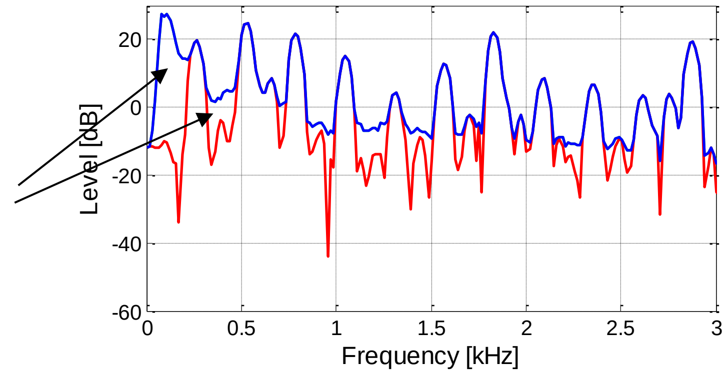
Applications (II): Pitch adaptive post filter

Quanto mais ruídoso, menor a minha capacidade de distinguir a freq fundamental

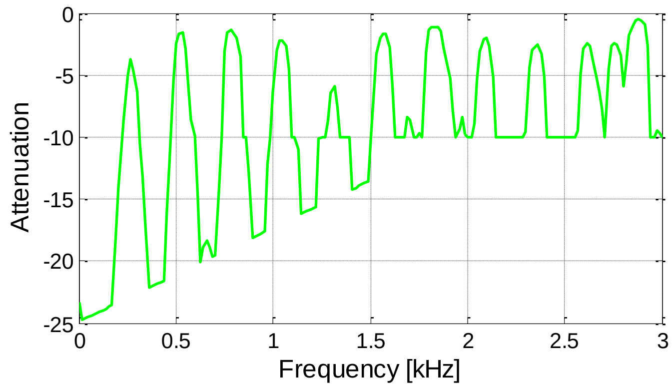
- Wiener filter attenuation limited by the spectral floor (10 dB):



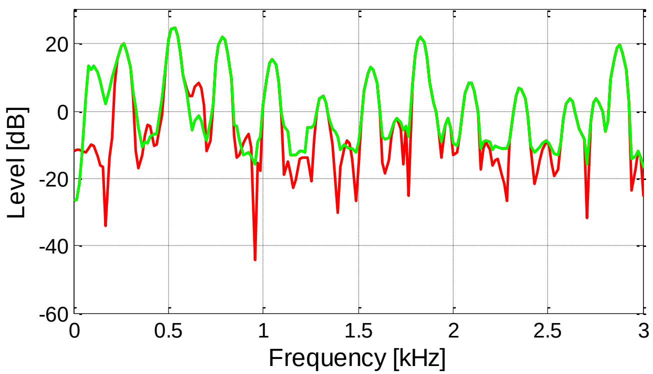
- Clean (red) and estimated (blue) speech spectrum => not enough attenuation



- Wiener filter attenuation combined with the post filter:



- Clean (red) and estimated (green) speech spectrum => strong attenuation between harmonic components



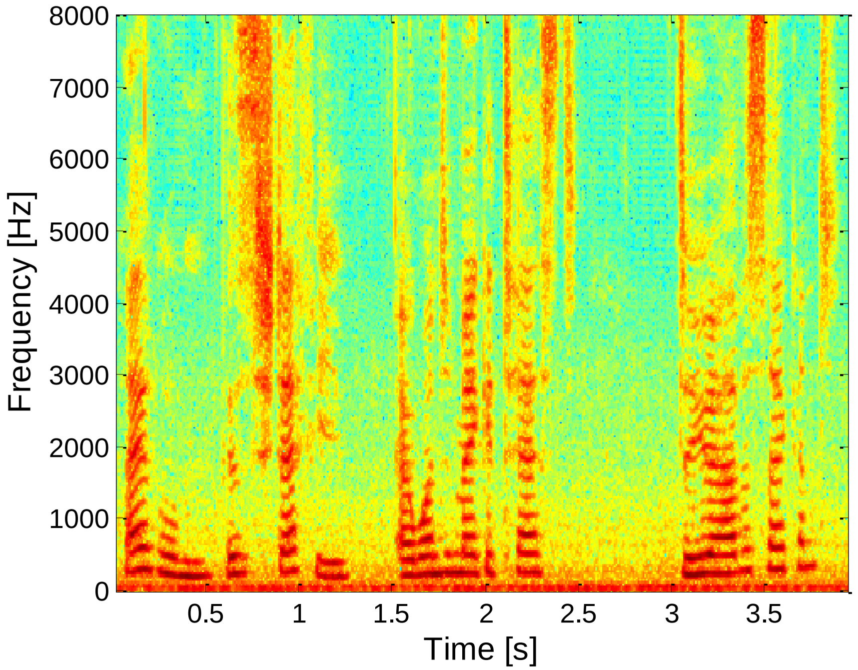
Applications (II): Pitch adaptive post filter

é mais fácil aplicar esse método para mulheres ja que o periodo fundamental é maior entao o ouvido é capaz de distinguir melhor

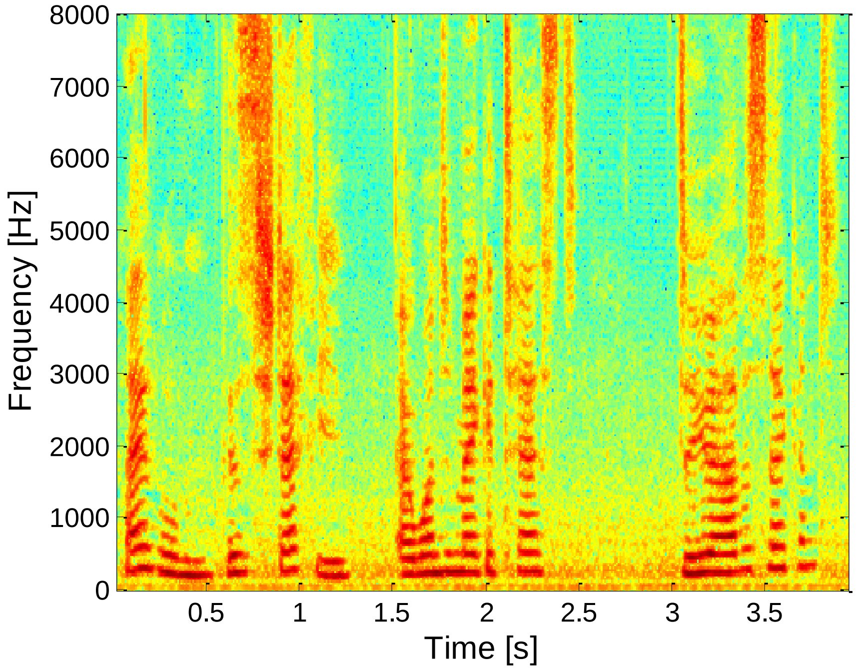
- Additional attenuation by the post filter is limited to frequency regions up to 1-1.5 kHz and reduced for increasing frequencies:
  Reason: No strong periodic harmonics => risk of target signal attenuations

- Listening example: Input signal:

Without post filter:



With post filter:



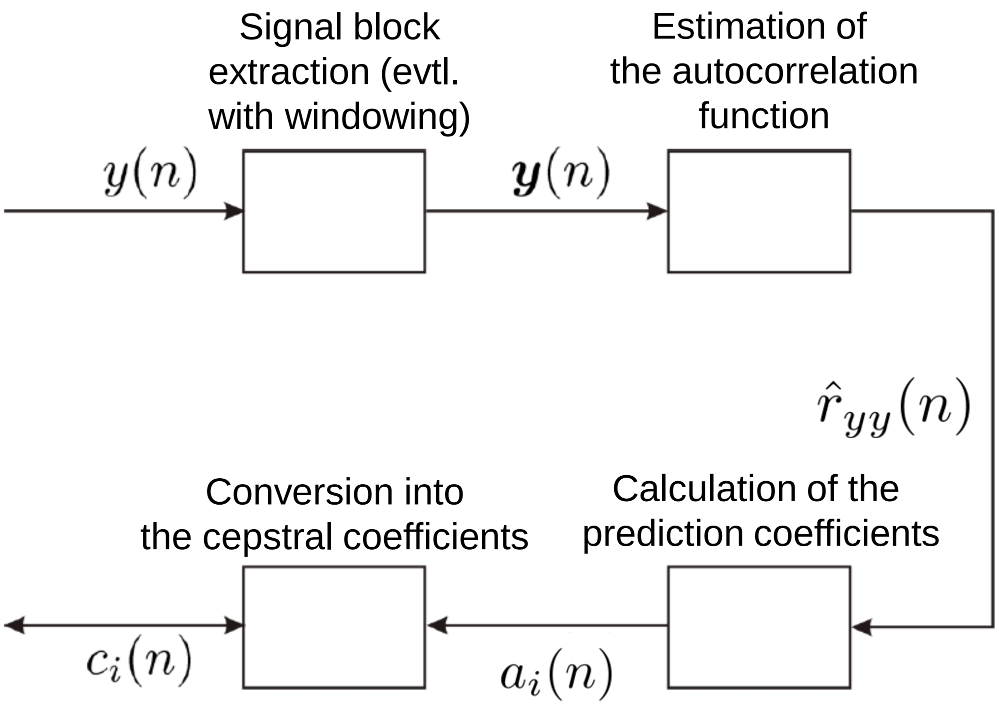
Simple blocks for speech based detection methods

- Speech based detection methods are typically designed based on two main blocks:
  - 1) Feature extraction, e.g., MFCCs
  - 2) Modelling by probability models
    - without memory, e.g., Bayes, Gaussian mixture models
    - with memory, e.g., Hidden-Markov models

NoiseReduction e Beamforming

Se usar MFCC, talvez de pra diminuir o tamanho da DNN

Representation of the spectral envelope by cepstral coefficients

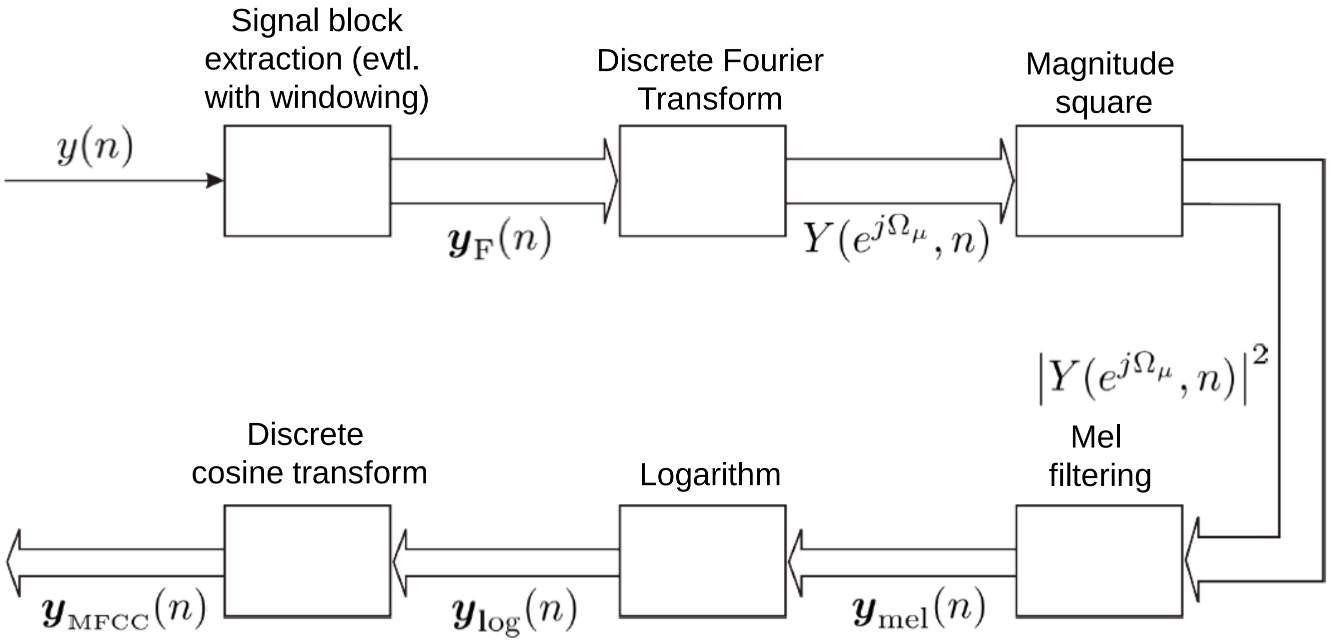


- Typically, every 5 to 20 ms 15 to 30 cepstral coefficients are calculated.
- First the coefficients of a prediction error filter are determined.
- The necessary autocorrelation values are estimated based on input signal blocks of a length of 20 to 50 ms.
- Finally, the cepstral coefficients are determined which are typically used in all applications based on spectral envelopes, e.g. coding, bandwidth extension, etc.

Mel-filtered cepstral coefficients (MFCCs)

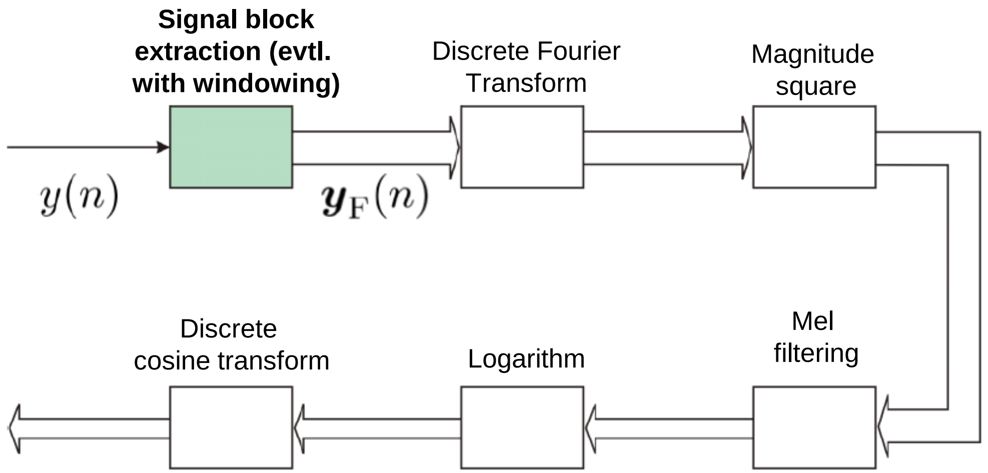
- MFCCs: Speech characteristics determined based on short-term signal periodograms with a frequency resolution adapted to the human ear.

- Overview:



Mel-filtered cepstral coefficients (MFCCs)

- Signal block extraction, subsampling, windowing:



- Block extraction:

$$\tilde{\boldsymbol{y}}(n) = \big[y(n),\, y(n-1),\, \ldots\, y(n-N+1)\big]^{\mathrm{T}}$$

- Block subsampling, e.g. half block overlap $r = N/2$.

$$\boldsymbol{y}(n) = \tilde{\boldsymbol{y}}(nr)$$

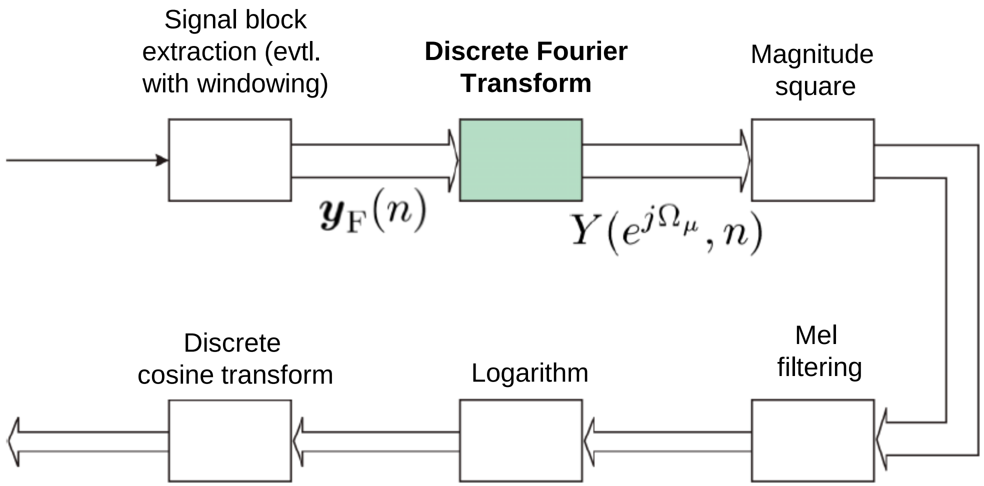
- Windowing:

$$\boldsymbol{y}_{\mathrm{F}}(n) = \boldsymbol{H}\,\boldsymbol{y}(n)$$

$$\boldsymbol{H} = \begin{bmatrix} h_0 & 0 & \ldots & 0 \\ 0 & h_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & h_{N-1} \end{bmatrix}$$

Mel-filtered cepstral coefficients (MFCCs)

- Discrete Fourier Transform:



- Discrete Fourier Transform:

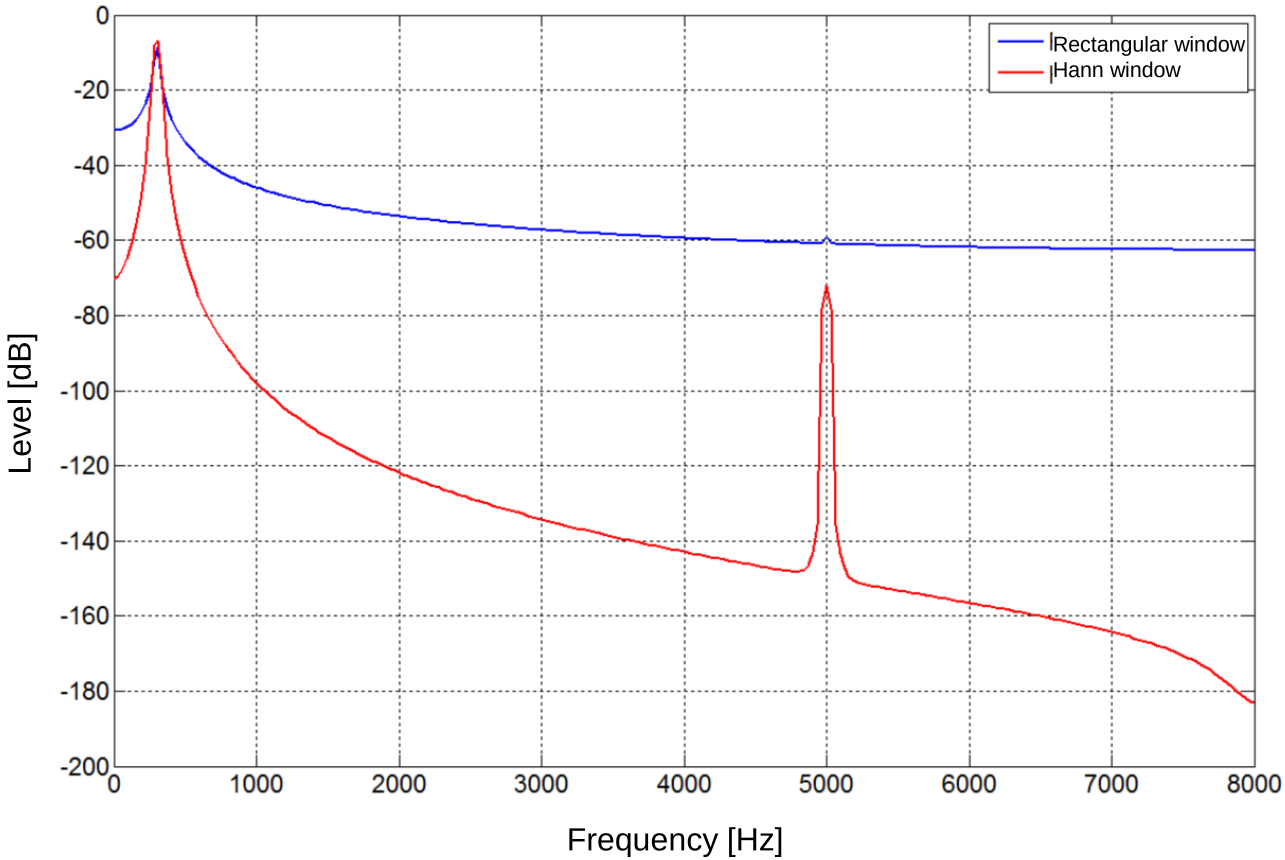
$$Y(e^{j\Omega_\mu}, n) = \sum_{k=0}^{N-1} y(nr-k)\, h_k\, e^{-j\frac{2\pi}{N}k\mu}$$

- Matrix vector notation:

$$\begin{aligned}\boldsymbol{y}(e^{j\Omega}, n) &= \left[Y(e^{j\Omega_0}, n),\, ...,\, Y(e^{j\Omega_{N-1}}, n)\right]^{\mathrm{T}} \\ &= \boldsymbol{T}_N\, \boldsymbol{H}\, \boldsymbol{y}(n)\end{aligned}$$

Mel-filtered cepstral coefficients (MFCCs)

- Influence of the window function:



- Two sinusoidal input signal components at 300 Hz and 5000 Hz

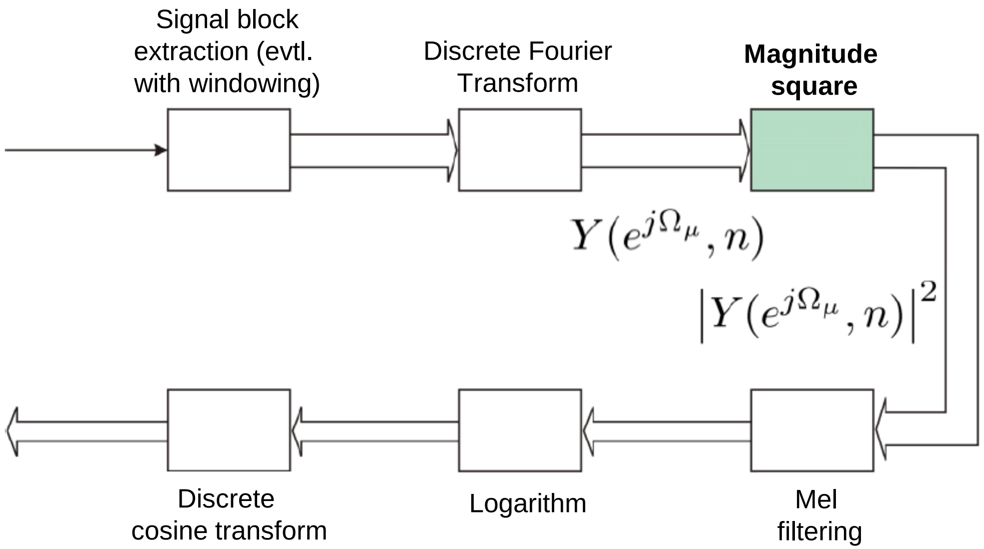
  Level ratio of 66 dB

- FFT size: 512

Mel-filtered cepstral coefficients (MFCCs)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Calculation of the magnitude (square):



- Calculation of the magnitude square:

$$\begin{aligned}\left|Y(e^{j\Omega_\mu}, n)\right|^2 &= \mathrm{Re}^2\{Y(e^{j\Omega_\mu}, n)\} \\ &\quad + \mathrm{Im}^2\{Y(e^{j\Omega_\mu}, n)\}\end{aligned}$$

- Approximation (reduced computational effort and reduced dynamic):

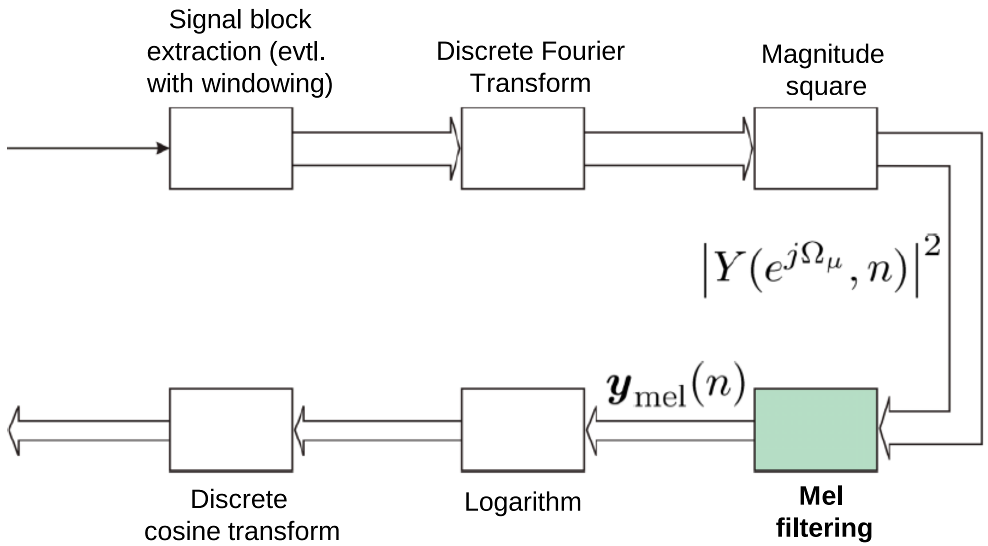
$$\begin{aligned}\left|Y(e^{j\Omega_\mu}, n)\right| &\approx K\left|\mathrm{Re}\{Y(e^{j\Omega_\mu}, n)\}\right| \\ &\quad + K\left|\mathrm{Im}\{Y(e^{j\Omega_\mu}, n)\}\right|\end{aligned}$$

- Matrix vector notation:

$$\boldsymbol{y}_{\mathrm{abs}}(n) = \left[\left|Y(e^{j\Omega_0}, n)\right|, \ldots, \left|Y(e^{j\Omega_{N-1}}, n)\right|\right]^{\mathrm{T}}$$

Mel-filtered cepstral coefficients (MFCCs)

- Mel filtering:



- Mel frequency relation according to the definition of Stanley Smith Stevens:

$$m = 2595\,\mathrm{Mel}\,\log_{10}\left\{\frac{f}{700Hz} + 1\right\}$$

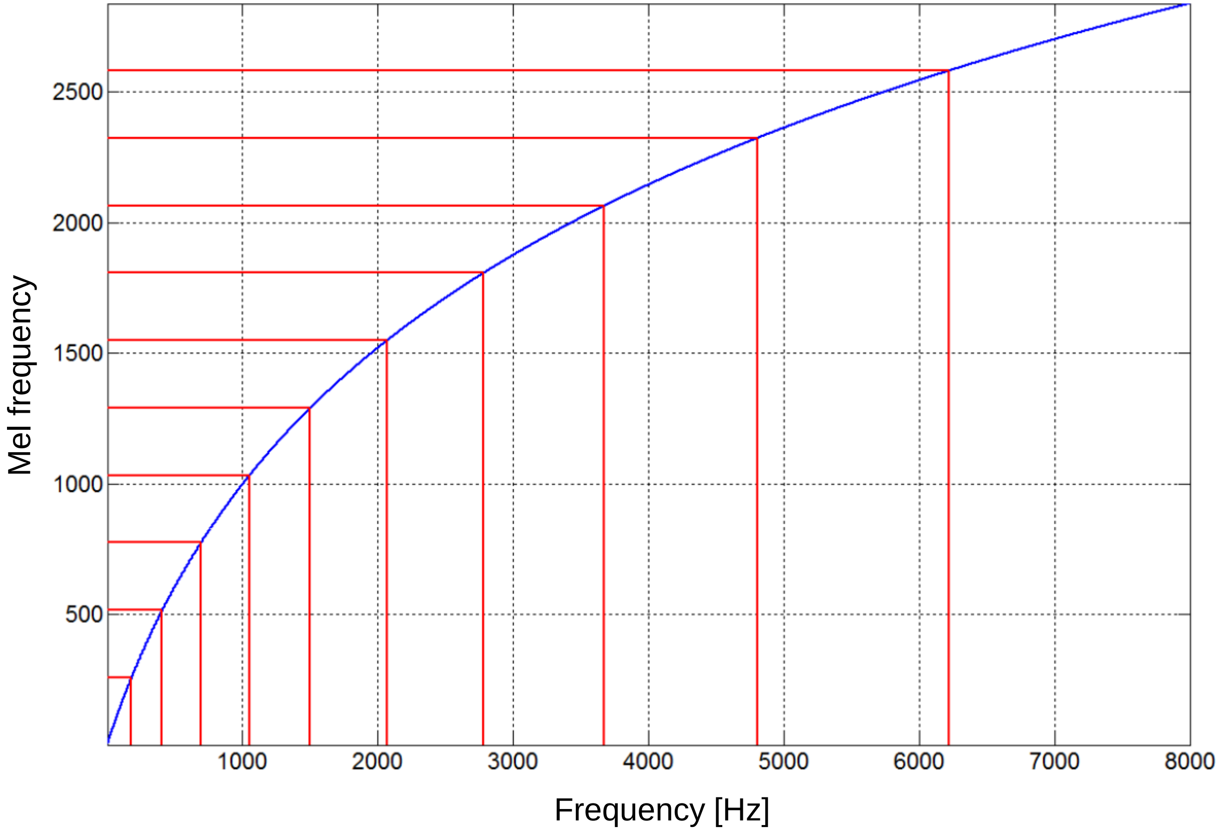
- Decomposition into equal intervals in the Mel frequency domain.
- Interval overlap of 50% of adjacent frequency bands.
- Typically, triangular filters are used.
- Filter normalization such that one obtains the same output power for each filter for white input signals.

Mel-filtered cepstral coefficients (MFCCs)

TECHNISCHE UNIVERSITÄT DARMSTADT

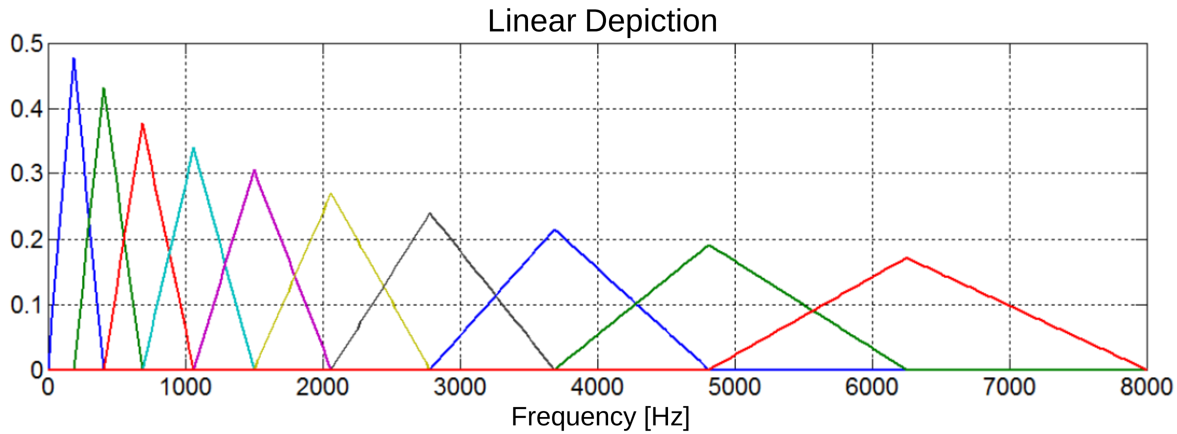
- Mel filtering:

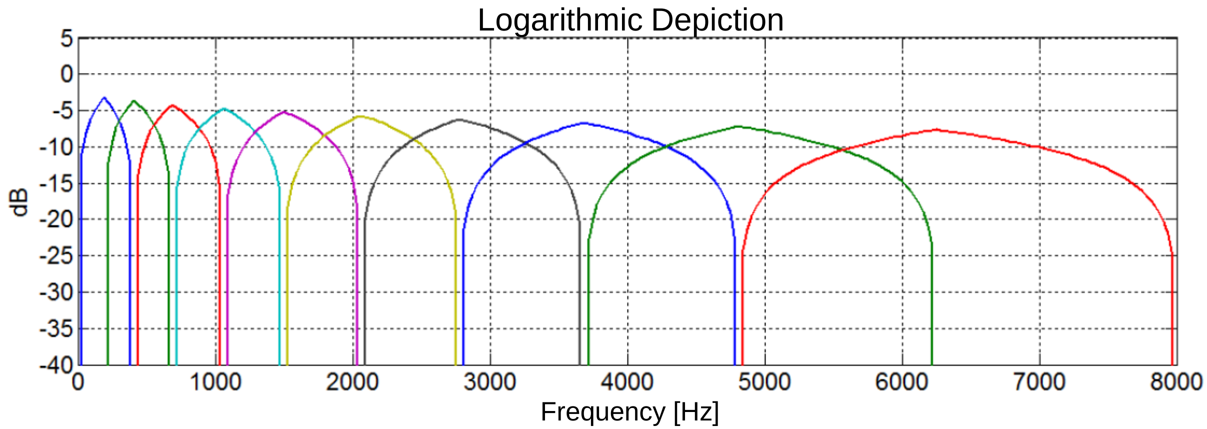
Decomposition into 10-20 equally spaced intervals in the Mel domain



Mel-filtered cepstral coefficients (MFCCs)

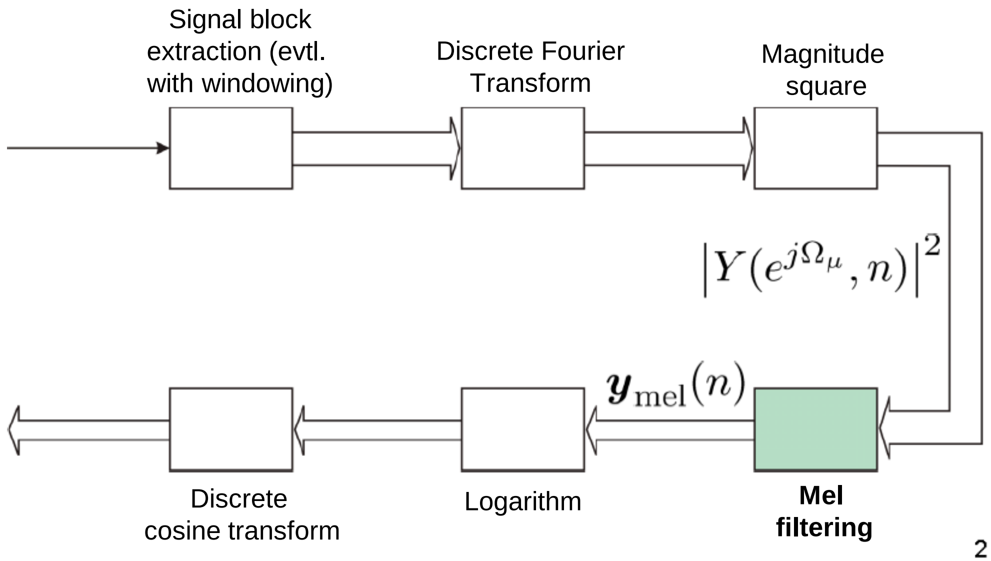
- Mel filtering (for 10 Mel frequency bands):





Mel-filtered cepstral coefficients (MFCCs)

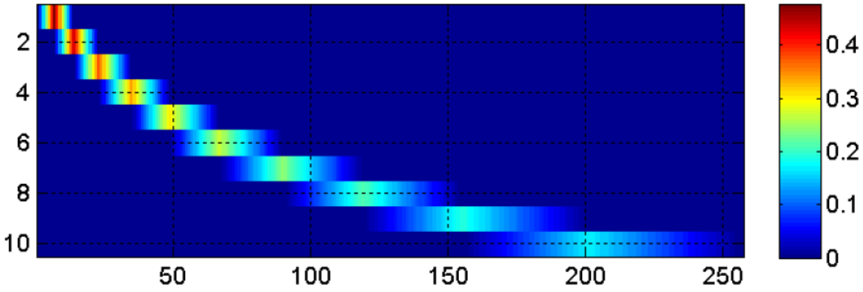
- Mel filtering:



- Typically 15 to 30 Mel filters are used for sample rates of 8 to 16 kHz. (Order 10 in the examples here).
- Matrix vector notation of the Mel filtering i.e. convolution with the Mel filters:

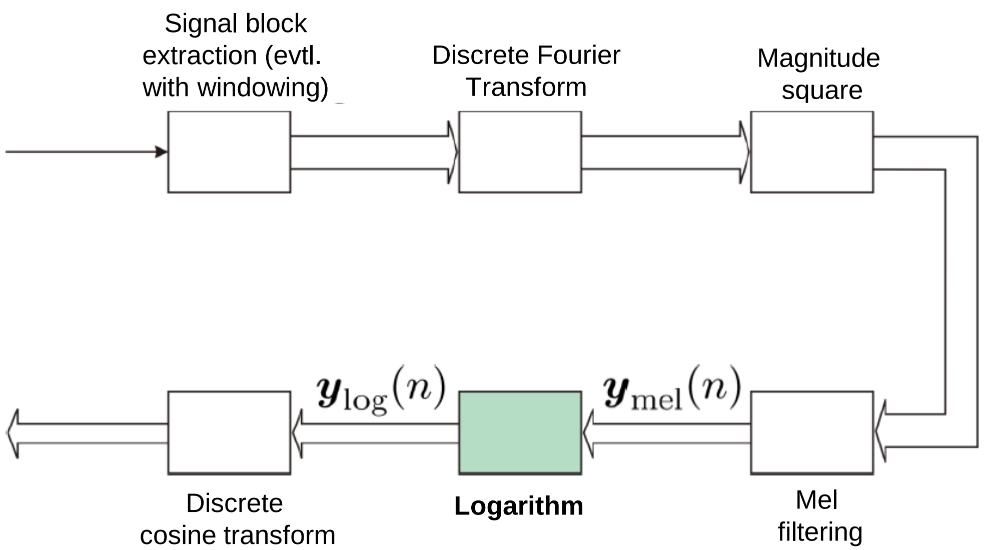
$$\boldsymbol{y}_{\mathrm{mel}}(n) \quad = \quad \boldsymbol{M}\, \boldsymbol{y}_{\mathrm{abs}}(n)$$

- Depiction of the matrix *M*:



Mel-filtered cepstral coefficients (MFCCs)

- Logarithm:



- Calculation of the (natural) logarithm:

$$\boldsymbol{y}_{\log}(n) = \log_e\{\boldsymbol{y}_{\mathrm{mel}}(n)\}$$

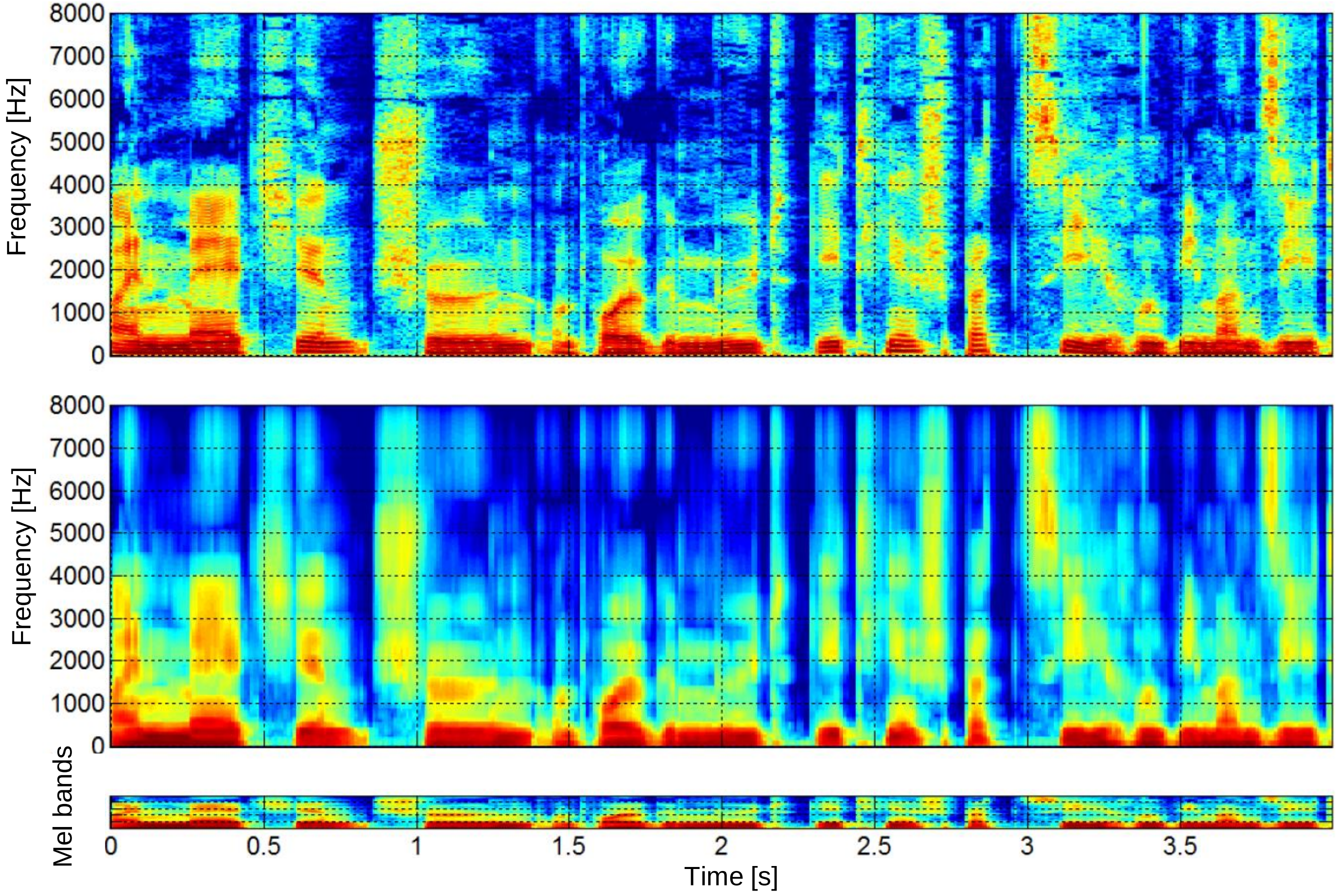
- As an alternative also logarithms at other algorithms can be used.

- Comparable to the Mel frequency resolution the logarithm is based on the human perception (human loudness perception).

Mel-filtered cepstral coefficients (MFCCs)

- Logarithm:

The size of the pictures is proportional to the data rate of the signals:



- Logarithmic spectrogram:

$$\log_e\{\boldsymbol{y}_{\text{abs}}(n)\}$$

- Logarithmic mel-filtered spectrogram at the original frequency resolution:

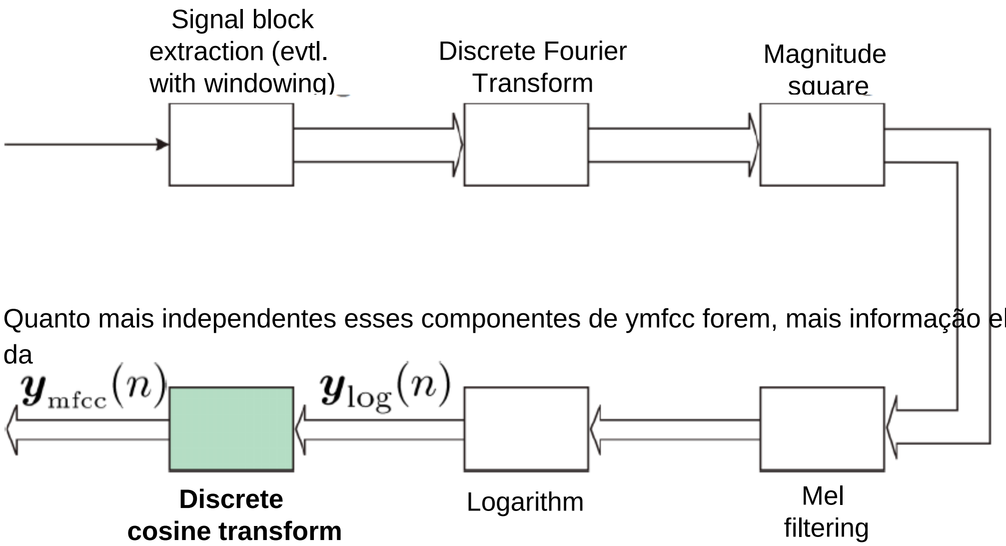
$$\log_e\{\boldsymbol{M}^{\mathrm{T}}\,\boldsymbol{y}_{\text{mel}}(n)\}$$

- Log. Mel coefficients:

$$\boldsymbol{y}_{\log}(n) = \log_e\{\boldsymbol{y}_{\text{mel}}(n)\}$$

Mel-filtered cepstral coefficients (MFCCs)

- Discrete Cosine Transform (DCT):



Quanto mais independentes esses componentes de ymfcc forem, mais informação ele da

- Definition of the most common version of the DCT:

$$X_\mu = 2 \sum_{n=0}^{M-1} x(n) \cos\left(\frac{\pi}{M}\mu(n+\frac{1}{2})\right)$$

$$\mu \in [0, \ldots, M-1]$$

- The real values of

$$\boldsymbol{y}_{\log}(n) = \log_e\{\boldsymbol{y}_{\text{mel}}(n)\}$$

motivate to use a real-valued transform: the DCT.

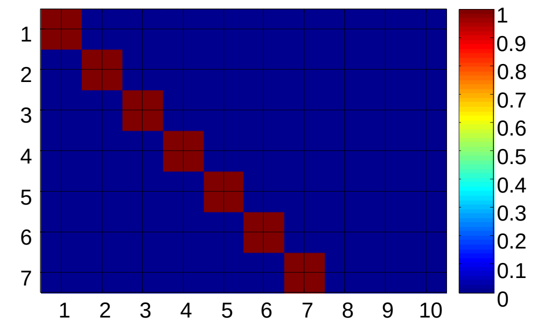
- The transformation can be noted in a matrix vector multiplication:

$$\tilde{\boldsymbol{y}}_{\text{mfcc}}(n) = \boldsymbol{T}_{\cos}\boldsymbol{y}_{\log}(n)$$

- Selection of the first elements of the MFCC vector in order reduce the influence of the fundamental frequency

$$\boldsymbol{y}_{\text{mfcc}}(n) = \boldsymbol{P}\,\boldsymbol{T}_{\cos}\boldsymbol{y}_{\log}(n)$$

$\boldsymbol{P}$:



Mel-filtered cepstral coefficients (MFCCs)

- The DCT generates a decorrelation of the vector elements

- It performs the transform into the cepstral domain where speech spectral envelopes can be efficiently coded.

- In order to analyse the „decorrelating effect“ of the DCT the ACF matrices are calculated. Therefore, first the mean of the feature vectors before and after the DCT is subtracted and they are normalized by the variance:

$$\begin{aligned} \boldsymbol{y}_{\mathrm{log,nor}}(n) &= \boldsymbol{N}_{\mathrm{log}}\left[\boldsymbol{y}_{\mathrm{log}}(n) - \boldsymbol{m}_{\mathrm{log}}\right], \\ \boldsymbol{y}_{\mathrm{mfcc,nor}}(n) &= \boldsymbol{N}_{\mathrm{mfcc}}\left[\boldsymbol{y}_{\mathrm{mfcc}}(n) - \boldsymbol{m}_{\mathrm{mfcc}}\right]. \end{aligned}$$

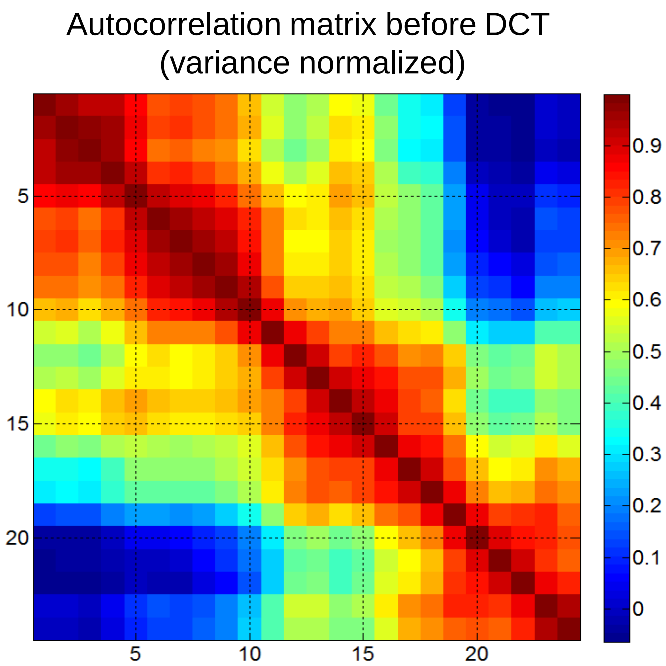
  The normalizing matrices are diagonal matrices with the standard deviations of each vector element on the main diagonal.

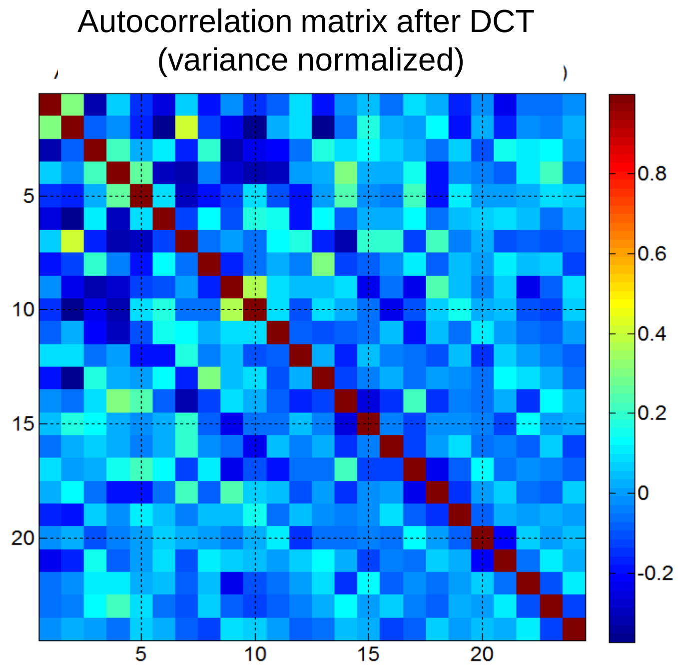
- Then the autocorrelation matrices of the normalized vectors before and after the DCT are calculated:

$$\begin{aligned} \boldsymbol{S}_{\mathrm{log}} &= \mathrm{E}\Big\{\boldsymbol{y}_{\mathrm{log,nor}}(n),\, \boldsymbol{y}^{\mathrm{T}}_{\mathrm{log,nor}}(n)\Big\}, \\ \boldsymbol{S}_{\mathrm{mfcc}} &= \mathrm{E}\Big\{\boldsymbol{y}_{\mathrm{mfcc,nor}}(n),\, \boldsymbol{y}^{\mathrm{T}}_{\mathrm{mfcc,nor}}(n)\Big\}. \end{aligned}$$

Mel-filtered cepstral coefficients (MFCCs)

- Discrete Cosine Transform (DCT):

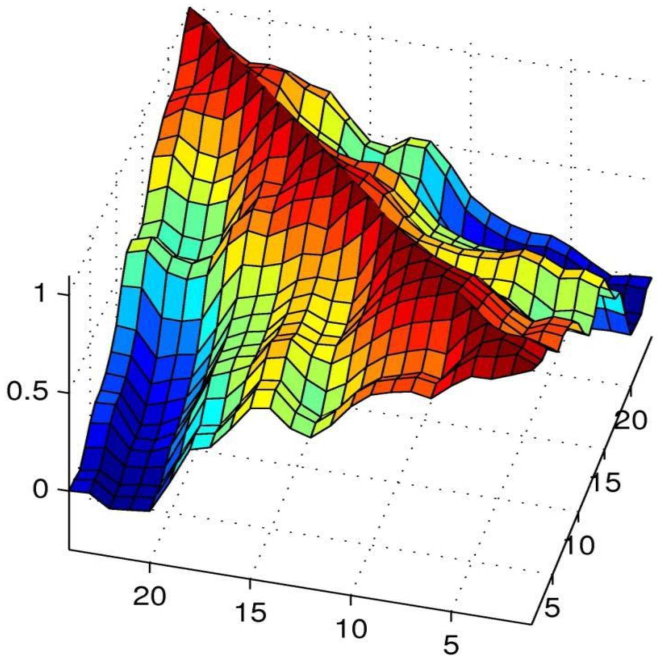




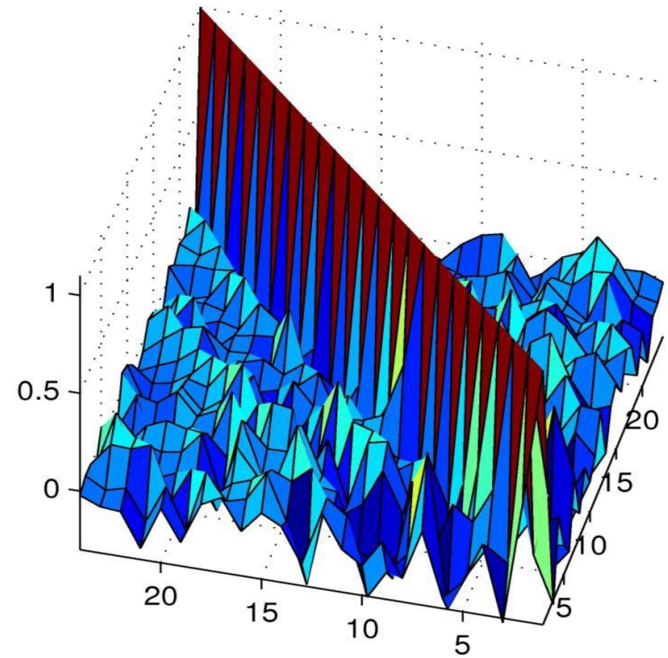
Mel-filtered cepstral coefficients (MFCCs)

- Discrete Cosine Transform (DCT):

Autocorrelation matrix before DCT
(variance normalized)



Autocorrelation matrix after DCT
(variance normalized)



Post processing for the generation of speech features

- **Combination with temporal characteristics of the MFFCs feature:**
  - Delta features: Difference of two consecutive MFCC feature vectors.
  - Delta-Delta features: Difference of two difference vectors.

- **Generation of “super vectors“:**
  - Concatenate consecutive feature vectors (which increases the vector sizes).
  - Typically, the redundancy of the elements of these “super vectors” is severely increased.

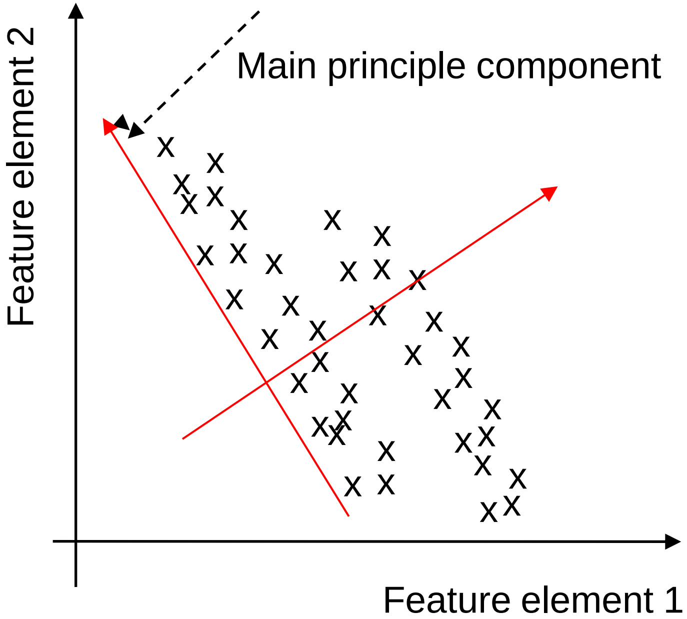
  - Methods for the reduction of the dimension of the feature space can then be applied.

- **Two different methods for the dimension reduction:**
  - **PCA** (Principal component analysis)
    => Orthonormal linear transform such that the data values can be best approximated at a lower dimension. Data driven analysis.
    Also known as Karhunen-Loève Transform (KLT)

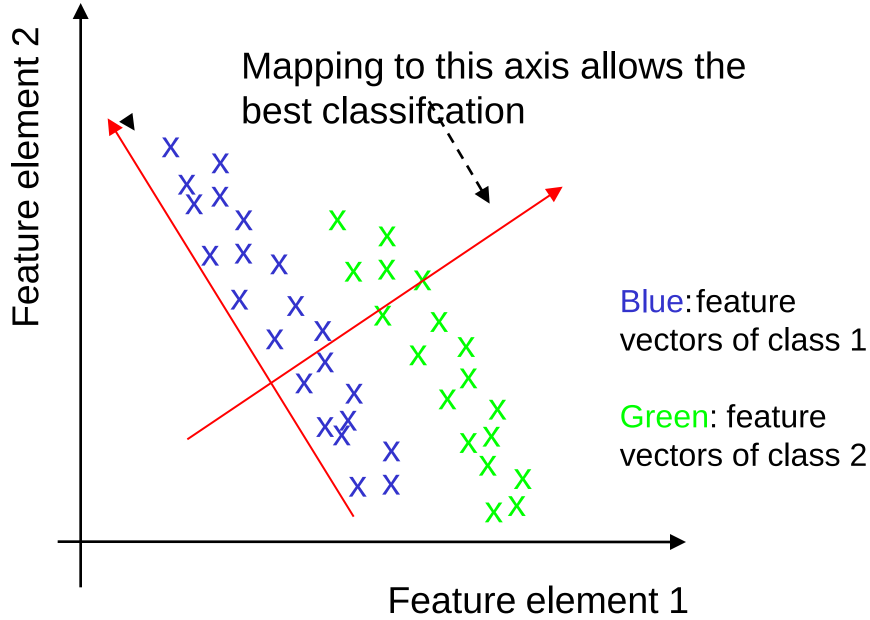
  - **LDA** (Linear discriminant analysis)
    => Projection into a space with lower dimensionality with respect to separate vectors of two different classes the best as possible.

PCA and LDA comparison

- PCA:
  Main axis according to the largest variance:



- LDA:
  Main axis in order to allow the best classification:



Summary

- Motivation for the estimation of the fundamental frequency
- Application scenarios
- Methods for the estimation of the fundamental frequency:
  - Autocorrelation based method
  - YIN procedure
- Voiced / unvoiced classification
- Applications:
  - Audio coding
  - Pitch adaptive post filter for noise reduction
- Cepstral Feature extraction (MFCC: Mel-filtered Cepstral Coefficients) as basis for Speech and Speaker Recognition

- **Next week:** Gaussian mixture models / Bayes decision theory.

References

- [1] L. Arevalo: Beiträge zur Schätzung der Frequenzen gestörter Schwingungen kurzer Dauer und eine Anwendung auf die Analyse von Sprachsignalen, Dissertation, Ruhr-Uni Bochum, April 1991.
- [2] H. Quast, O. Schreiner, and M.R. Schröder: *Robust Pitch Tracking in the Car Environment*, Intl. Conf. on Audio, Speech and Signal Processing (ICASSP), Orlando, USA, 2002.
- [3] Alain de Cheveigne, Hideki Kawahara: *YIN, a fundamental frequency estimator for speech and music*, Journal of the Acoustical Society of America, 111, pp. 1917-1930.
- [4] J. Tilp: *Verfahren zur Verbesserung gestörter Sprachsignale unter Berücksichtigung der Grundfrequenz stimmhafter Laute*, Fortschritt-Berichte VDI, Reihe 10, Nr. 703, Darmstadt, 2002.
- [5] N. Madhu, C. Breithaupt, and R. Martin: Temporal Smoothing of Spectral Masks in the Cepstral Domain for Speech Separation, Intl. Conf. on Audio, Speech and Signal Processing (ICASSP), Las Vegas, USA, 2008.
- [6] C. Breithaupt, T. Gerkmann, R. Martin: A Novel a Priori SNR Estimation Approach Based on Selective Cepstro-Temporal Smoothing, Intl. Conf. on Audio, Speech and Signal Processing (ICASSP), Las Vegas, USA, 2008.
- [7] T. Rosenkranz, H. Puder: Integrating recursive minimum tracking and codebook-based noise estimation for improved reduction of non-stationary noise, Signal Processing, Vol. 92, Issue 3, pp. 767-779, March 2012

References

Cepstral processing:

[1] A.V. Oppenheim, R.W. Schafer: Zeitdiskrete Signalverarbeitung, 2. Auflage, Oldenbourg Verlag, München Wien, 1994

Mel-filtered cepstral coefficients:

[2] E. Schukat-Talamanzzini: Automatische Spracherkennung – Grundlagen, statistische Modelle und effiziente Algorithmen, Vieweg, 1995

[3] L. Rabiner, B.-H. Juang: Fundamentals of Speech Recognition, Prentice-Hall, 1993