Tampere University

# Speech processing and features

SGN 14007

Lecture 8

Annamaria Mesaros

# Content

- Cepstrum
- Pitch detection
- Mel-frequency cepstral coefficients

# Cepstrum

# Analyzing the log-magnitude spectrum

- The source-filter model of speech production views speech spectrum *S(f)* as a product of the vocal tract transfer function *H(f)* and glottal excitation *G(f)*:

  $$S(f) = H(f) \times G(f)$$

  and the corresponding magnitude spectrum is

  $$|S(f)| = |H(f)| \times |G(f)|$$

- Taking logarithm of both sides renders the product into a sum :

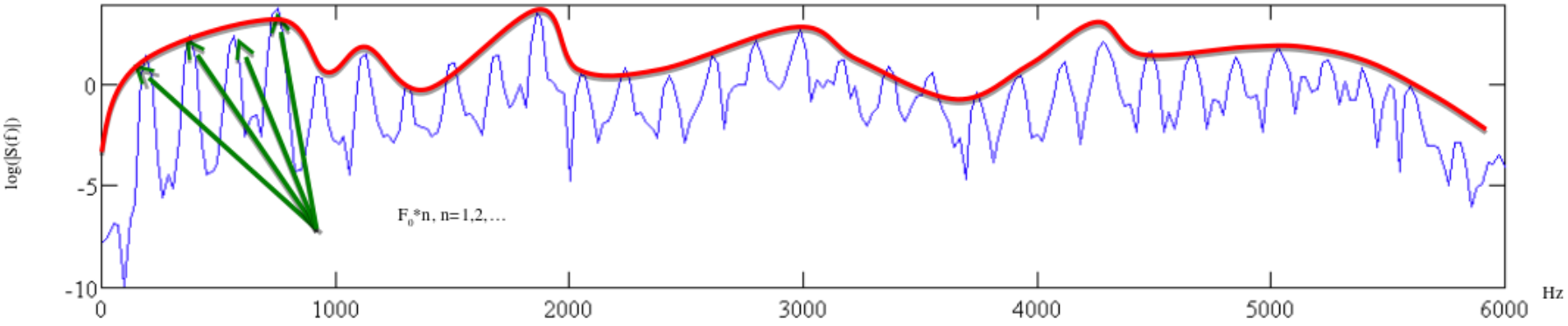
  $$log\ |S(f)| = log|H(f)| + log|G(f)|$$

# Spectral periodicity

- The spectrum of a periodic sound involves frequency components at the fundamental frequency F0 and its whole-number multiples (harmonics) n*F0
  - In the magnitude spectrum, the F0 + harmonics appear as ”high frequency” components.
  - Vocal tract response is a slowly varying component in contrast.

- Cepstral coefficients are obtained by inverse Fourier transform = $F^{-1}\{\}$ of the log-magnitude spectrum:

$$c(k)=F^{-1}\{ \log |S(f)| \} = F^{-1}\{ \log |G(f)| \} + F^{-1}\{ \log |H(f)| \}$$



# From spectrum to cepstrum

- **Cepstrum** is obtained by applying another ”frequency transform” (inverse DFT) on the log-magnitude spectrum.
- The new variable *k* in *c(k)* is not frequency, but time; **’quefrency’** – pseudo time.
- |*G(f)*| can now in principle be removed by **”liftering”** the cepstrum of high-frequency components (liftering / filtering).

# Cepstrum: formal definition

- Definition of cepstrum:

$$c_x[k] = \frac{1}{N}\sum_{n=0}^{N-1} \log\left|S(n)\right| e^{\frac{j2\pi k}{N}n}$$

where *S(n)* is the Fourier spectrum of a speech frame, *n* is the discrete frequency index, *k* is cepstrum coefficient index.

- The inverse Fourier transform of the log-magnitude (Fourier) spectrum
- Because the magnitude spectrum |*S(n)*| of a real-valued signal is always symmetric, the cepstrum is real-valued (imaginary parts of the inverse-Fourier basis vanish and only the cosine terms remain).

# Cepstrum via the cosine transform

$$c_s[k] = \frac{1}{N}\sum_{n=0}^{N-1} \log|S(n)| e^{\frac{j2\pi k}{N}n}$$

$$= \frac{1}{N}\sum_{n=0}^{N-1} \log|S(n)| \left(\cos(\tfrac{2\pi k}{N}n) + i\sin(\tfrac{2\pi k}{N}n)\right)$$

Because S(n) is symmetric, complex sine terms vanish:

$$c_s[k] = \frac{1}{N}\sum_{n=0}^{N-1} \log|S(n)| \cos(\tfrac{2\pi k}{N}n)$$

DCT II cosine transform is defined as:

$$\mathbf{c}_x[k] = \frac{1}{N}\sum_{n=0}^{N-1} x(n)\cos(\tfrac{\pi k}{N}(n+\tfrac{1}{2}))$$

$$\mathbf{c}_{\log|S(n)|}[k] = \frac{1}{N}\sum_{n=0}^{N-1} \log|S(n)| \cos(\tfrac{\pi k}{N}(n+\tfrac{1}{2}))$$

For those who want more details, not important for exam

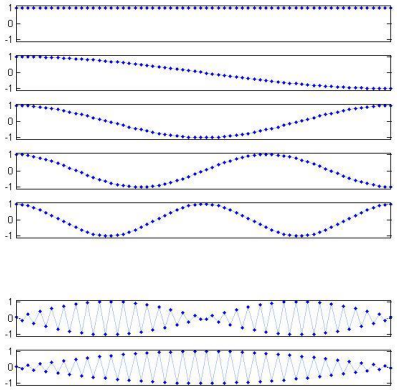
Only small difference freq/2 + phase difference

# DFT vs. DCT

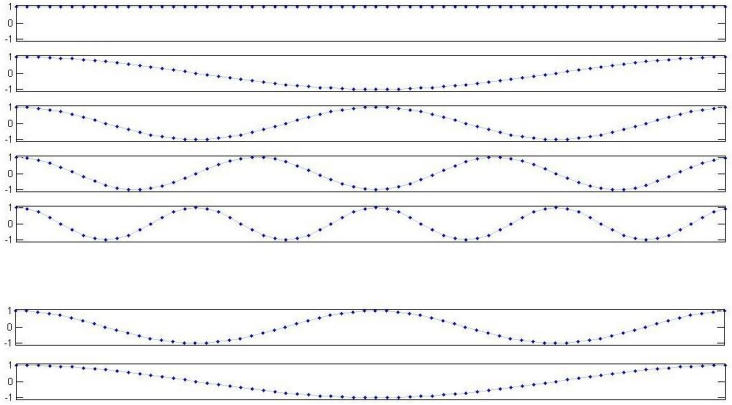
- Fourier transform can be replaced with the cosine transform and thereby reduce the computational complexity.

- The basis functions of the DCT are well-suited to represent a real-valued symmetric signal.

DCT basis functions



Real part of DFT basis functions



# Speech spectrum and cepstrum

- For speech, we assume that the vocal tract exhibits **wide resonance regions** whereas the glottal excitation consists of a harmonic **comb spectrum** (for voiced phonemes).

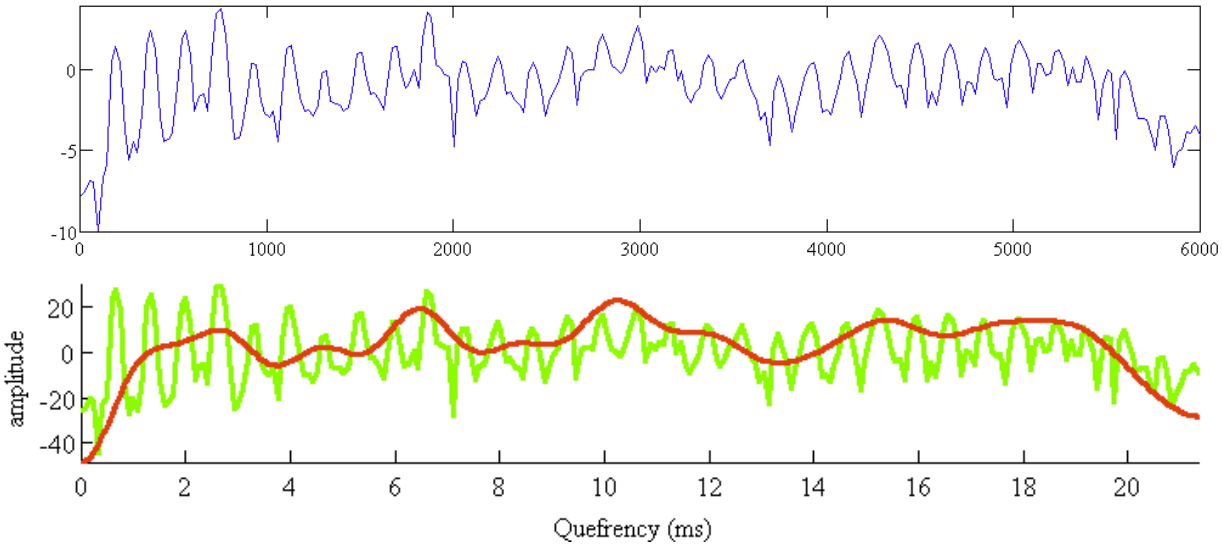
- Therefore if **we frequency transform the spectrum** (to get the cepstrum)…

- …we can think that **smallest cepstral coefficients** (low quefrencies) **correspond to the spectral properties of the vocal tract**

- And **higher-order cepstral coefficients** (large quefrencies) **represent the glottal excitation.**

# Spectral envelope extraction

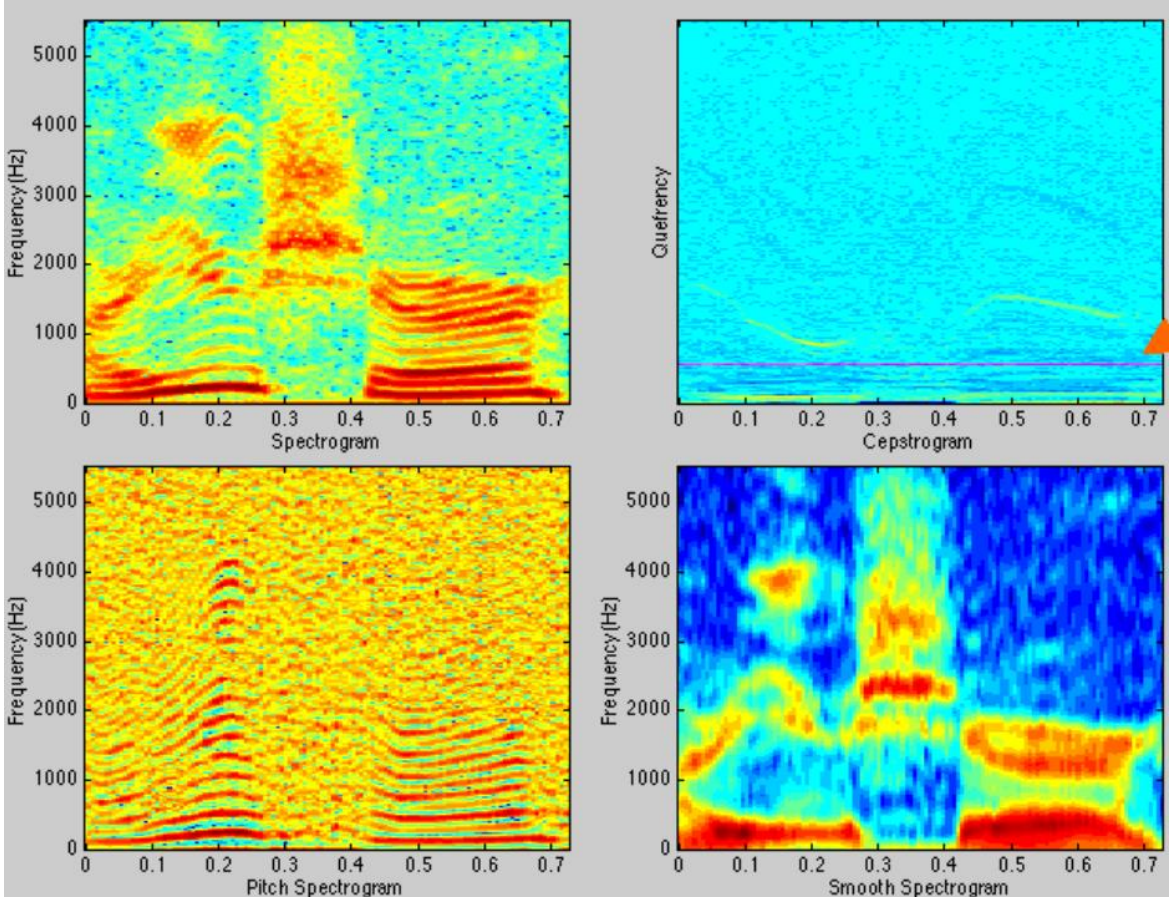
- Removing high quefrency components (liftering), and taking DFT results in spectral envelope of the signal
- Removing low quefrencies and taking DFT we get the spectrum of glottal excitation



# Example

- An estimate of the spectrogram of the glottal excitation is obtained by **zeroing-out the lowest 15-20** cepstral coefficients and going back to spectral domain

- Similarly, an estimate of the time-varying vocal tract response is obtained by **zeroing-out the higher-order cepstral coefficients** and then going back to the spectral domain



# Thinking break (2 min)

# Pitch detection

# Pitch detection

- Perceptual pitch = subjective sensation where one matches a test tone to a sound from which the pitch is estimated.
- Pitch detection: refers to an algorithm for determining the fundamental frequency corresponding to the lowest vibration mode.
- Pitch detection is often called fundamental frequency or F0 estimation
- Speech: The glottal vibration period (in voiced speech)

- Challenges:
  - Wide range (60Hz – 800Hz)
  - Voiced vs. unvoiced detection

# Pitch detection

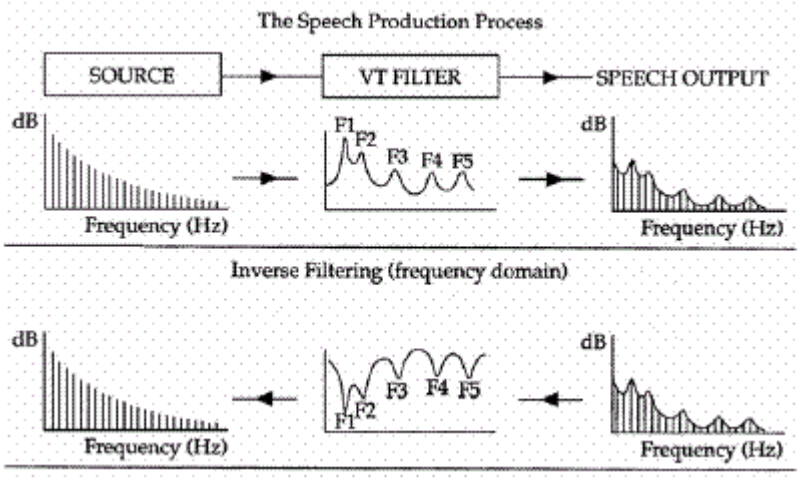
- In speech, changes in pitch are related to prosody (aspects of speech that have duration of groups of syllables or words)
- Pitch variation over sentence = intonation

- Pitch detection approaches
  - Simple Inverse Filter Tracking (SIFT)
  - YIN – a normalized autocorrelation method
    - Close to “normalized autocorrelation method”
  - Cepstral pitch detection
  - Machine learning approaches
    - Pitch detection from a noisy signal is difficult
    - Solution with recurrent neural networks result in state-of-the-art in pitch detection:

      B. Liu, J. Tao, D. Zhang and Y. Zheng, "A novel pitch extraction based on jointly trained deep BLSTM Recurrent Neural Networks with bottleneck features," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

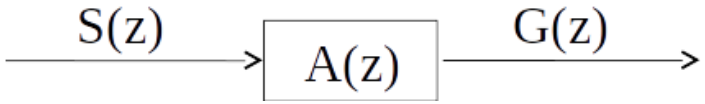
# Simple Inverse Filter Tracking (SIFT)

- Uses the residual signal of LP analysis (i.e. the prediction error)
- Use the LP coefficients (poles) to build a filter with zeros instead of poles.



# Simple Inverse Filter Tracking (SIFT)

- Formants can be removed and the glottal excitation G(z) solved by filtering a speech frame with an inverse filter A(z) that involves the parameters from LP analysis:




$$S(z) = G(z)H(z) = \frac{G(z)}{A(z)}$$

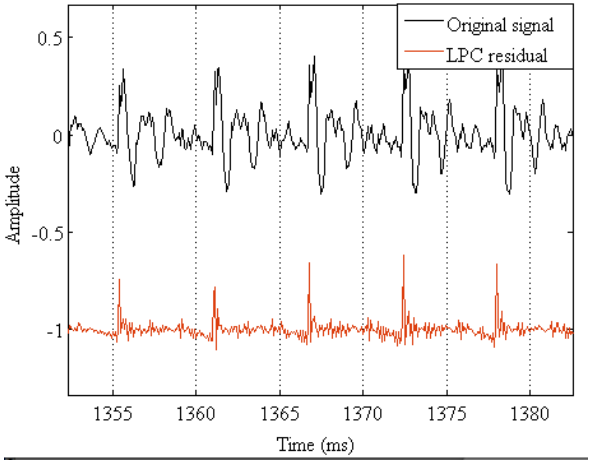
$$=> G(z) = S(z)A(z)$$

- Fundamental frequency can be more reliably computed from the glottal excitation signal G(z) than from speech frame S(z) directly
  - Influence of formants is reduced

- Often autocorrelation function of G(z) is computed within the frame and the maximum of the autocorrelation function is sought in the feasible range of fundamental frequencies

# Simple Inverse Filter Tracking (SIFT)

Glottal excitation signal g(n) using inverse filtering



# Pitch detection with YIN

- Pitch is visible in the autocorrelation function (ACF)
  - τ is time delay in samples, $x_j$ is the jth sample of x
- Peaks will appear at lag values τ with signal periodicity

$$r_t(\tau) = \sum_{j=t+1}^{t+W} x_j x_{j+\tau},$$

- YIN algorithm [1] is a robust F0 estimation method
  - key idea is to obtain a difference function dt'(τ) and find it's minimum index value (in samples)

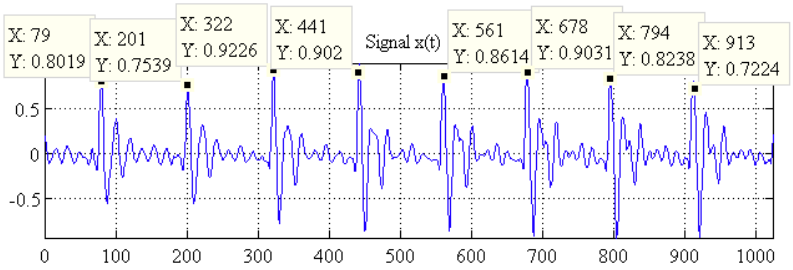
$$d'_t(\tau) = \begin{cases} 1, & \text{if } \tau = 0, \\ d_t(\tau) \Big/ \left[ (1/\tau) \sum_{j=1}^{\tau} d_t(j) \right] & \text{otherwise.} \end{cases}$$

$$d_t(\tau) = \sum_{j=1}^{W} (x_j - x_{j+\tau})^2,$$

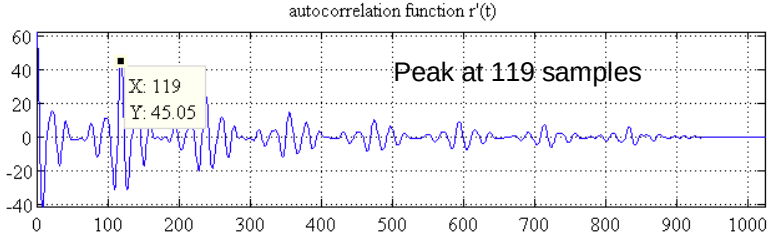
- Then, convert that sample delay T into frequency $F_0$=1/T

[1] A. de Cheveigné and H. Kawahara. YIN, a fundamental frequency estimator for speech and music. The Journal of the Acoustical Society of America, 111:1917, 2002. doi:10.1121/1.1458024

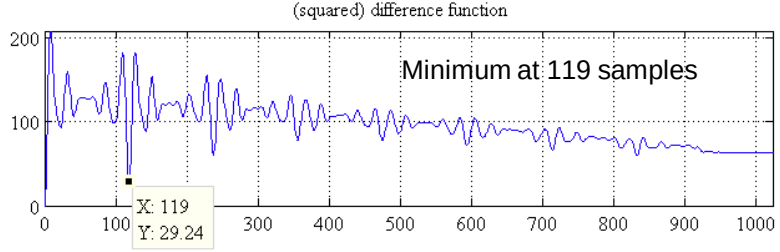
# Example



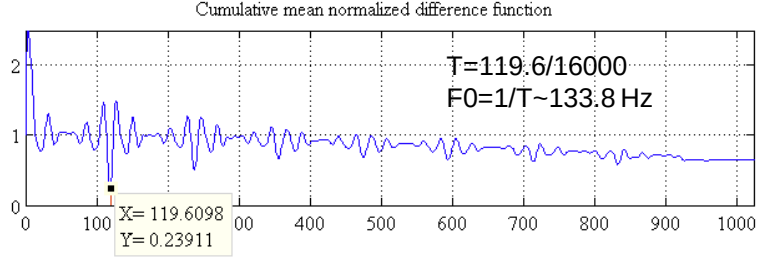

Signal frame $x_j$ at 16 kHz sampling rate of a vowel (1024 samples). Glottal vibration period is visible as interval between adjacent peaks




The modified autocorrelation function $r'_t(\tau) = \sum_{j=t+1}^{t+W-\tau} x_j x_{j+\tau}$.




Difference function $d_t(\tau) = \sum_{j=1}^{W} (x_j - x_{j+\tau})^2$,




Cumulative mean normalized difference function, peak at 119.6 samples with parabolic interpolation

$$d'_t(\tau) = \begin{cases} 1, & \text{if } \tau = 0, \\ d_t(\tau) \Big/ \left[ (1/\tau) \sum_{j=1}^{\tau} d_t(j) \right] & \text{otherwise.} \end{cases}$$

# Thinking break (2 min)

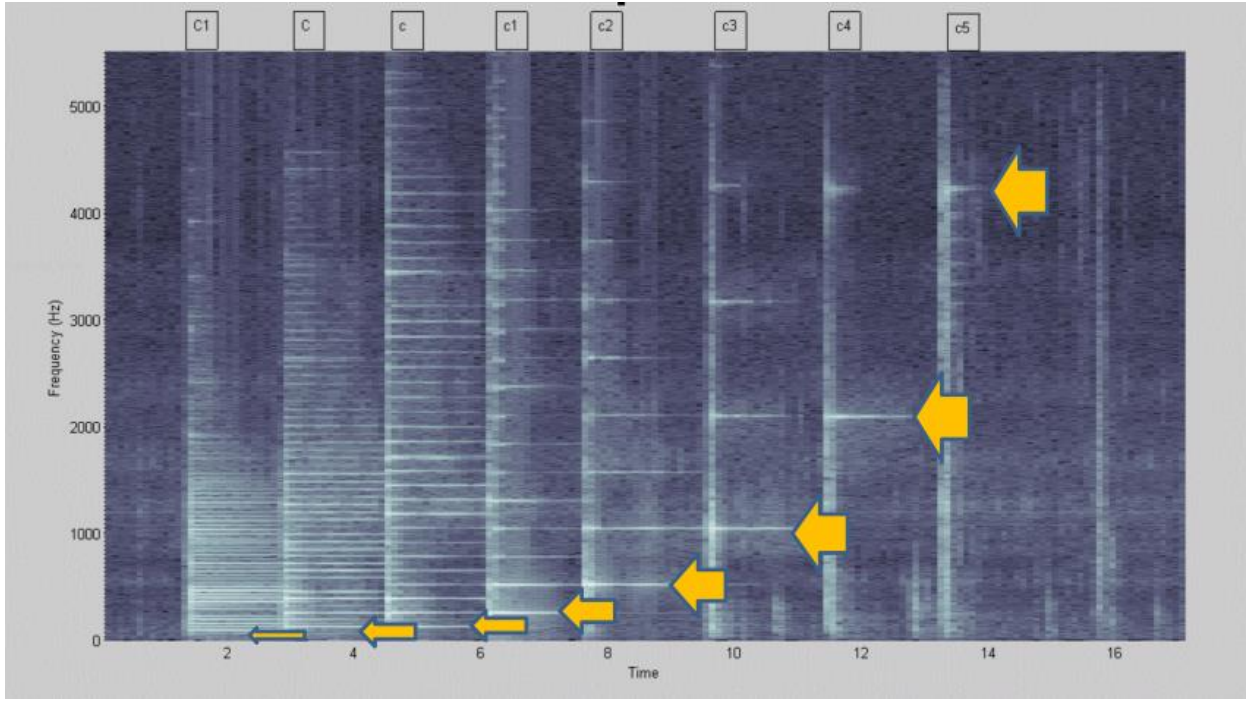
# MFCC

Mel-frequency cepstral coefficients

# Mel-frequency cepstral coefficients

- Model the spectral energy distribution in a perceptually meaningful way
- MFCCs are a widely-used acoustic feature for speech recognition, speaker recognition, and audio classification
  - Although, recent deep learning results in fields of automatic speech recognition already prefer simple mel energies
  - Trend: ”end2end” system, which does not rely on hand crafted features that are classified, but that the machine learning method is able to find a better internal representation of the data.
  - Still, MFCCs are a good starting point (since they are low dimensional)
- MFCCs take into account certain properties of the human auditory system
  - Critical-band frequency resolution (approximately)
  - Log-power (dB magnitudes)

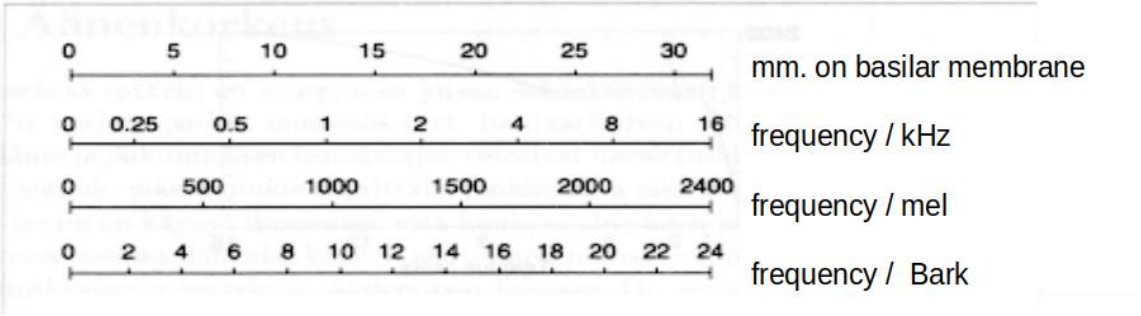
# Spectrogram of piano notes C1 – C8




- Note that the fundamental frequency 16,32,65,131,261,523,1045,2093,4186 Hz doubles in each octave and the spacing between harmonic partials doubles too.
- Such octave change is perceived as ”doubling the height of the note”

# Mel scale

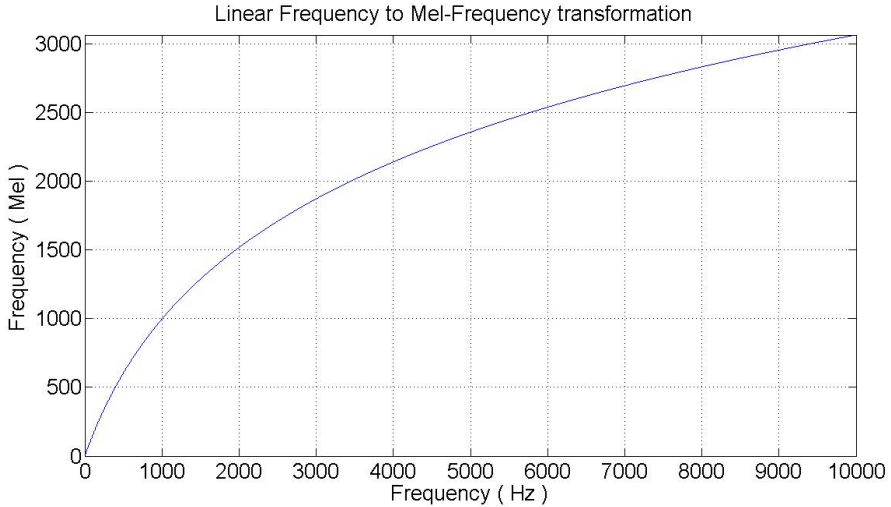
- Mel-frequency scale represents subjective (perceived) pitch. It is one of the perceptually motivated frequency scales
  - See the ”Hearing” lecture slides, where Mel scale was discussed.
  - Models the non-linear perception of frequencies in the human auditory system.
- For comparison, the Bark critical-band scale has been constructed based on the masking properties of nearby frequency components.
  - Constructed by filling the audible bandwidth with adjacent critical bands
- Note that all the scales are related and $f_{Mel} \approx 100 f_{Bark}$ (roughly)



# Mel scale

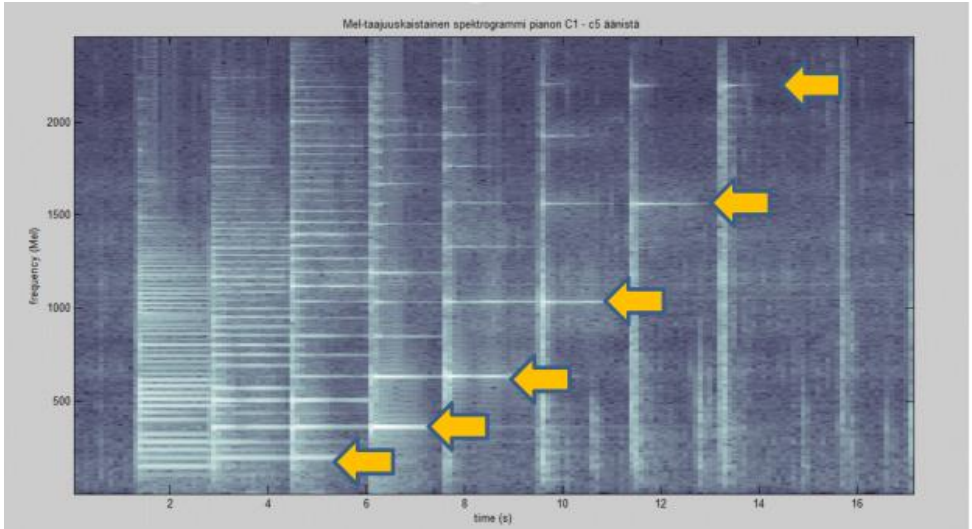
$$f_{Mel} = 2595 \log_{10}(1 + \frac{f_{Hz}}{700})$$

The anchor point for Mel scale is chosen so that 1000 Hz = 1000 Mel

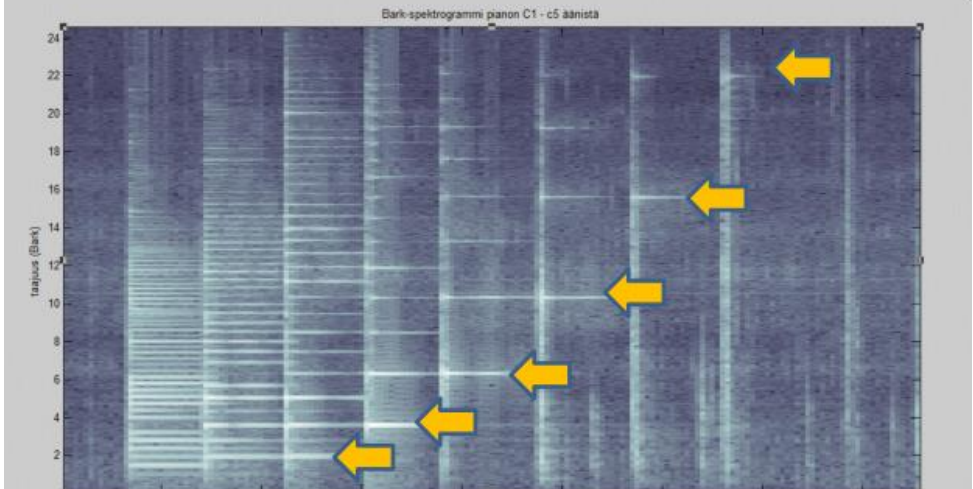


Tampere University

Mel-frequency spectrogram

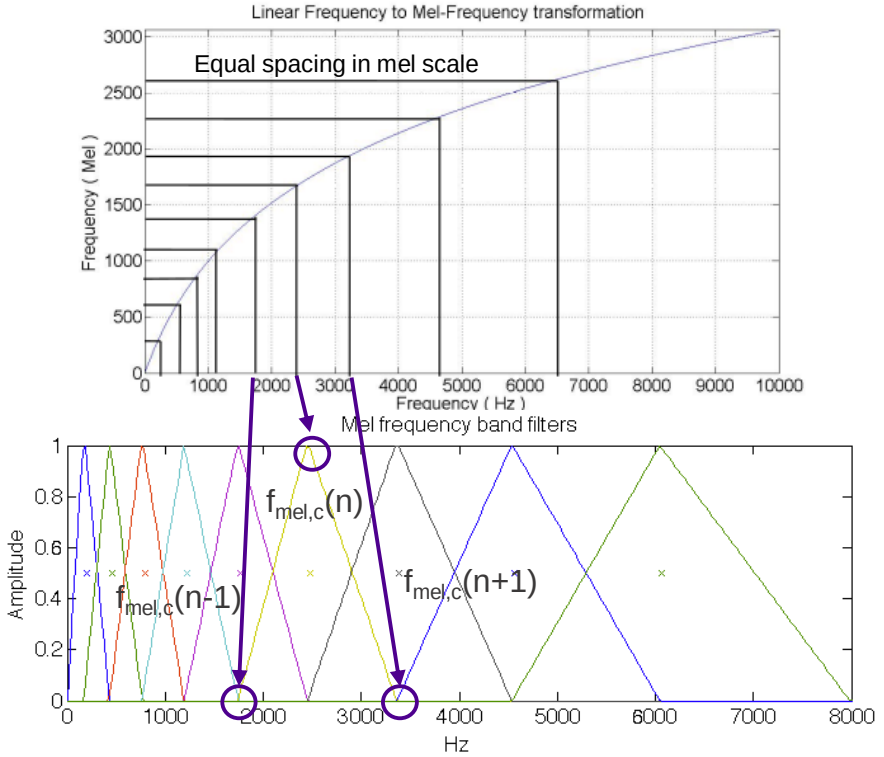



Bark-scale spectrogram



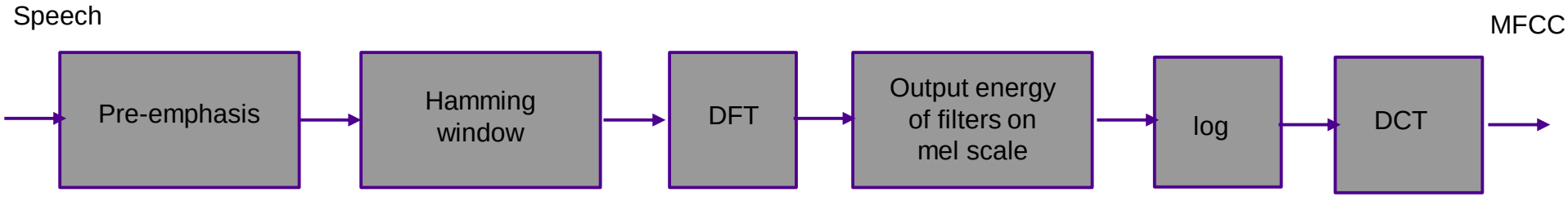
# Mel filterbank

- Triangular ”bandpass filters” uniformly distributed on the mel scale (usually about 40 filters in range 0…8kHz).
  - Note that the mel filter bank has overlap between adjacent frequency bands.
  - The center (mel scale) frequency of band n is $f_{mel,c}(n)$
- Mel filter of band n:
  - starts at 0 amplitude at $f_{mel,c}(n-1)$
  - has maximum amplitude at $f_{mel,c}(n)$
  - decays to zero at $f_{mel,c}(n+1)$
- Note: to have a flat spectrum in mel domain for a DFT magnitude spectrum, the mel bands need to be scaled (bottom figure).



# Mel-frequency cepstral coefficients calculation

- Pre-emphasize the signal, i.e., filter with $H(z)=1-az^{-1}$, $0.95<a<0.99$
- The signal is processed in short windows of x(n).
- Window the short signal x(n) with a window function w(n)
- Take DFT of x(n) → X(f)
- Obtain MFCC
- Proceed to next window



# Spectrum to MFCC

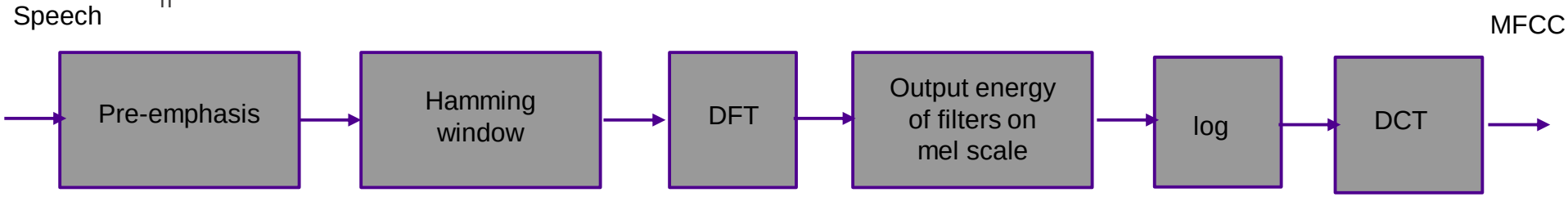
- Define triangular ”bandpass filters” $W_k$, $k$ = 1,…,$K$ uniformly distributed on the mel scale
  - (usually $K$ = 40 filters in range 0…8kHz).
- DFT bin energies $|X(f)|^2$ of each filter are weighted with $k$th band’s filter shape $W_k(f)$ and accumulated

$$E(k) = \sum_f W_k(f)\left|X(f)\right|^2$$

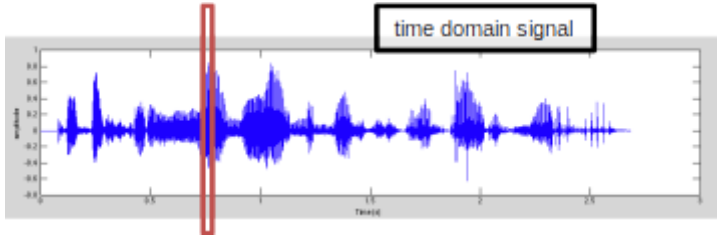
- Take logarithm of each $E(k)$, k=1,2,…$K$
- Calculate discrete cosine transform (DCT II) of log energies

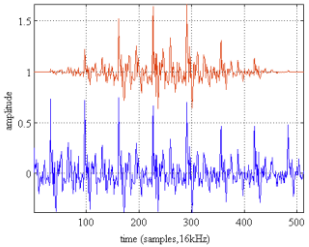
$$c_n = \sum_{k=1}^{K} \log(E(k)) \cos\left[ n\left( k - \frac{1}{2} \right) \frac{\pi}{K} \right], \qquad \text{for} \quad n = 1,\ldots, K$$

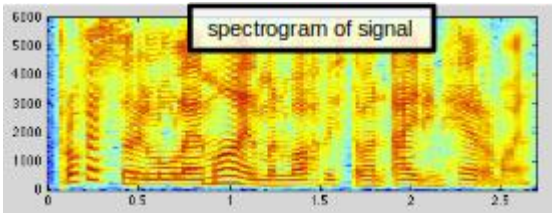
- $c_n$ are called MFCCs



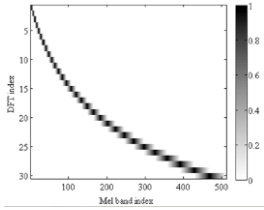
# Log-mel energies



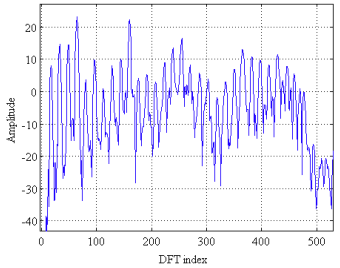



1. Time domain: take one window of data x(n)



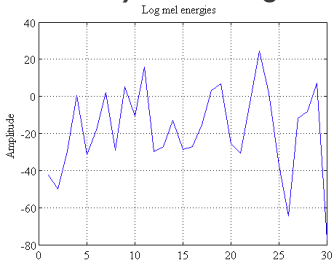
30x512

X



512x1

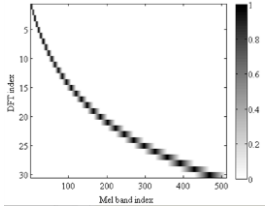
=




30x1

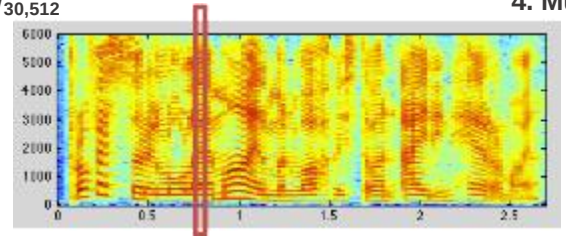
2. Use (pre-emphasis and) windowing

3. Mel scale coefficients in matrix $W_{30,512}$



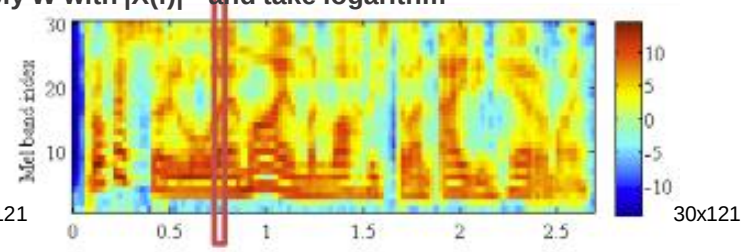
30x512

X



512x121

=



30x121

4. Multiply W with $|X(f)|^2$ and take logarithm

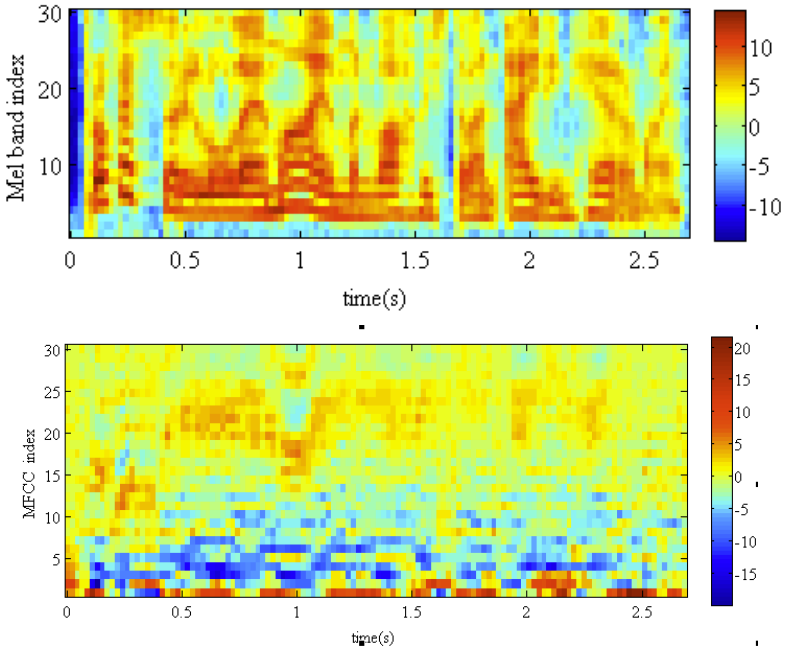
# Log-mel energies to MFCCs

- Apply DCT to log mel energy spectrum of each frame

Log mel spectrum

DCT

MFCC



# Why MFCCs for classification?

- Perceptually-motivated (near log-f) frequency resolution
- Perceptually-motivated decibel-magnitude scale
- Convenient control of the model order:
  - picking only the lowest N coefficients gives lower-resolution approximation of the spectral energy distribution (vocal tract etc.)

- Discrete cosine transform **decorrelates the features** (improves statistical properties by removing correlations between the features)
  - Was highly successful with traditional machine learning tools
  - Current deep learning favors more ”elementary” features (e.g. Mel band energies, STFT bin energies, even raw audio input and let the network find a meaningful representation)
  - MFCCs still a good feature to use (due to low dimensionality)

# Python tools

- http://librosa.github.io/
  - Basic and advanced audio processing functions

- scikits.talkbox
  - LPC implementation (an old library)

# Summary

- Cepstrum processing
  - Relationship to vocal tract shape estimation
- Pitch detection
  - Different methods: inverse filtering, autocorrelation / YIN
- MFCCs - the traditional acoustic feature
  - How to calculate MFCCs
  - Perceptual basis
  - Mel scale