Pasi Pertilä
SGN-14006
2015

SGN-14006 Audio and Speech Processing

Mel-frequency cepstral coefficients (MFCCs) and gammatone filter banks

Slides for this lecture are based on those created by Katariina Mahkonen for TUT course "Puheenkäsittelyn menetelmät" in Spring 2013.

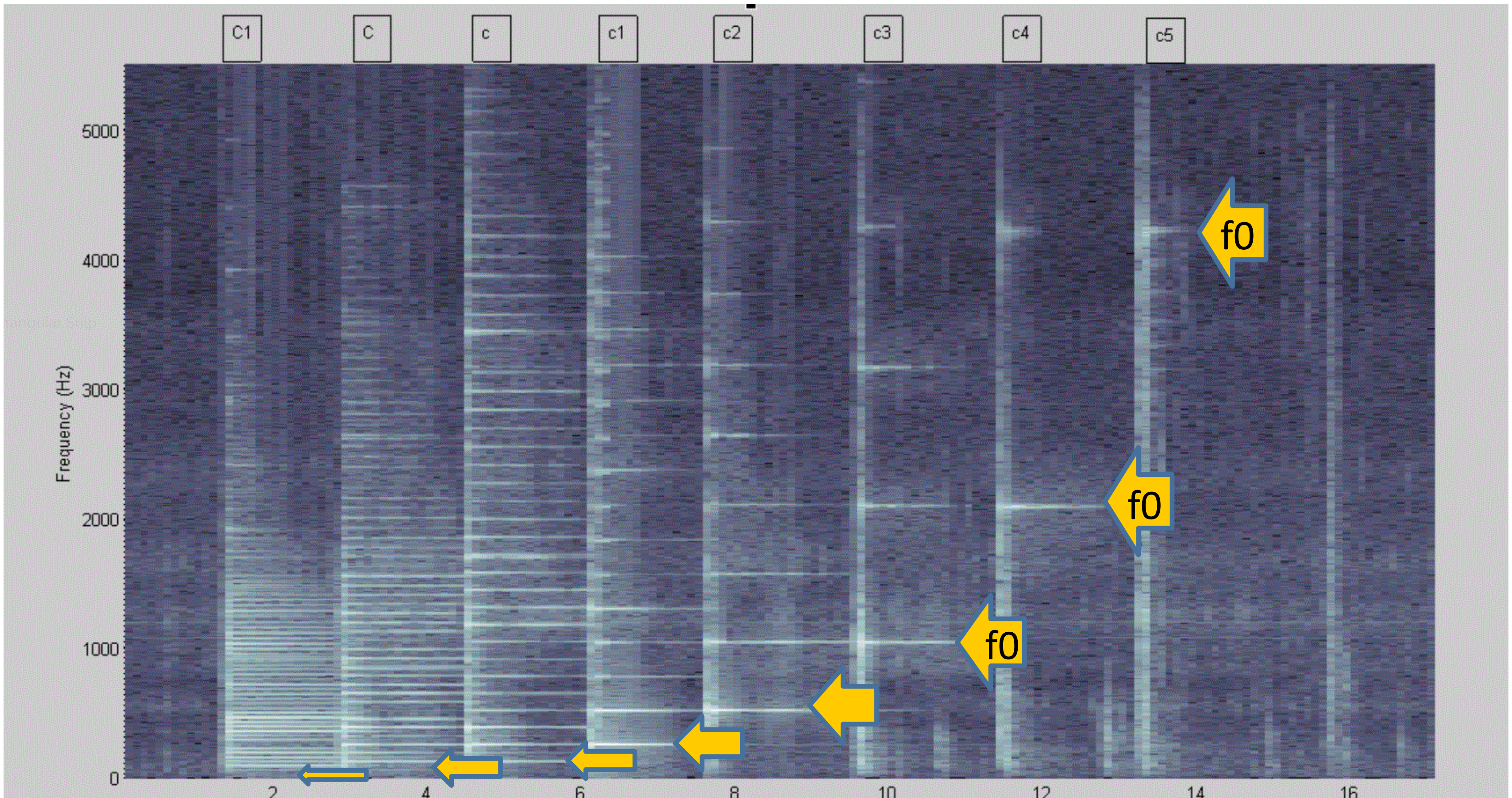
Introduction

- MFCC coefficients model the spectral energy distribution in a perceptually meaningful way

- MFCCs are the most widely-used acoustic feature for speech recognition, speaker recognition, and audio classification

- 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

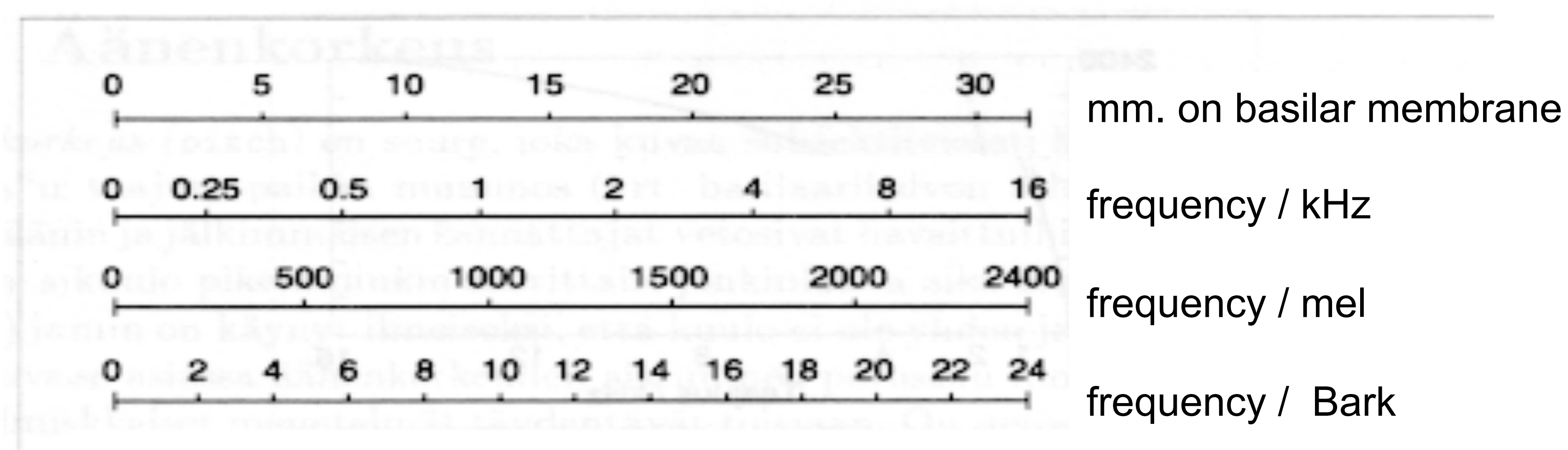


Note that the fundamental frequency 16,32,65,131,261,523,1045,2093,4186 Hz doubles in each octave and the spacing between harmonic partials doubles too.

- Such octave change is perceived as ”doubling the height of the note”

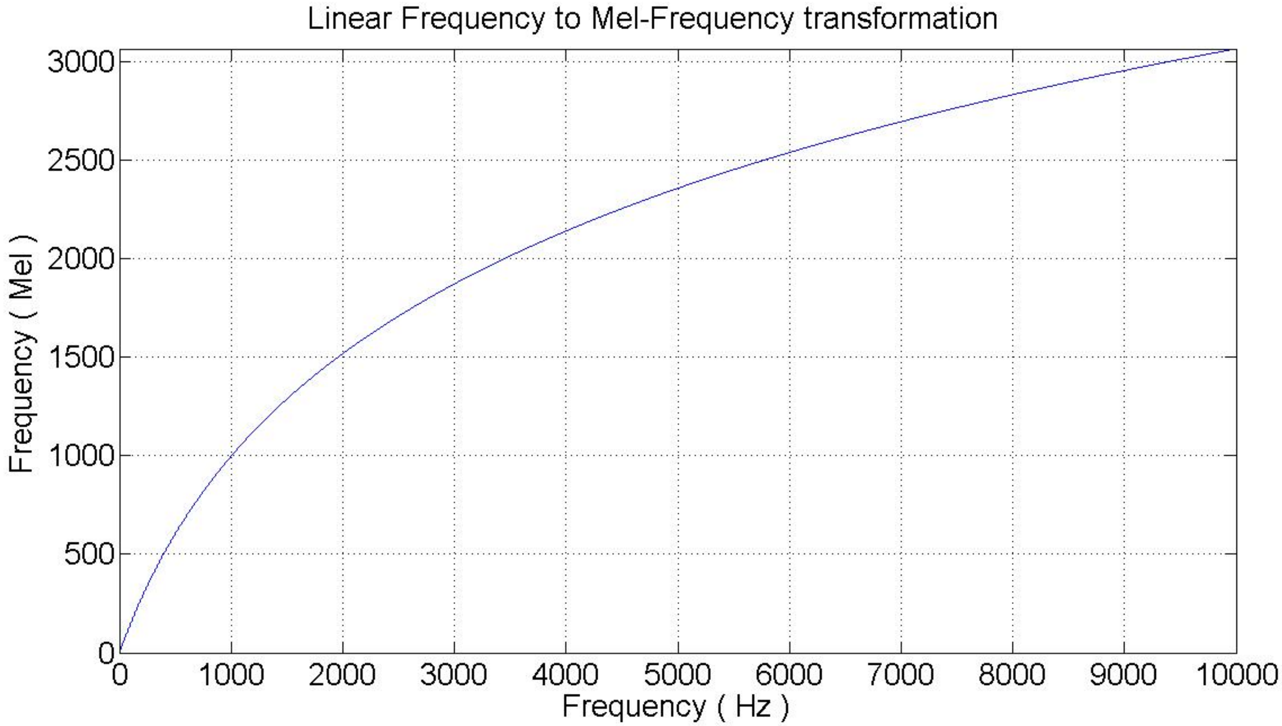
Mel scale

- Mel-frequency scale represents subjective (perceived) pitch. It is one of the perceptually motivated frequency scales (see figure below).
  - Mel-scale is constructed using pairwise comparisons of sinusoidal tones: a reference frequency is fixed and then a test subject (human listener) is asked to adjust the frequency of the other tone to be two times higher or half times lower.
  - 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 1…26
- Note that all the scales are related and: $f_{\mathrm{Mel}} \approx 100 f_{\mathrm{Bark}}$ (very roughly)



Mel scale

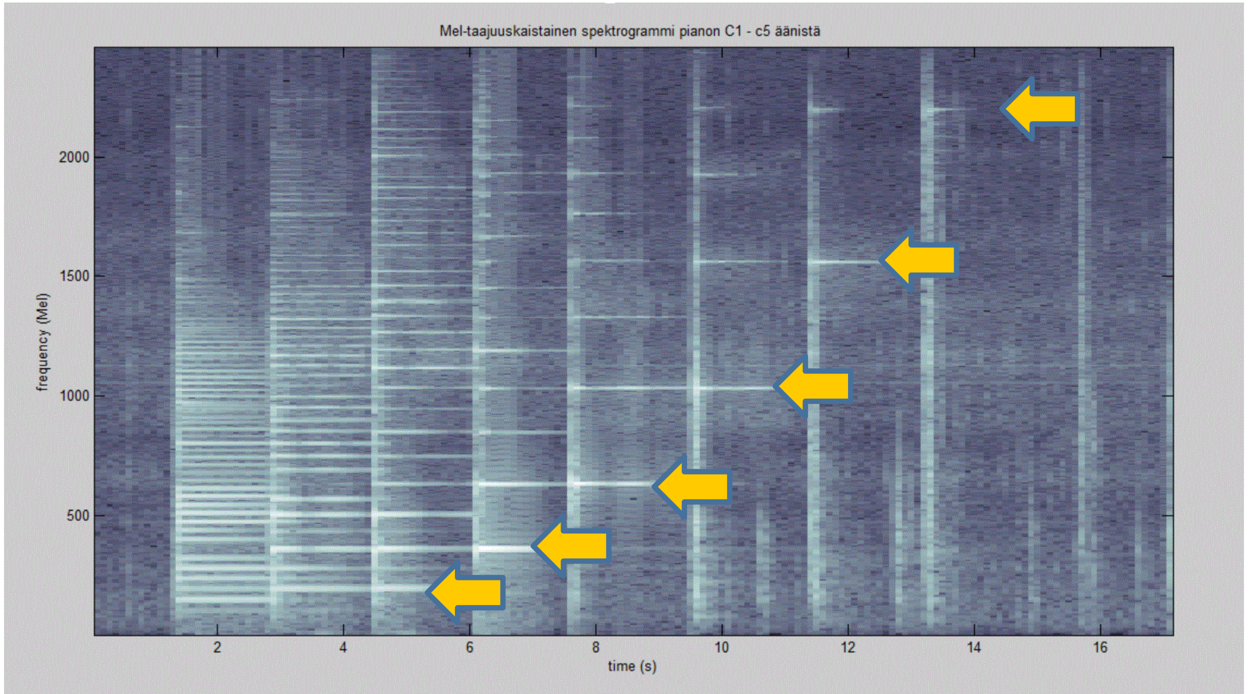
$$f_{Mel} = 2595 \log_{10}(1 + \frac{f_{Hz}}{700})$$



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

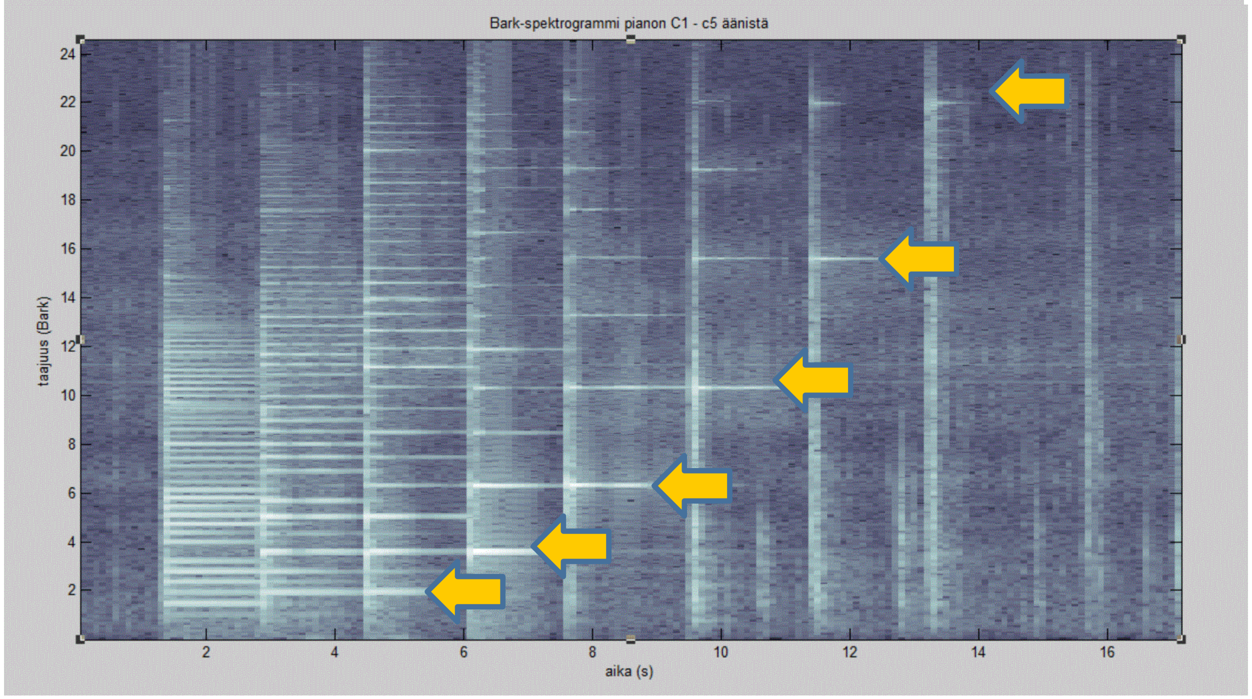
Piano tones C1 – C5

Mel-frequency spectrogram



and

Bark-scale spectrogram



Properties of human hearing – perception of loudness differences

- Weber rule says that the perceived change in a physical quantity is proportional to the relative change:

$$dy \propto \frac{dx}{x}$$

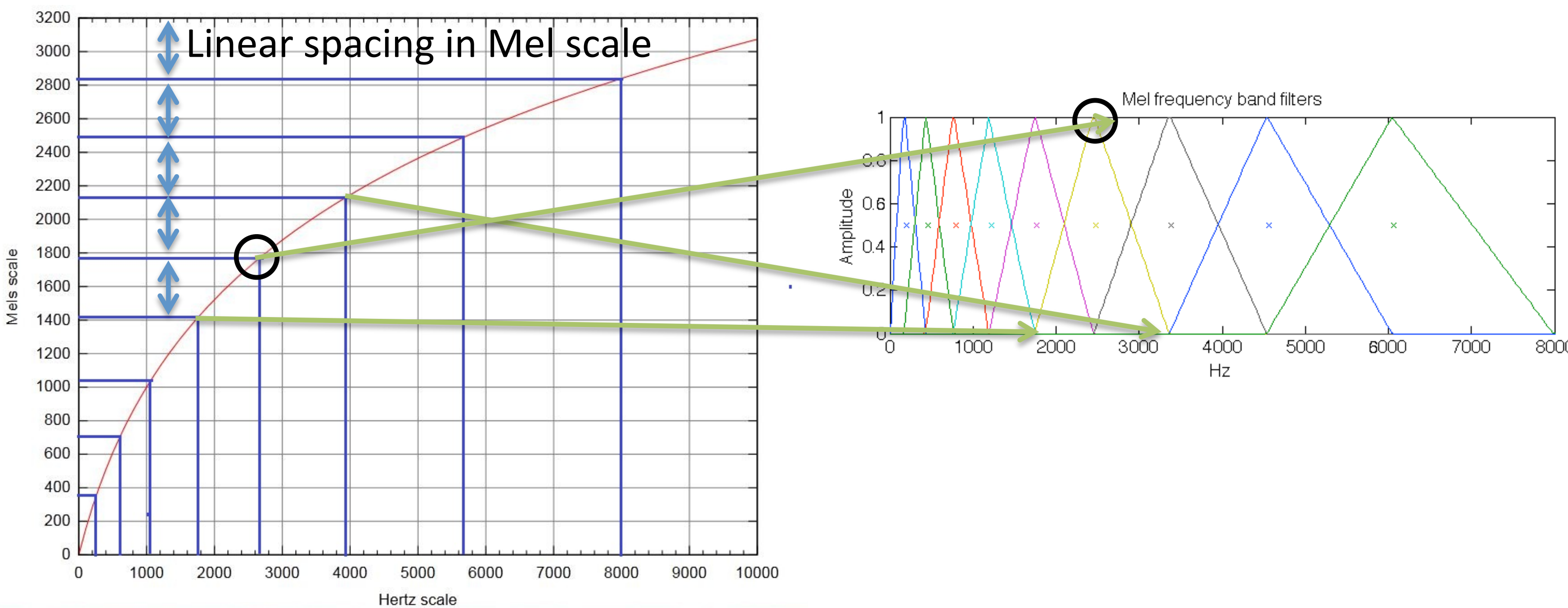
$$y(x) \propto \log(x) + C$$

- Therefore it makes sense to measure sound levels in decibels: $L_I = 10\log10(I)$

Now let's get back to the calculation of MFCC coefficients… The most widely-used acoustic feature used to represent a speech frame (in speech recognition for example)

Calculation of MFCC coefficients

– Define 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 $f_{Mel,c}(n)$ and decays to zero at $f_{Mel,c}(n+1)$

Calculation of MFCC coefficients

- 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

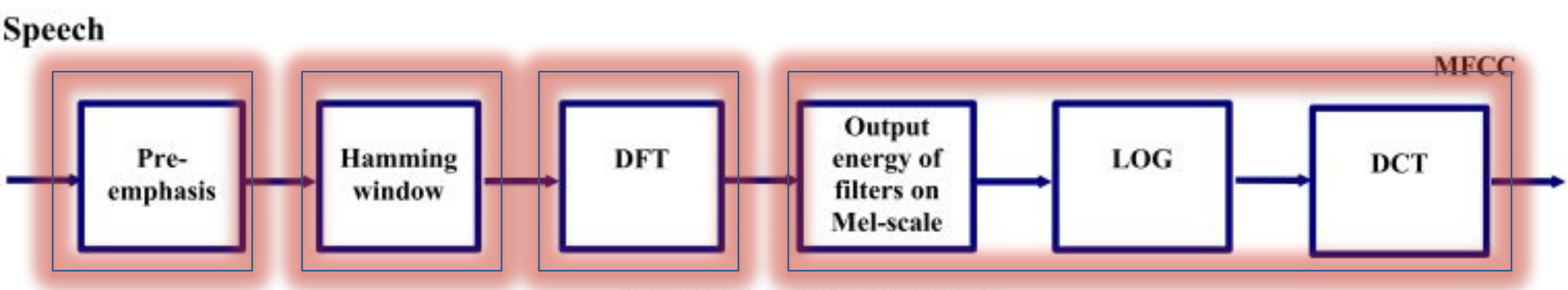


Figure 3- MFCC calculation.

Calculation of MFCC coefficients

– Define triangular "bandpass filters" W_k, $k=1,\ldots,K$ uniformly distributed on the Mel scale (usually $K=40$ filters in range $0\ldots 8kHz$).

– DFT bin energies $|X(f)|^2$ of each filter are weighted with kth 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 $E(k)$, $k=1,2,\ldots K$

– Calculate discrete cosine transform (DCT II) of log energies

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

→ c_n are called MFCCs

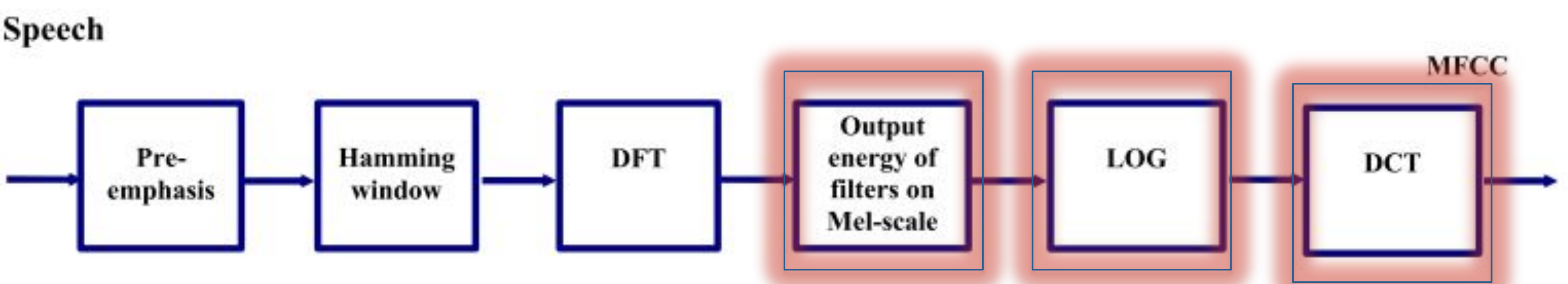
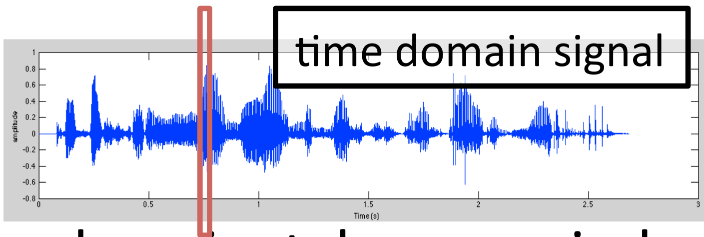
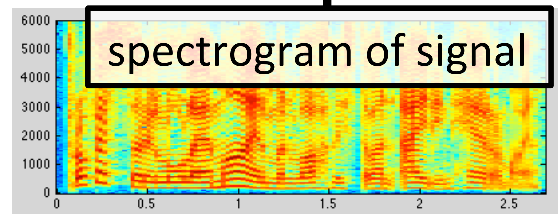


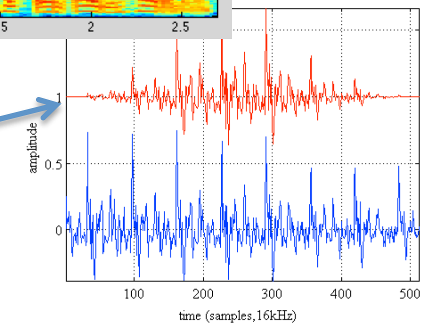
Figure 3- MFCC calculation.

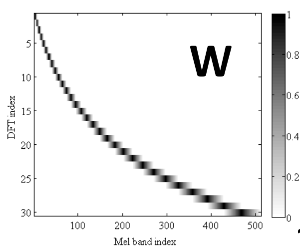
Log-Mel energies example





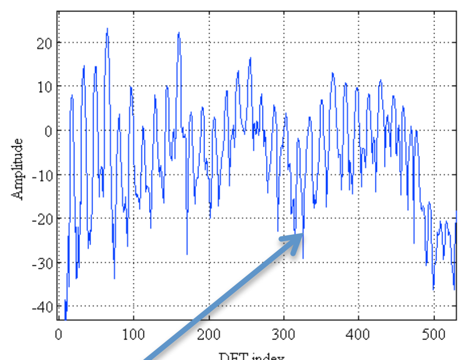
- Time domain: take one window of data x(n)
- Use (pre-emphasis and) windowing
- Mel scale coefficients in matrix $\mathbf{W}_{30,512}$
- Multiply $\mathbf{W}$ with $|X(f)|^2$ and take logarithm





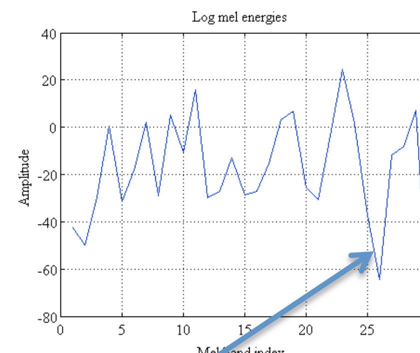
30x512

x

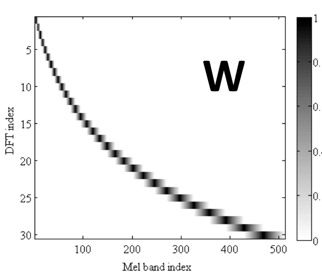


512x1

=

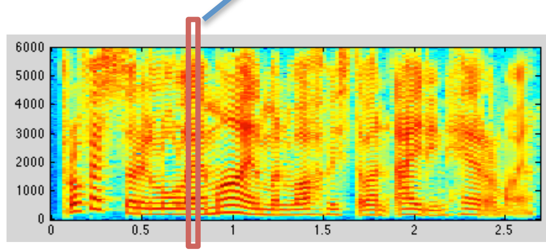


30x1



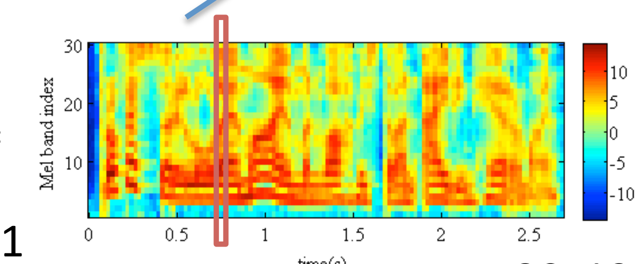
30x512

x



512x121

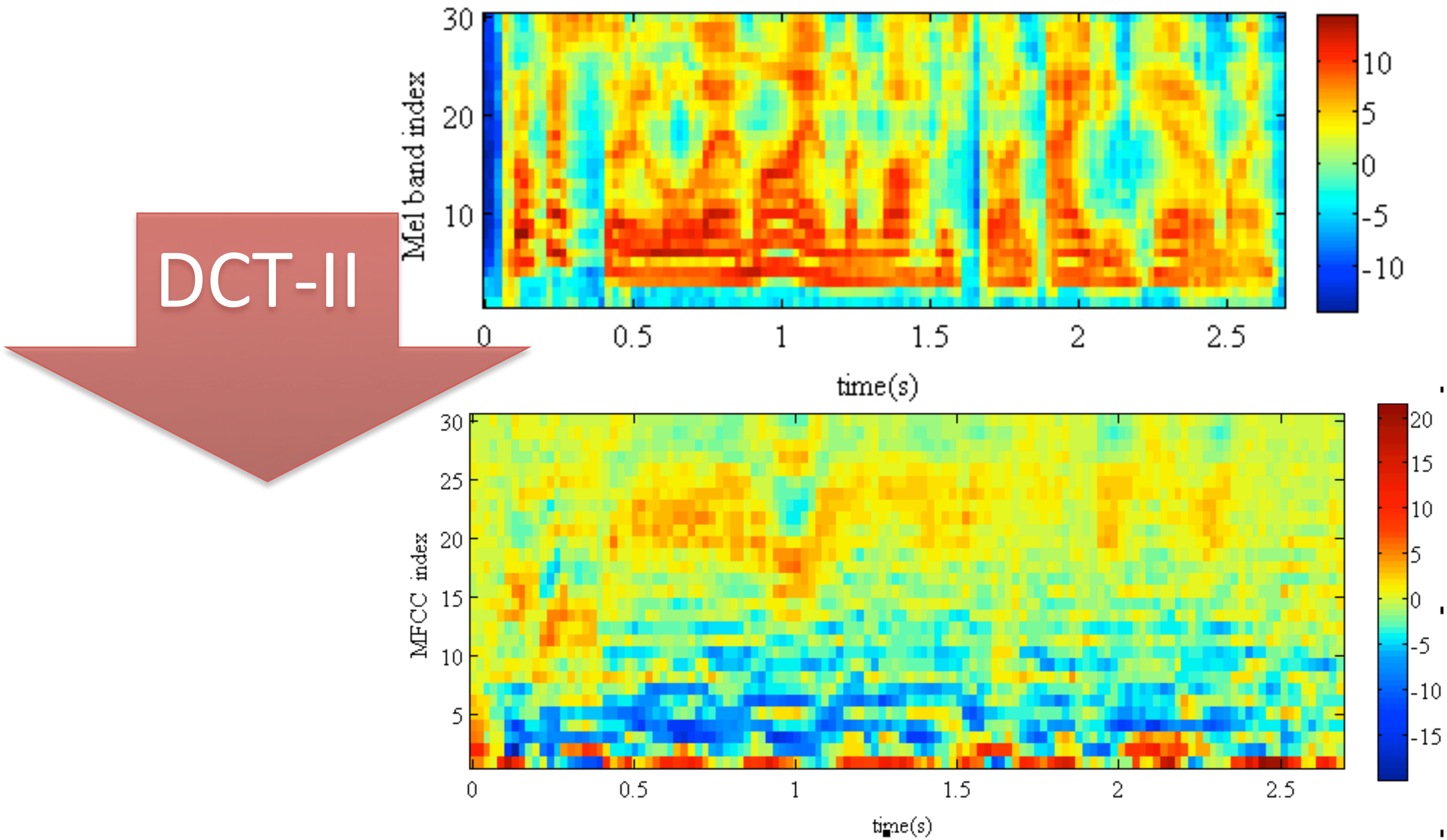
=



30x121

MFCCs from Log-Mel energies example

- Apply DCT to log Mel energy spectrum of each frame



Why are MFCC coefficients successful in audio classification?

– Perceptually-motivated (near log-f) frequency resolution
– Perceptually-motivated decibel-magnitude scale
– Discrete cosine transform decorrelates the features (improves statistical properties by removing correlations between the features)
– Convenient control of the model order: picking only the lowest *N* coefficients gives lower-resolution approximation of the spectral energy distribution (vocal tract etc.)

Gammatone filter bank

- Gammatone filter bank emulates human hearing by simulating the impulse response of the auditory nerve fiber.
- Shape resembles a tone modulated with a gamma-function.

$$g(t) = at^{n-1} \boxed{e^{-2\pi b(f_c)t}} \boxed{\cos(2\pi f_c t + \phi)}$$

- a is peak value, t^{n-1} time onset, $exp()$ –term defines bandwidth and decay, f_c is characteristic frequency, and ϕ is initial phase.
- Typically 42 bands, from 30Hz to 18kHz
- Drawback: Does not emulate level-dependent characteristics of auditory filters.

Example: Gammatone filter bank

http://ltfat.sourceforge.net/

a) response shape (time domain

b) magnitude responses

of 40 filters on ERB scale

c) log output of 30 filters

d) conventional

spectrogram