Feature extraction of speech with Mel frequency cepstral coefficients (MFCCs)

Generic audio classifier

Speech pressure wave

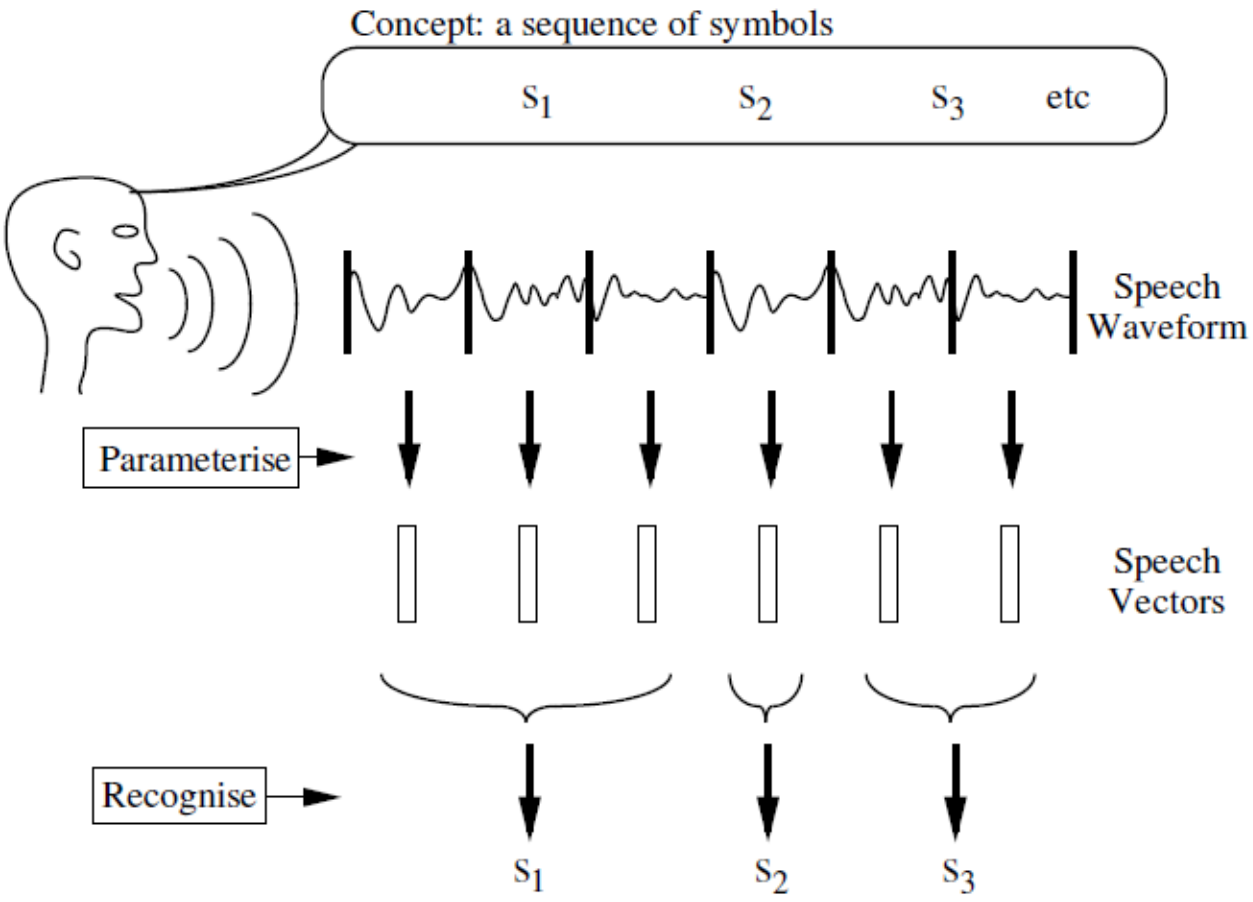
Feature extraction

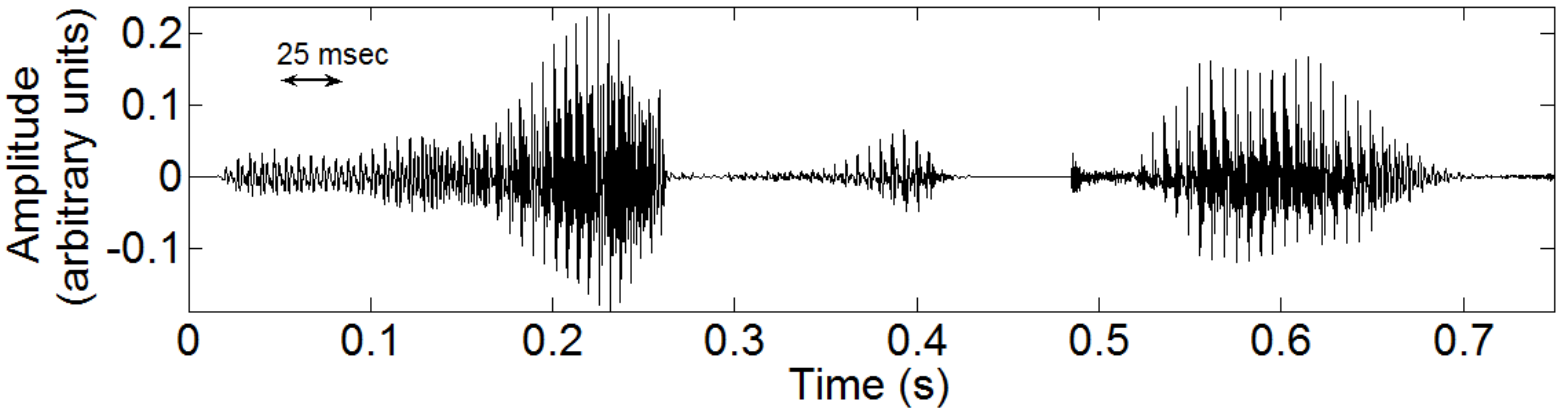
Classifier

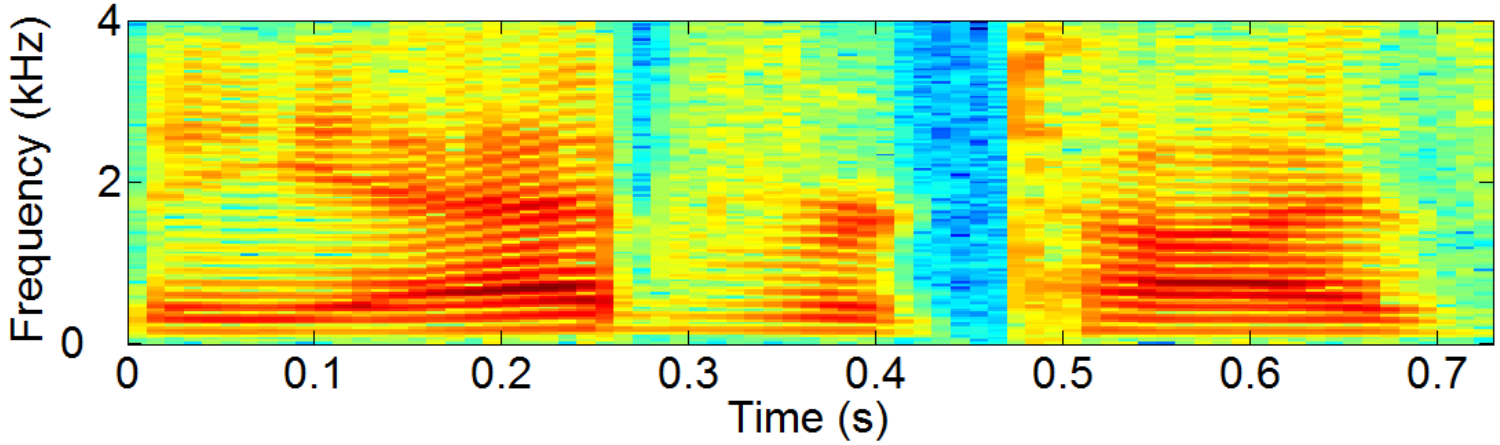
Predicted class(es)

Stored acoustic models

Short-term speech processing

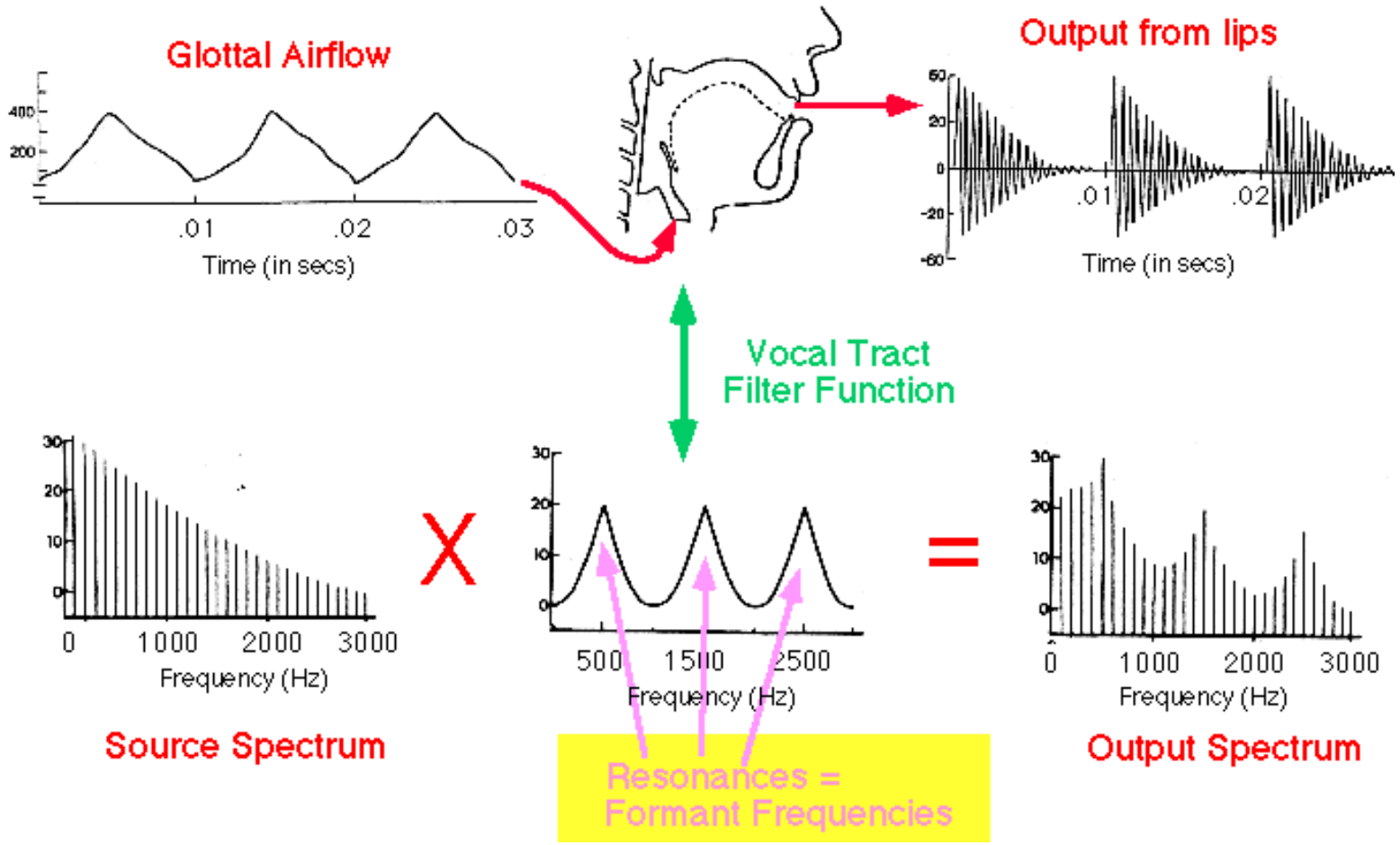




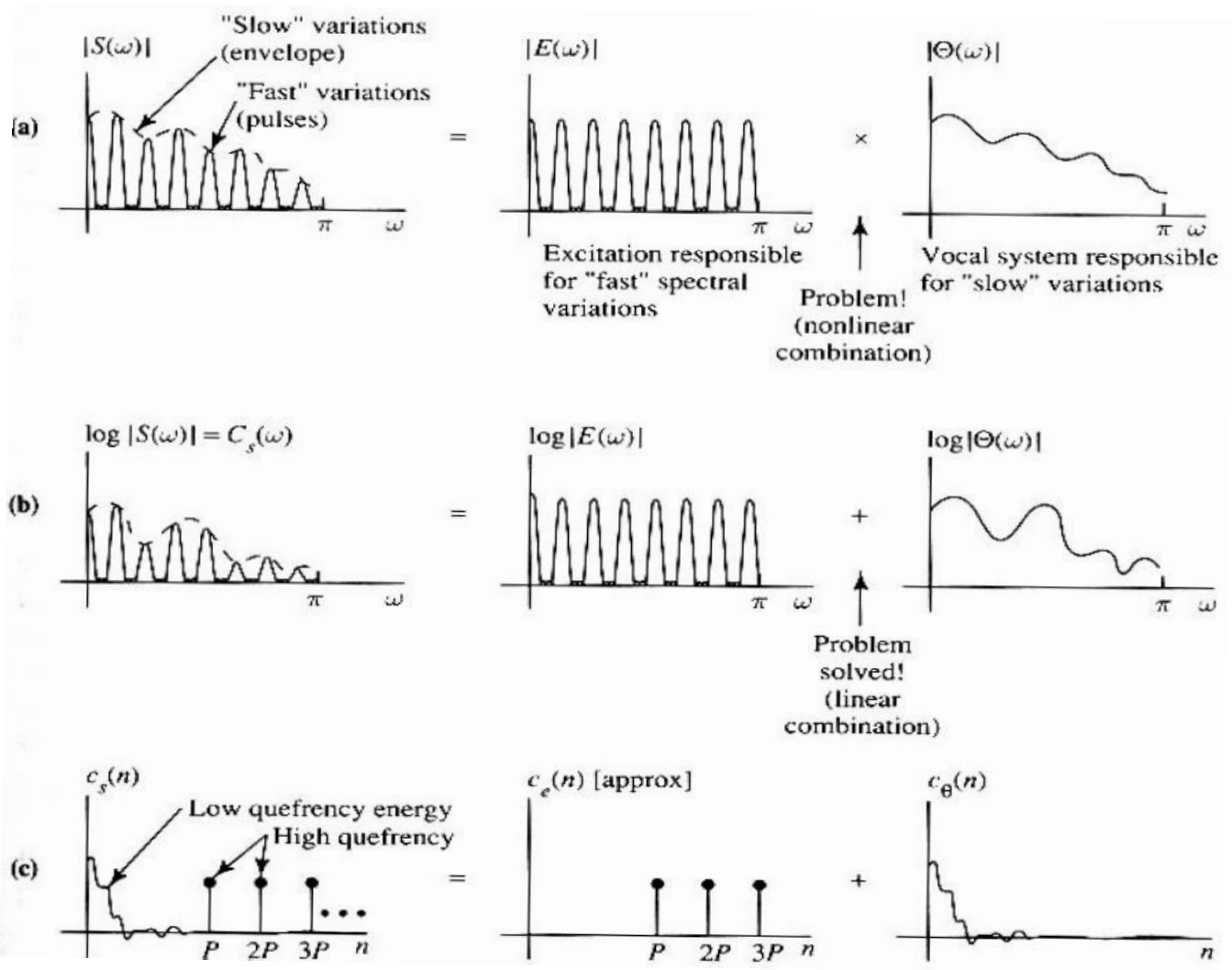


$$X[k] = \left| \sum_{n=0}^{N-1} x[n] w[n] e^{-2\pi j n k / N} \right|^2$$

The source and the filter

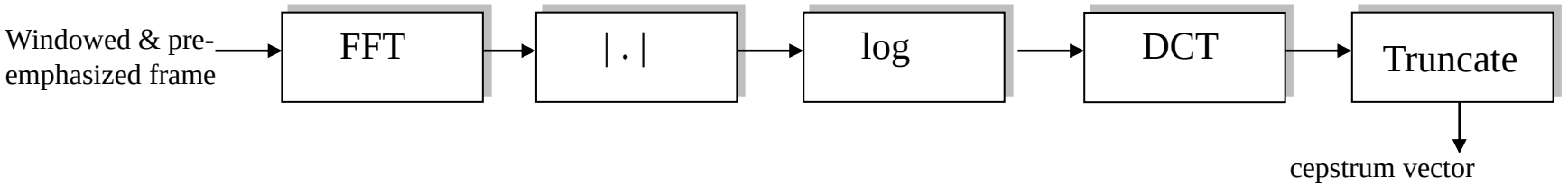


Intuition to cepstral processing



In cepstral processing, we are interested in the “frequency content” of the magnitude spectrum. We apply Fourier techniques to the magnitude spectrum, just like we usually apply them to time-domain signals.

From theory to practice

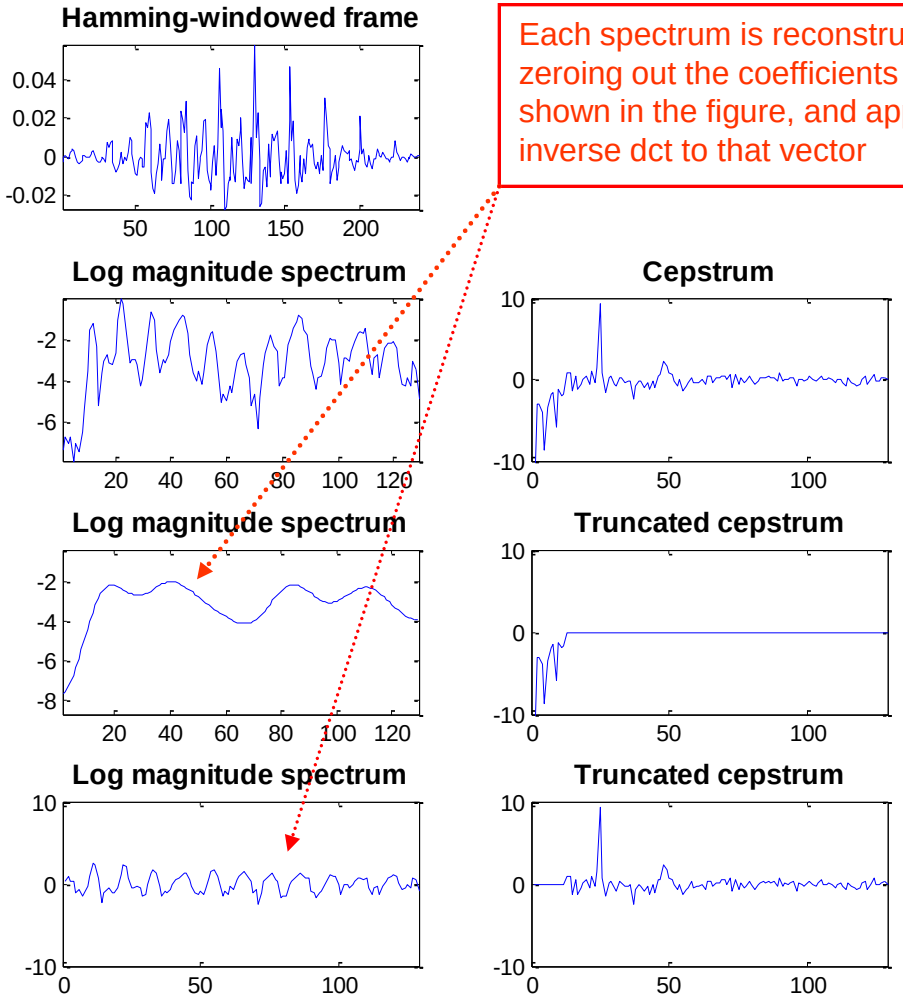


➔The inverse Fourier transform is replaced by the discrete cosine transform (DCT)

```
% frame is a short-term chunk of speech
NFFT = 2^nextpow2(length(frame));
frame = filter([1, -0.97],[1],frame);
frame = frame .* hamming(length(frame));
spectrum = fft(frame,NFFT);
logmagspec = log(abs(spectrum(1:NFFT/2+1))+eps);
cepstrum = dct(logmagspec);
cepstrum = cepstrum(2:numcoefficients+1);
```

We retain only a few of the lowest coefficients. Note that we drop c[0] since it is proportional to log-energy and therefore depends on the distance to microphone, the vocal effort of the speaker, etc.

Matlab demo: Reconstruction of spectra from cepstral coefficients

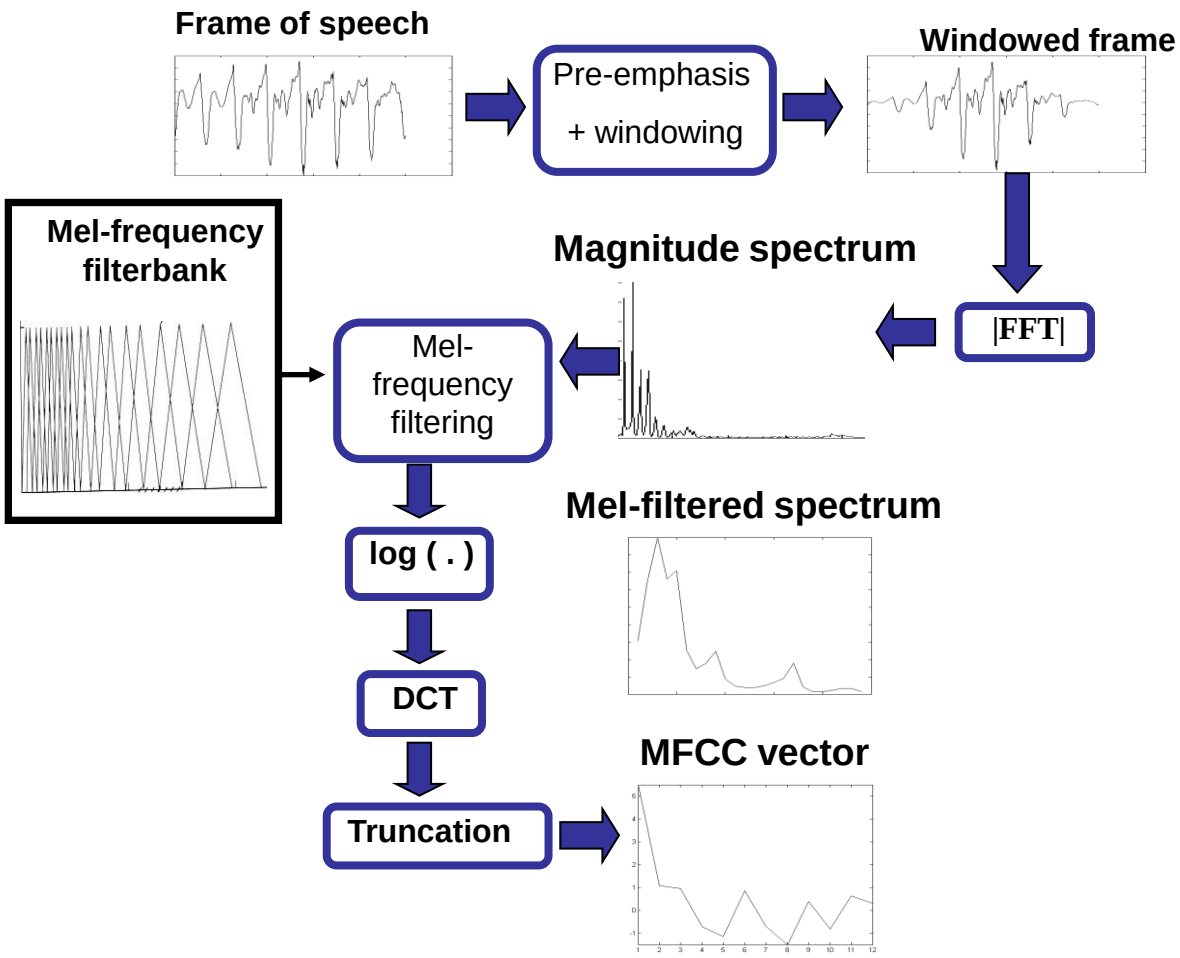


Lower coefficients ➔ spectral envelope, smooth spectral variations, the ”low-frequency” component of the log magnitude spectrum.

Higher coefficients ➔ the ”periodic” and other ”high-frequency” components of the log magnitude spectrum.

Mel-frequency cepstral coefficients (MFCCs)

- A combination of a psycho-acoustically motivated mel-frequency warped filterbank and cepstrum



Dimensionality of the intermediate vectors:

Frame: *N*=240 samples

↓

FFT magnitude spectrum:
NFFT/2+1=129 bins

↓

Mel filter outputs:

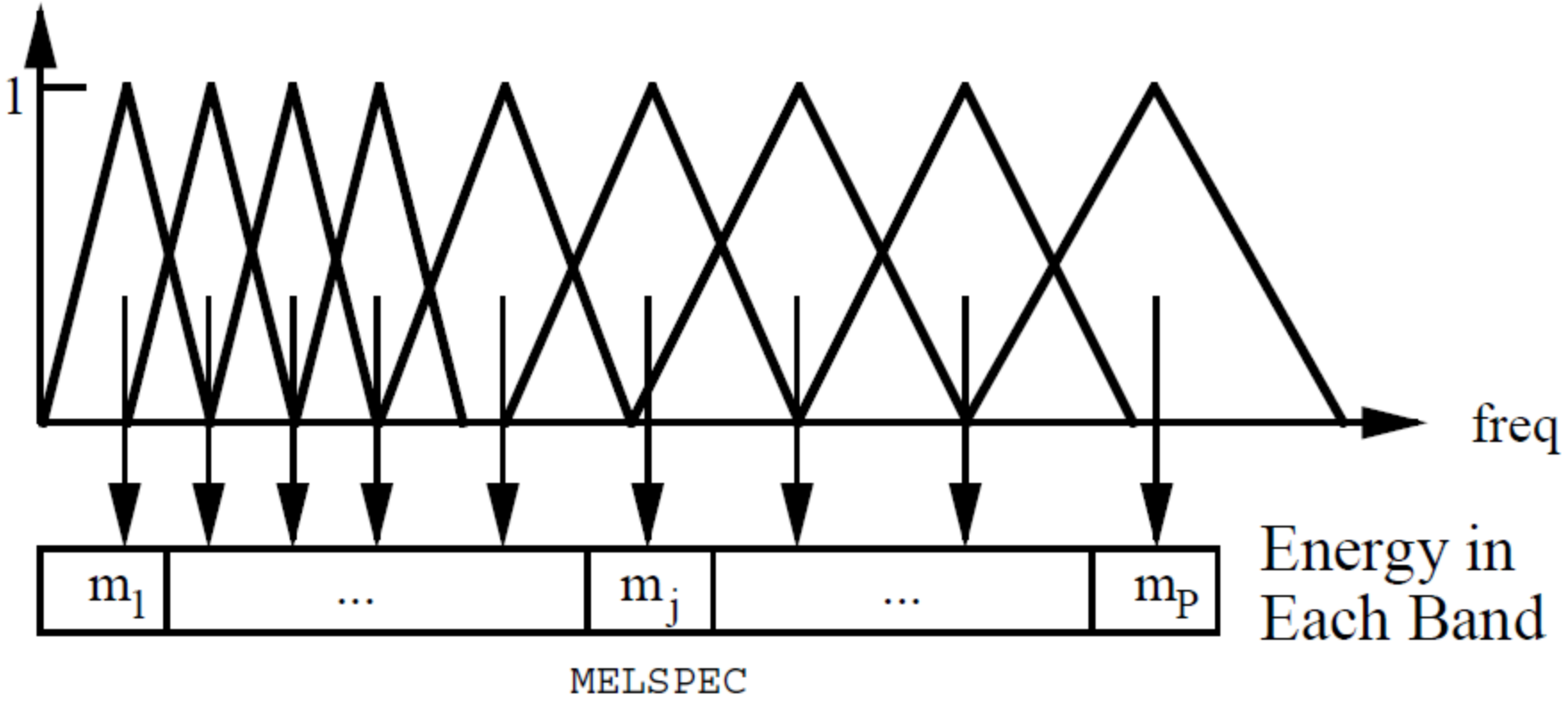
M=27 filters

↓

MFCC vector:

d=12 coefficients

Mel-scale filterbank

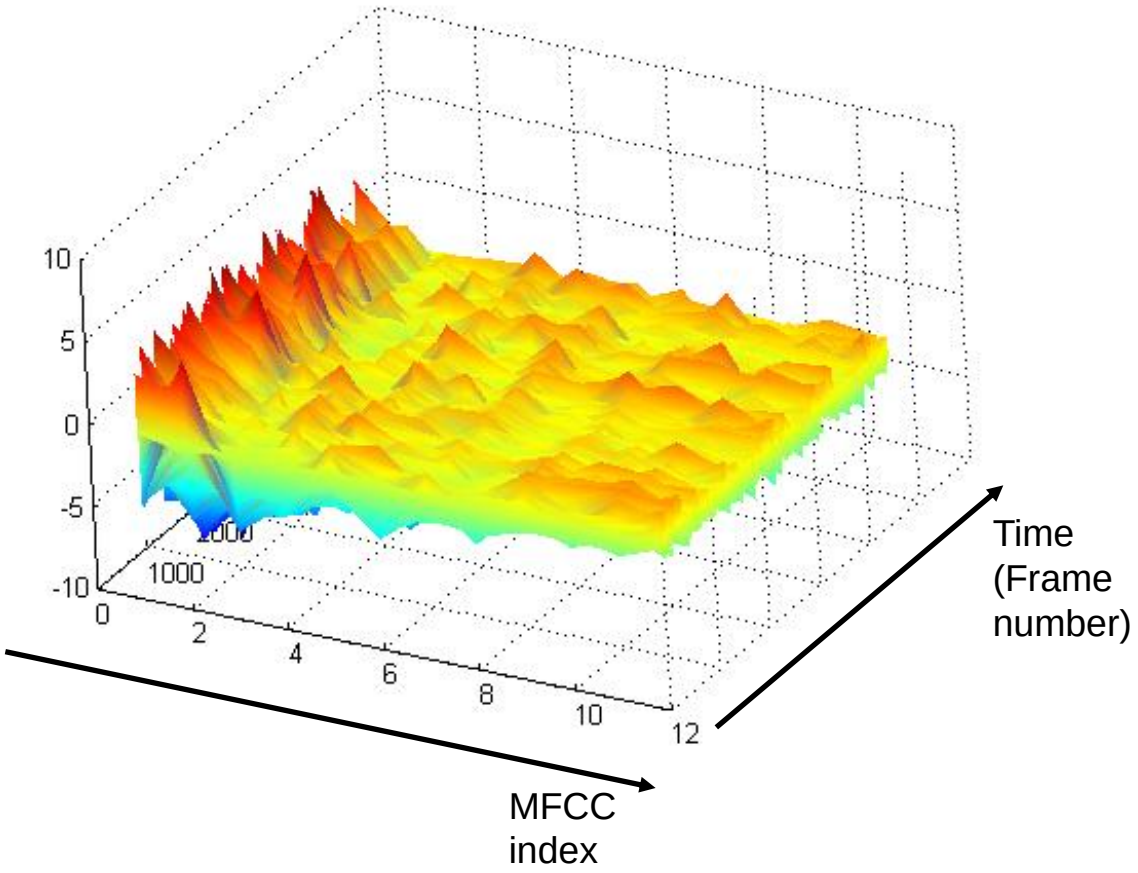


$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos\left(\frac{\pi i}{N}(j - 0.5)\right)$$

Typical MFCC parameters

- 25 or 30 msec frame, 50% frame overlap
- FFT size 512, 1024 or 2048
- P=24 to P=30 filters in the Mel bank
- Retain ≈ P/2 cepstral coefficients
- Delta and double delta coefficients and cepstral mean subtraction (CMS)

How do the MFCCs look like?



The energy of the cepstral coefficients is concentrated on the low coefficients (low ”quefrencies”)

MFCCs – the Swiss Army Knife of Speech and Audio Signal Processing

- **Speech classification**
  - Automatic speech recognition
  - Speaker identification
  - Language identification
  - Emotion recognition
- **Music information retrieval**
  - Musical instrument recognition
  - Music genre identification
  - Singer identification
- **Speech synthesis, coding, conversion**
  - Statistical parametric speech synthesis
  - Speaker conversion
  - Speech coding
- **Others**
  - Speech pathology classification
  - Identification of cell phone models
- etc …

Example MFCC implementations

- Voicebox (Matlab)
  http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
- Rastamat (Matlab)
  http://labrosa.ee.columbia.edu/matlab/rastamat/
- Bob toolkit
  http://idiap.github.io/bob/
- Hidden Markov model (HTK) toolkit
  http://htk.eng.cam.ac.uk/

Recommended literature

- X. Huang, A. Acero, H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall, 2001.
- R. M. Schafer, “Homomorphic systems and cepstrum analysis of speech”. In J. Benesty, M. Sandhi, Y. Huang (Eds), *Springer Handbook of speech processing*, pp. 161—180.