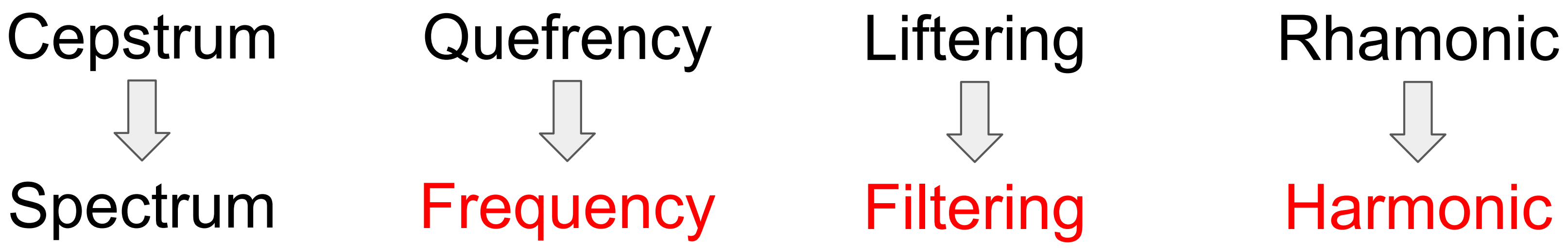
Mel-Frequency Cepstral Coefficients

Cepstrum

↓

Spectrum



Cepstrum → Spectrum

Quefrency → Frequency

Liftering → Filtering

Rhamonic → Harmonic

An historical note on Cepstrum

- Developed while studying echoes in seismic signals (1960s)
- Audio feature of choice for speech recognition / identification (1970s)
- Music processing (2000s)

Computing the cepstrum

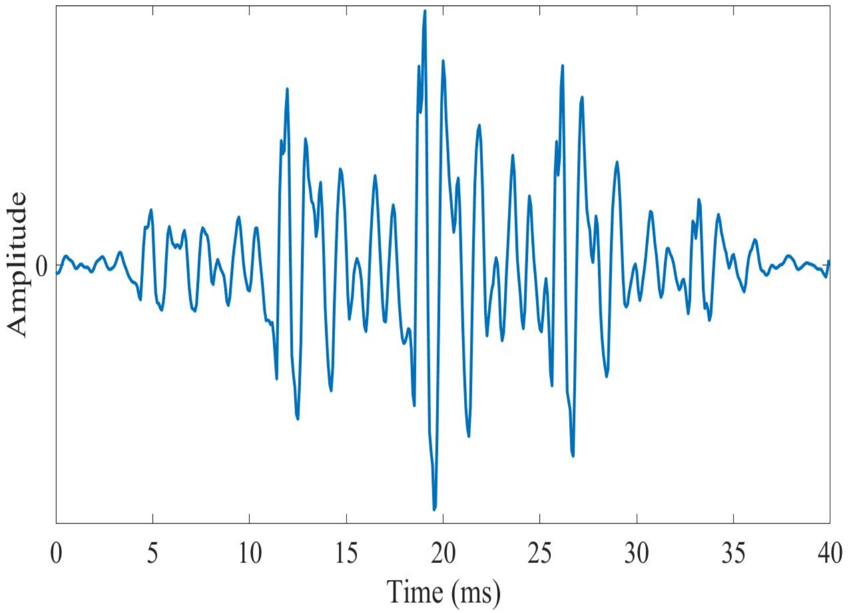
$$C(x(t)) = F^{-1}[log(F[x(t)])]$$

Computing the cepstrum

$$C(\overbrace{x(t)}^{\text{Time-domain signal}}) = \underbrace{F^{-1}[\underbrace{log(\overbrace{F[\boxed{x(t)}]}^{\text{Spectrum}})}_{\text{Log spectrum}}]}_{\text{Cepstrum}}$$

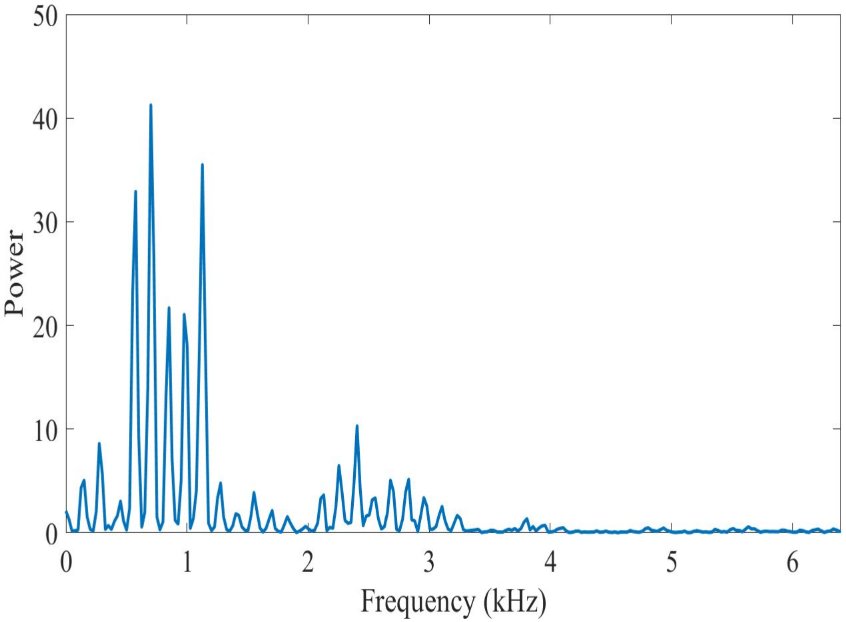
Visualising the cepstrum

Signal



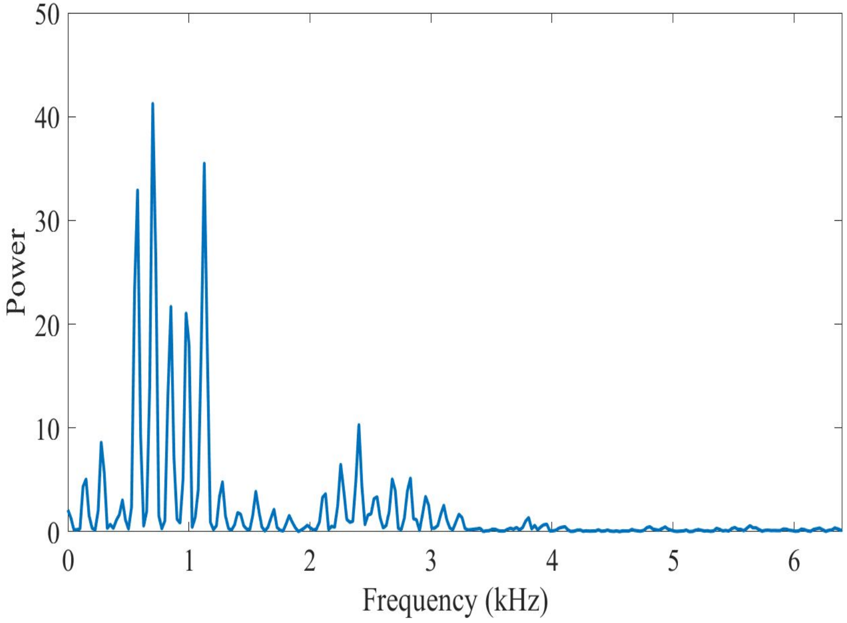
DFT

Power spectrum



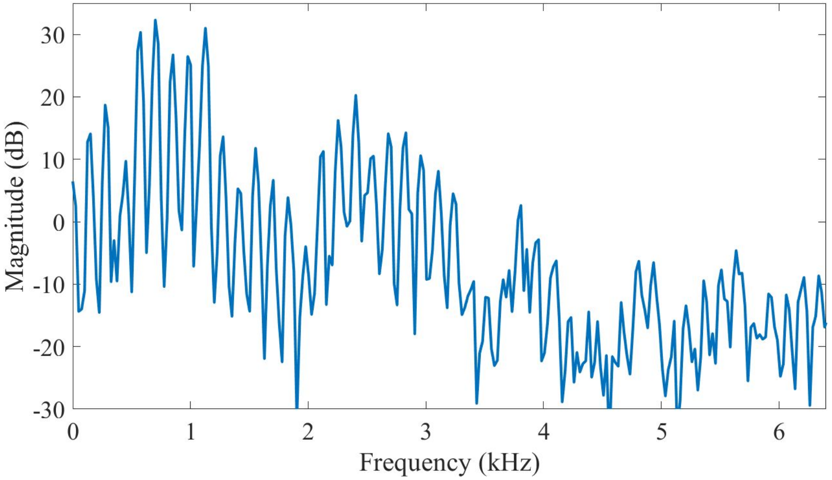
Visualising the cepstrum

Power spectrum



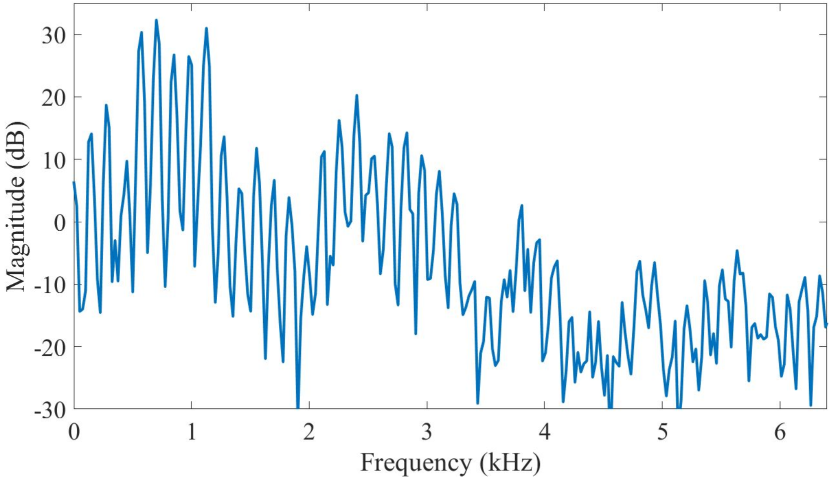
log

Log power spectrum



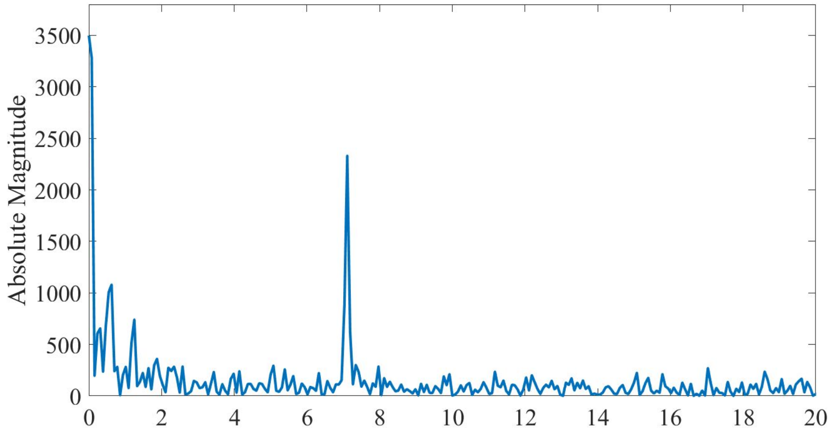
Visualising the cepstrum

Log power spectrum



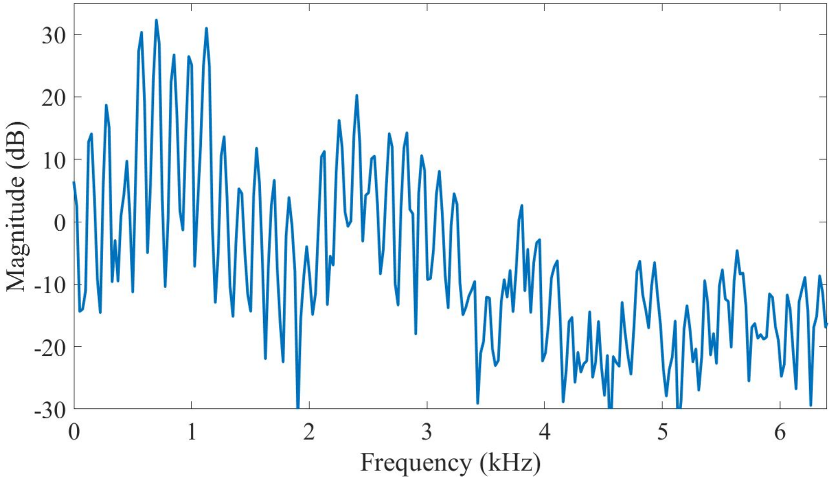
IDFT

Cepstrum

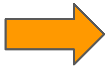


Visualising the cepstrum

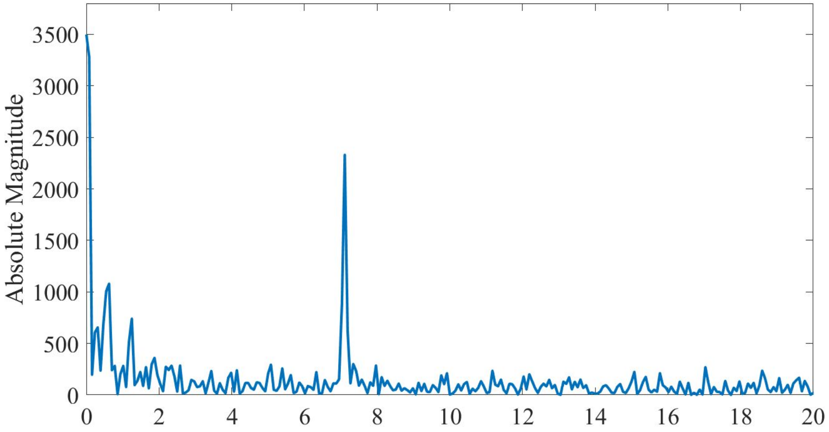
Log power spectrum



IDFT



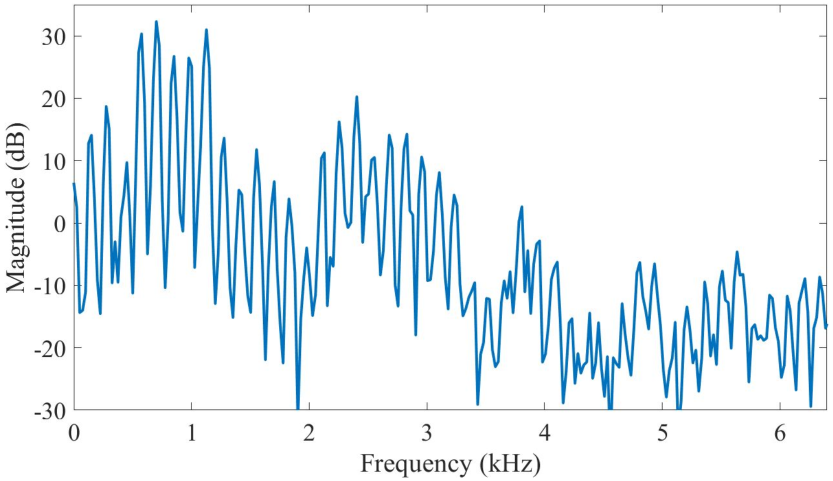
Cepstrum



????

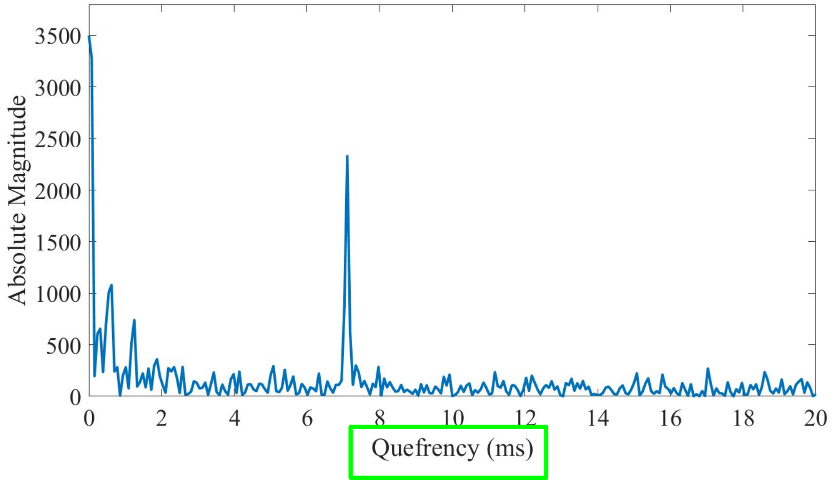
Visualising the cepstrum

Log power spectrum



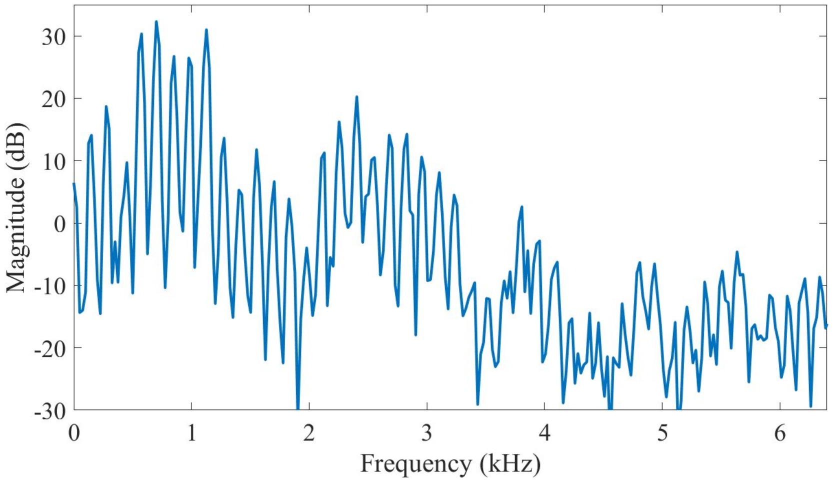
IDFT

Cepstrum



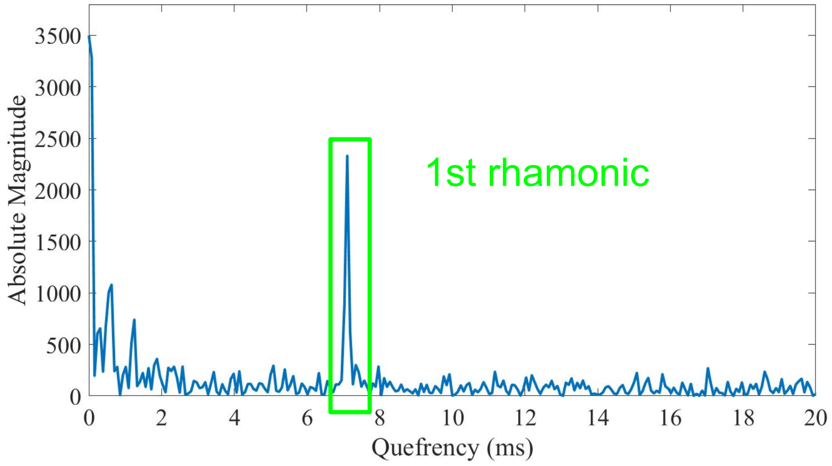
Visualising the cepstrum

Log power spectrum

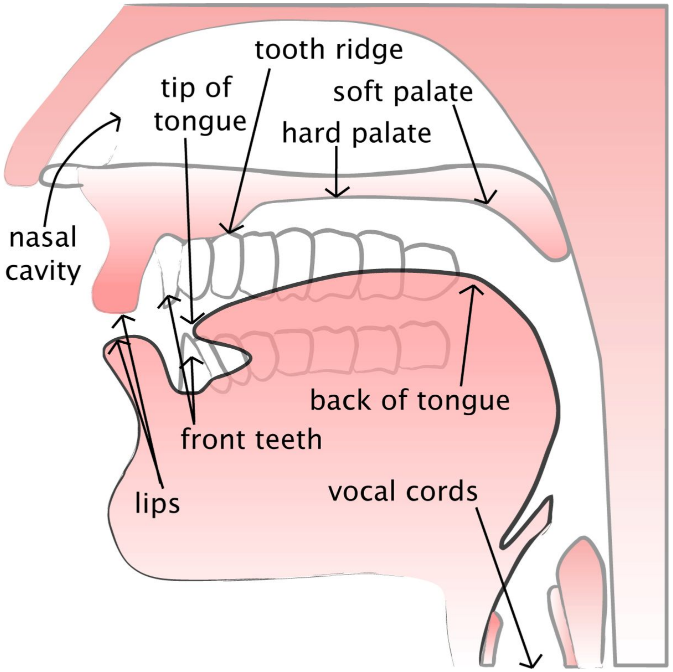


IDFT

Cepstrum

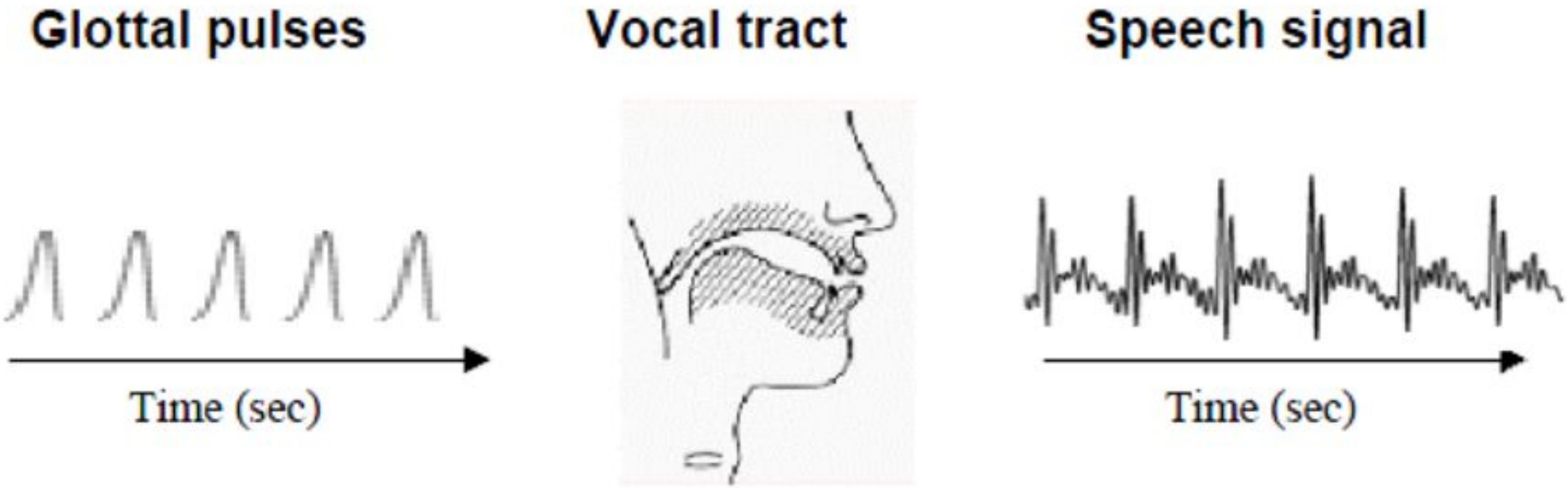


The vocal tract



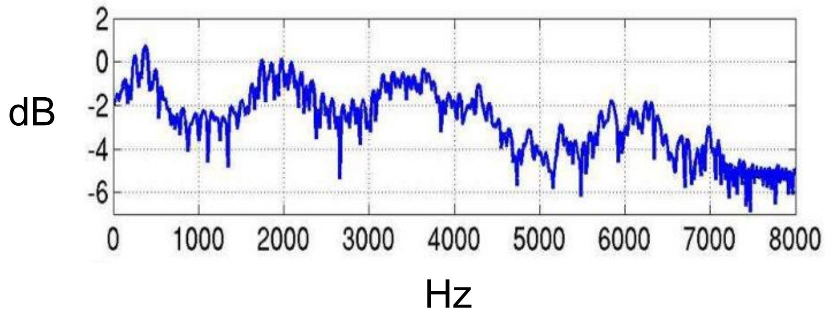
Vocal tract acts as a filter

Speech generation



Understanding the cepstrum

Log-spectrum

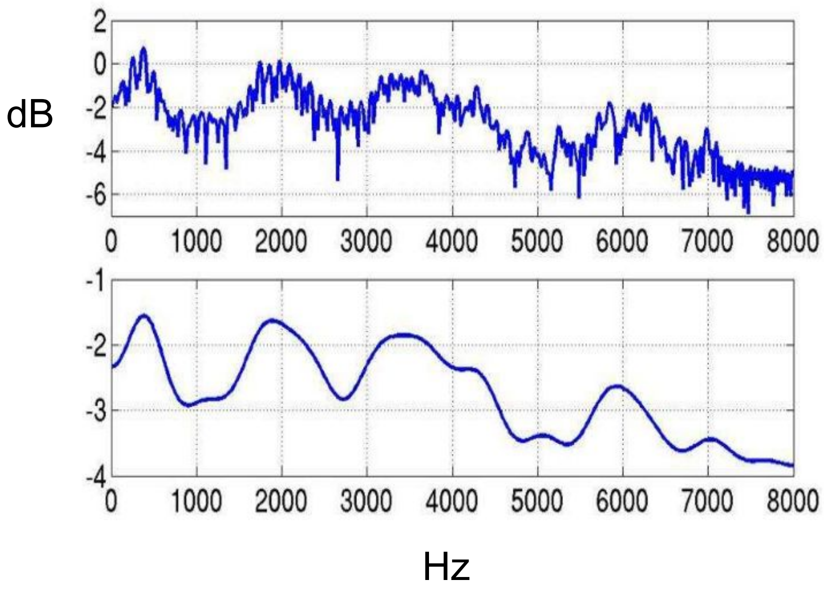


Speech

Understanding the cepstrum

Log-spectrum

Spectral envelope



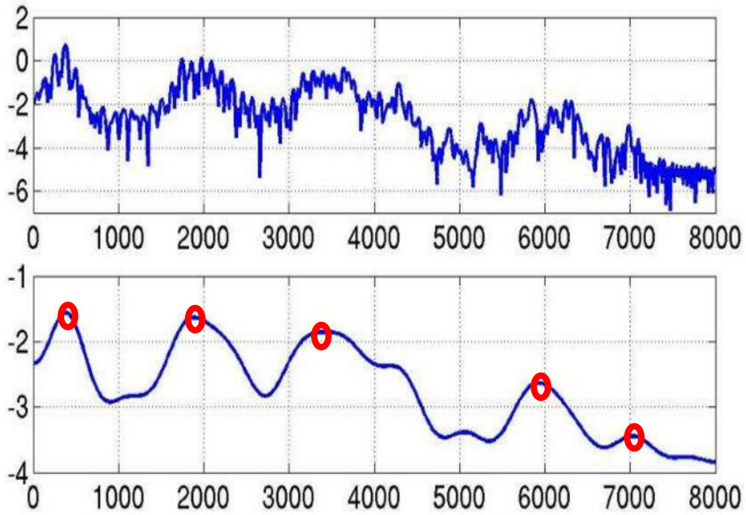
Speech

Understanding the cepstrum

Log-spectrum

dB

Speech



Spectral envelope

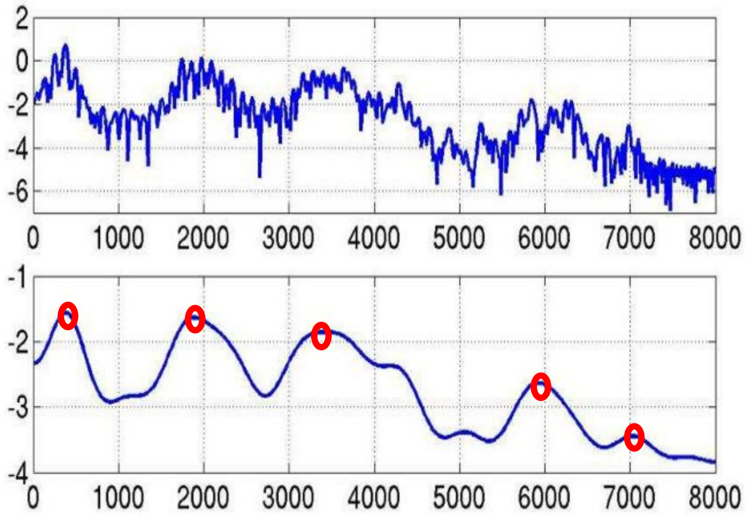
Hz

Understanding the cepstrum

Log-spectrum

Spectral envelope

dB



Hz

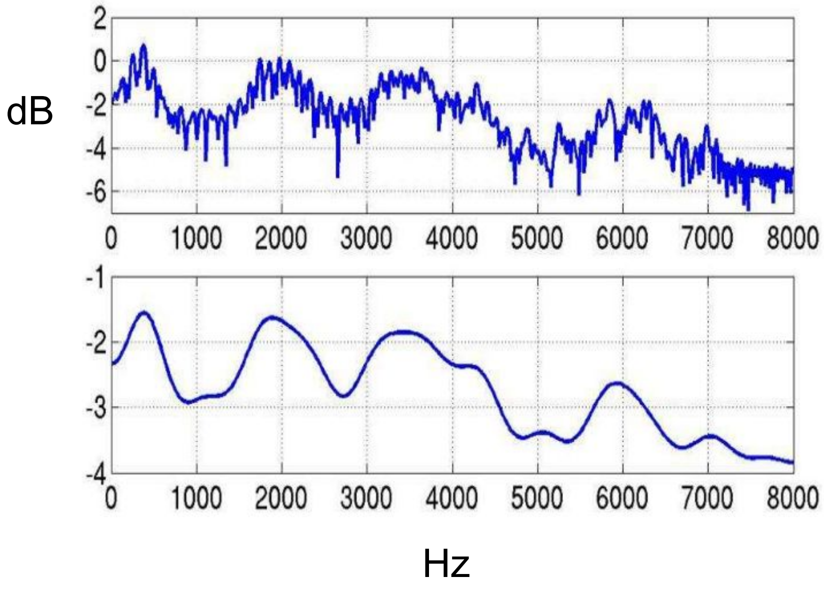
Speech

Formants = Carry identity of sound

Understanding the cepstrum

Log-spectrum

Spectral envelope



Speech

Vocal tract frequency response

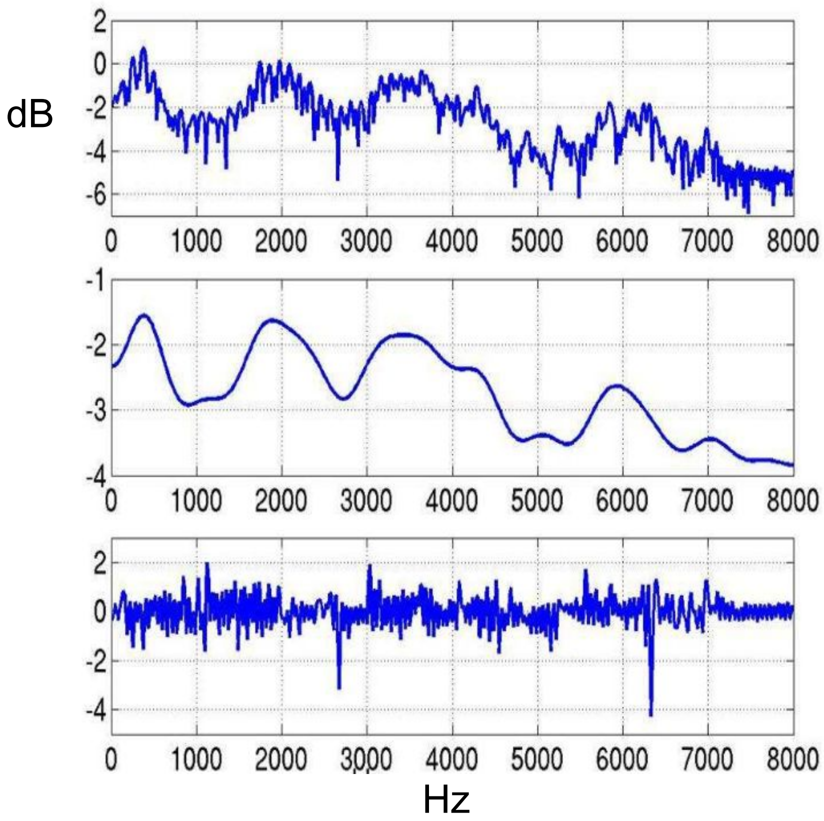
Understanding the cepstrum

Log-spectrum

Speech

Spectral envelope

Vocal tract frequency response



Understanding the cepstrum

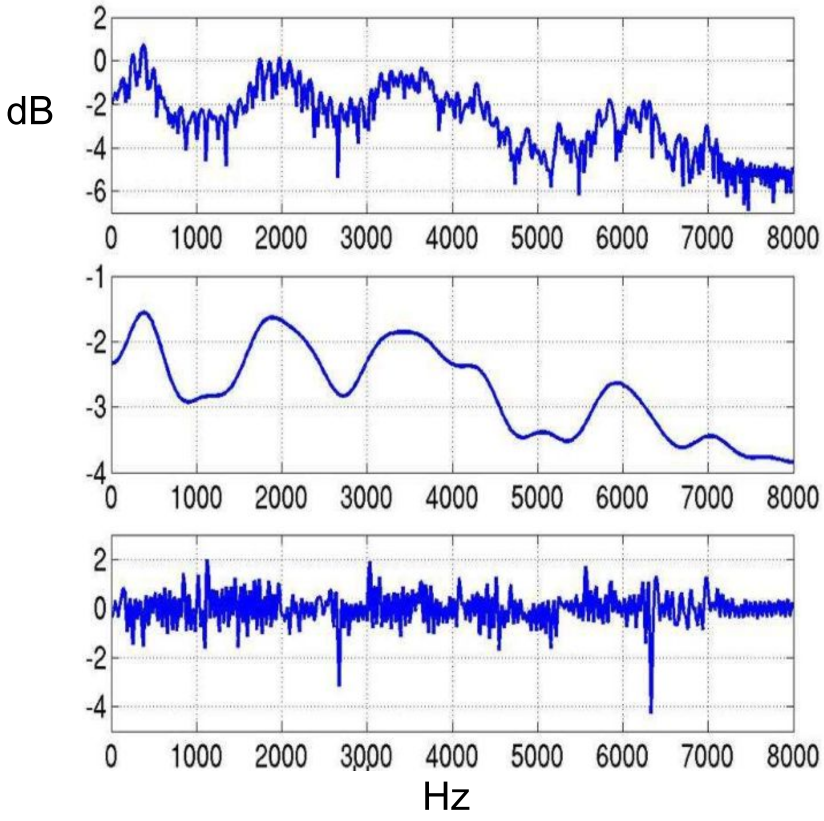
Log-spectrum

Speech

Spectral envelope

Vocal tract frequency response

Spectral detail

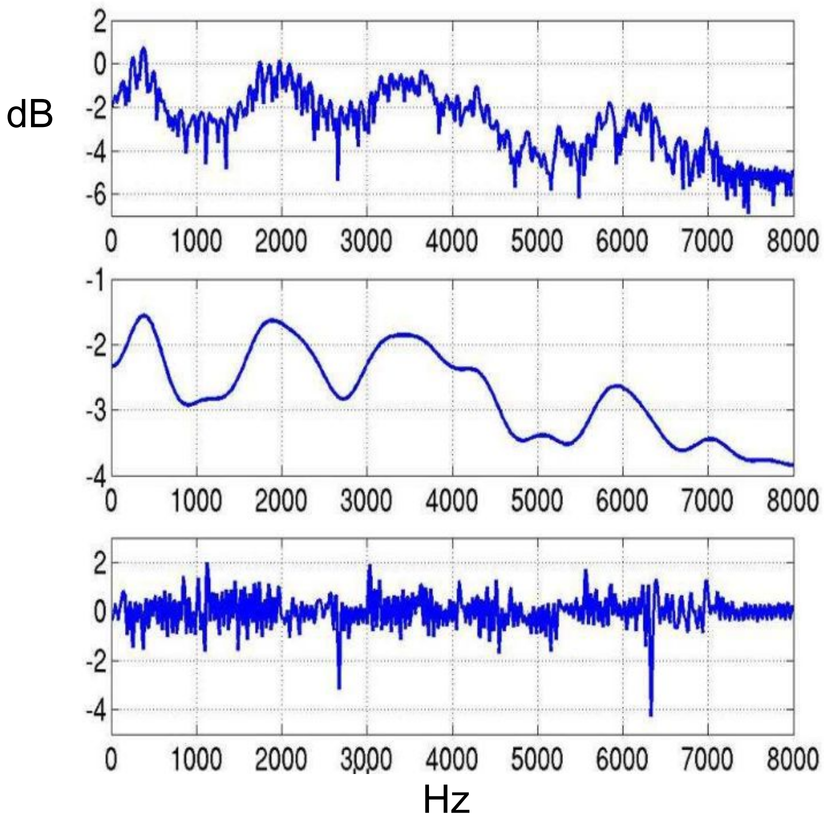


Understanding the cepstrum

Log-spectrum

Spectral envelope

Spectral detail



Speech

Vocal tract frequency response

Glottal pulse

Speech

=

Convolution of vocal tract frequency response with glottal pulse

Formalising speech

$$x(t) = e(t) \cdot h(t)$$

Formalising speech

$$x(t) = e(t) \cdot h(t)$$

$$X(t) = E(t) \cdot H(t)$$

Formalising speech

$$X(t) = E(t) \cdot H(t)$$

Formalising speech

$$X(t) = E(t) \cdot H(t)$$

$$\Downarrow$$

$$log(X(t)) = log(E(t) \cdot H(t))$$

Formalising speech

$$X(t) = E(t) \cdot H(t)$$

$$\Downarrow$$

$$log(X(t)) = log(E(t) \cdot H(t))$$

$$\Downarrow$$

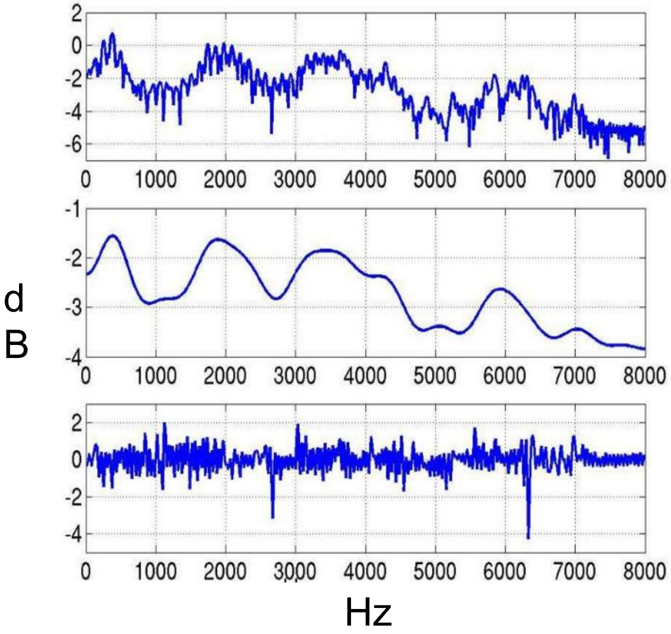
$$log(X(t)) = log(E(t)) + log(H(t))$$

Formalising speech

$$log(X(t)) = log(E(t)) + log(H(t))$$

Formalising speech

$$\boxed{log(X(t))} = \boxed{log(E(t))} + \boxed{log(H(t))}$$

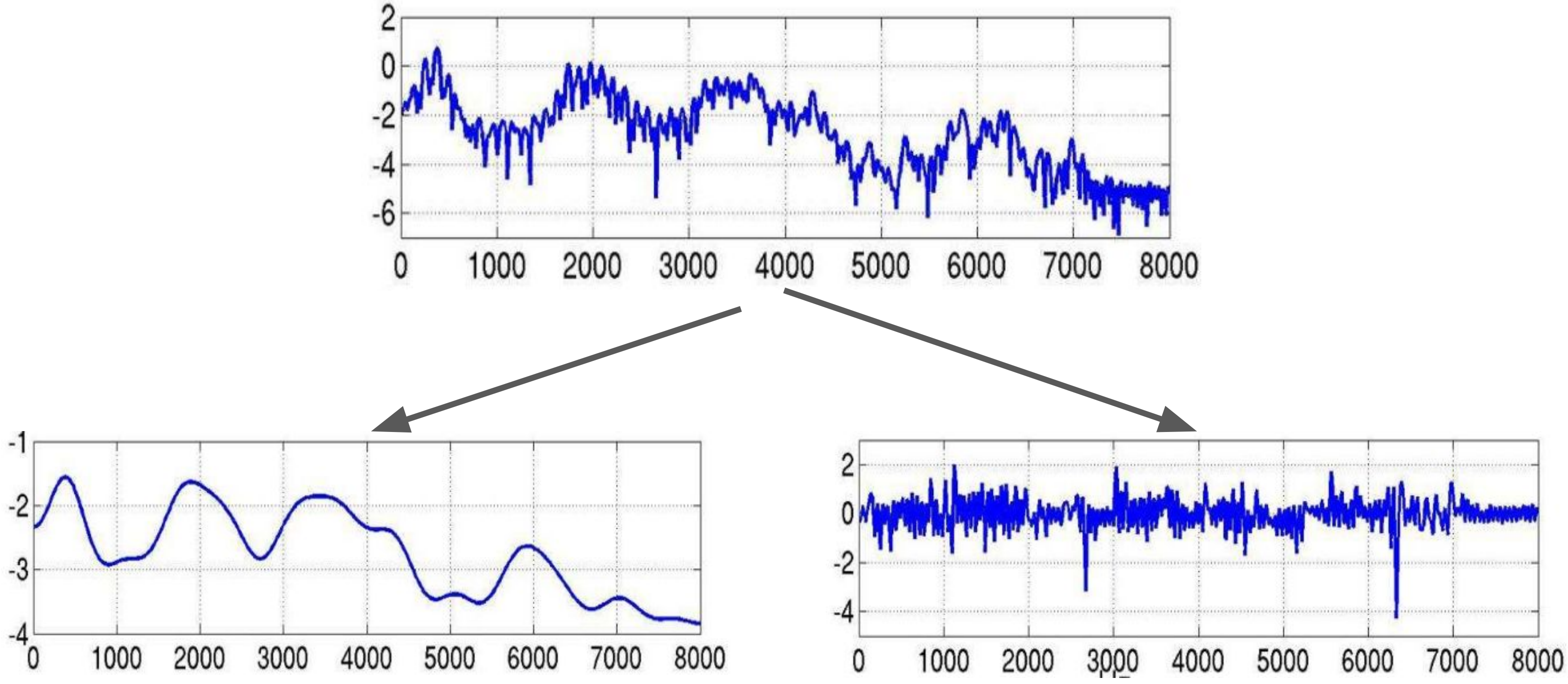


Speech

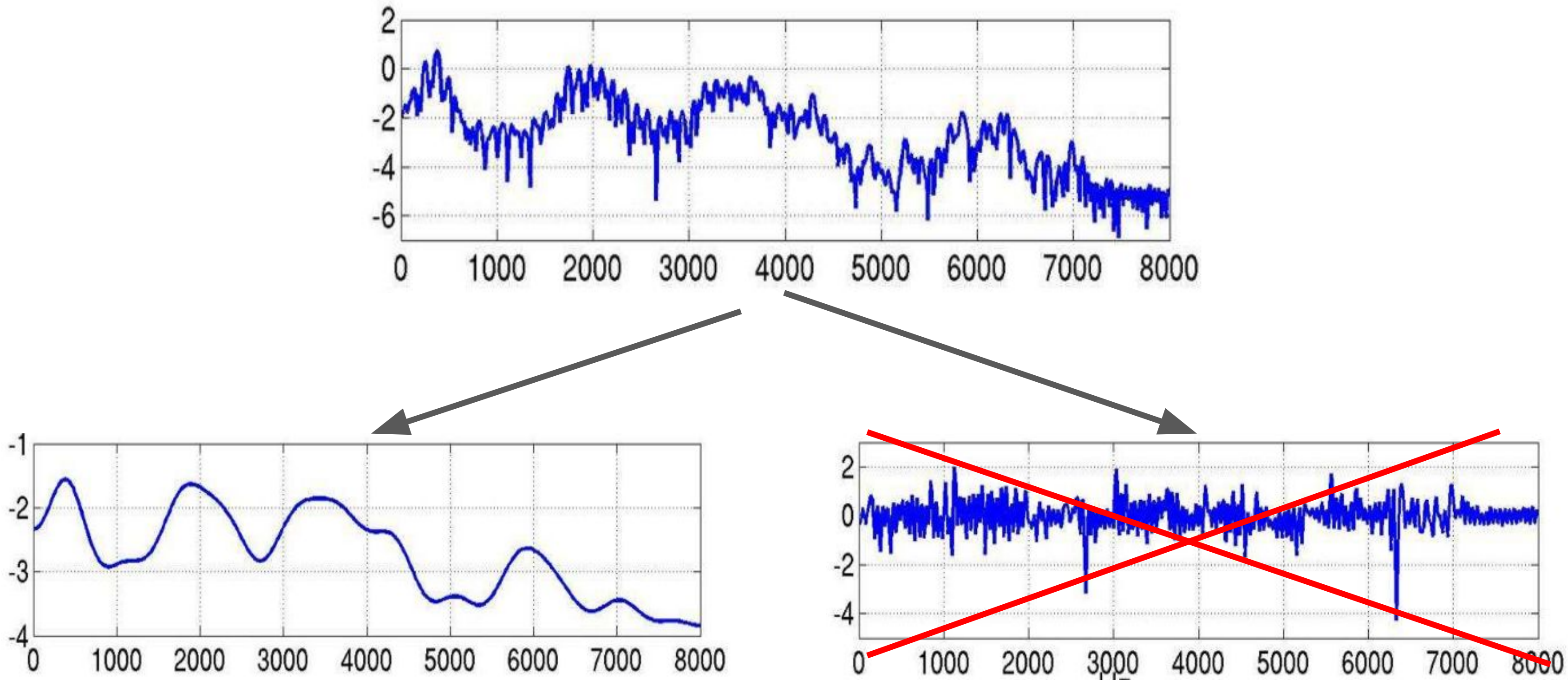
Vocal tract frequency response

Glottal pulse

The goal: Separating components

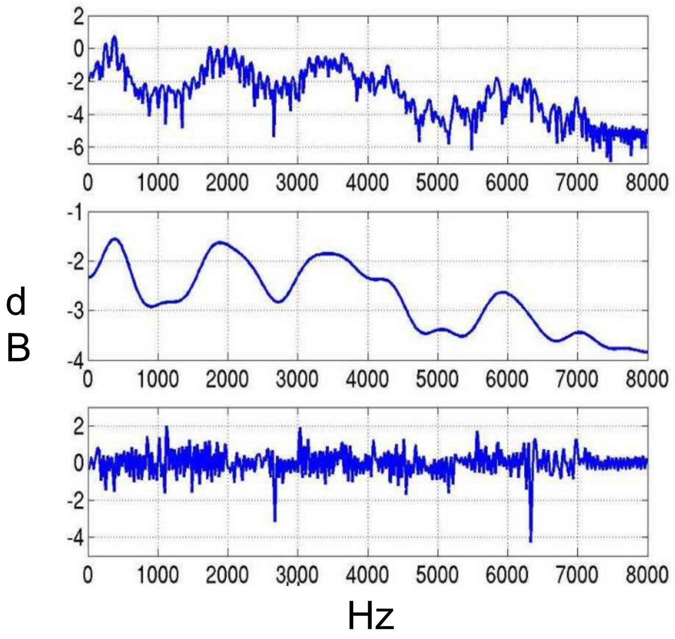


The goal: Separating components



Separating components

$$log(X(t)) = log(E(t)) + log(H(t))$$

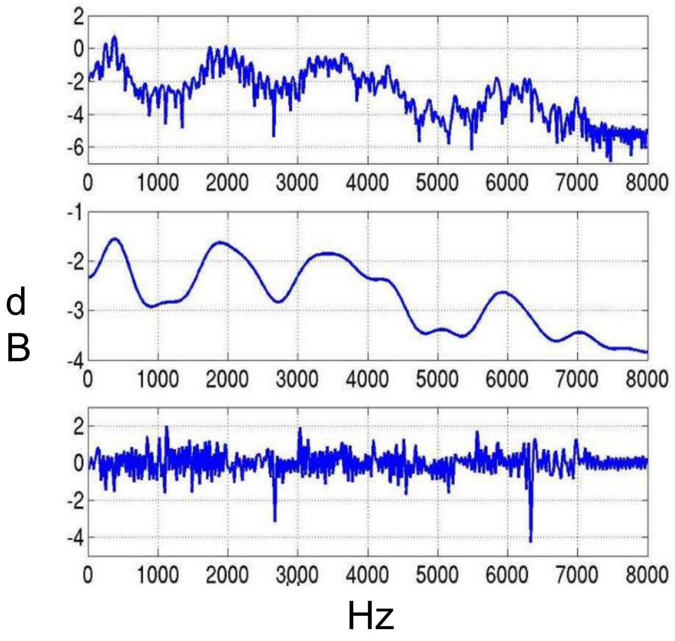


Separating components

$$log(X(t)) = log(E(t)) + log(H(t))$$

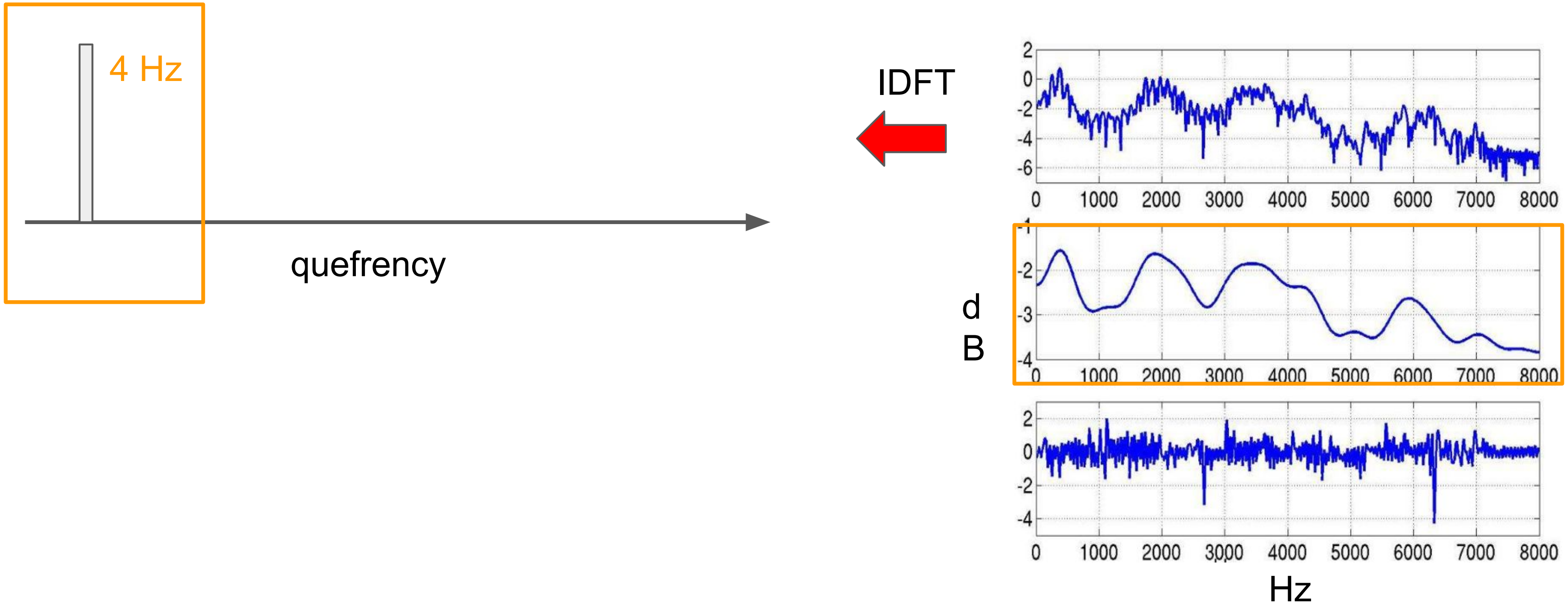
IDFT

quefrency



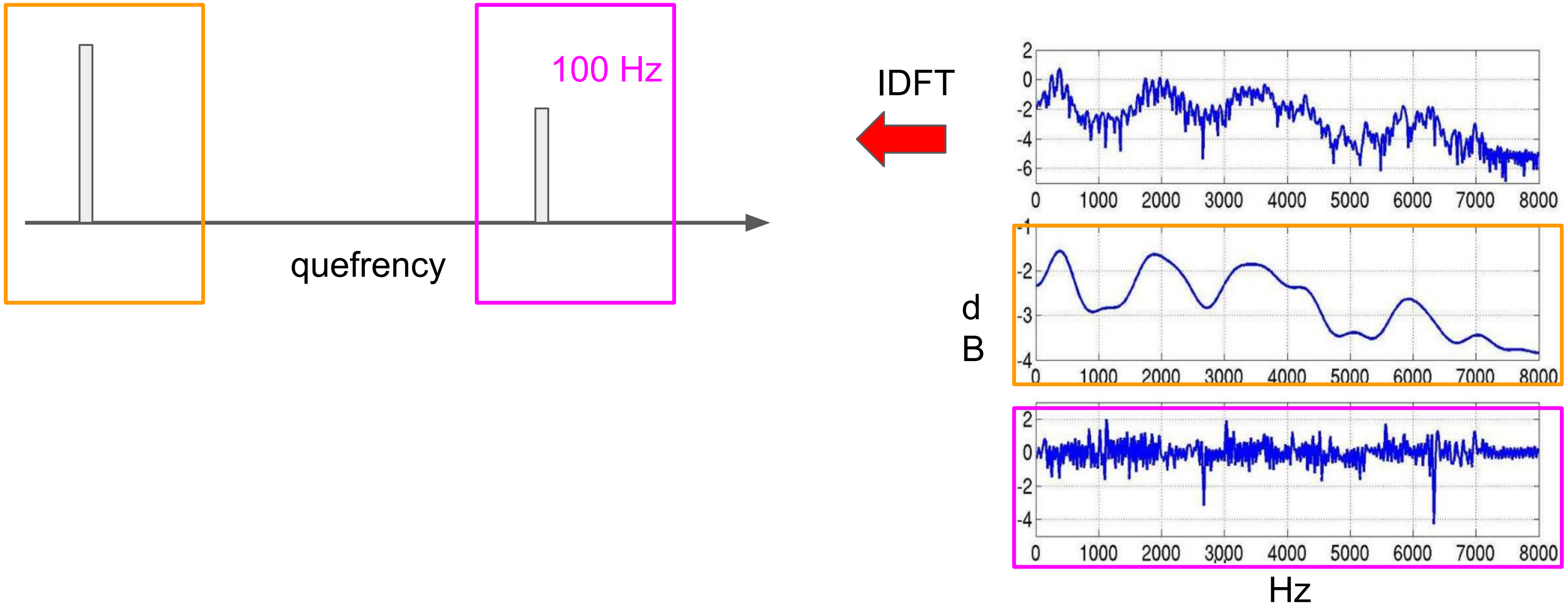
Separating components

$$log(X(t)) = log(E(t)) + log(H(t))$$



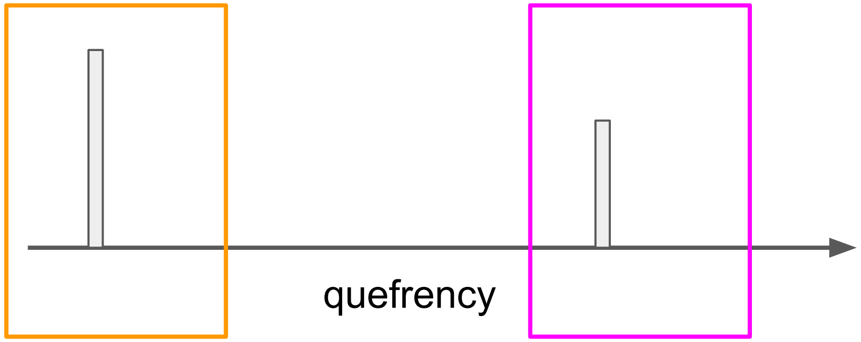
Separating components

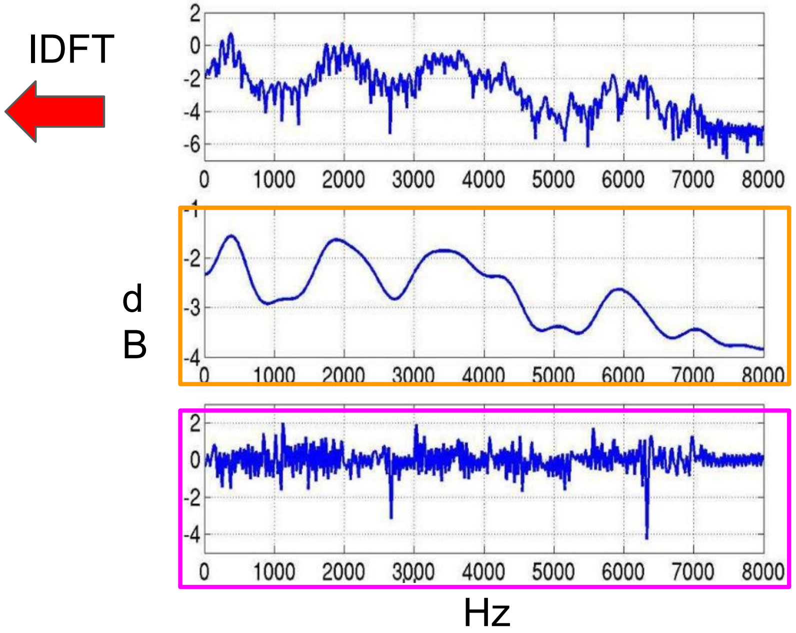
$$log(X(t)) = log(E(t)) + log(H(t))$$



Separating components

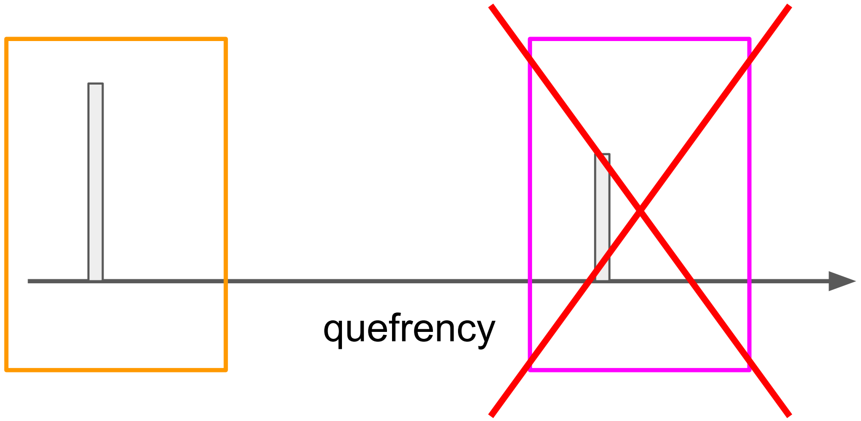
$$log(X(t)) = log(E(t)) + log(H(t))$$





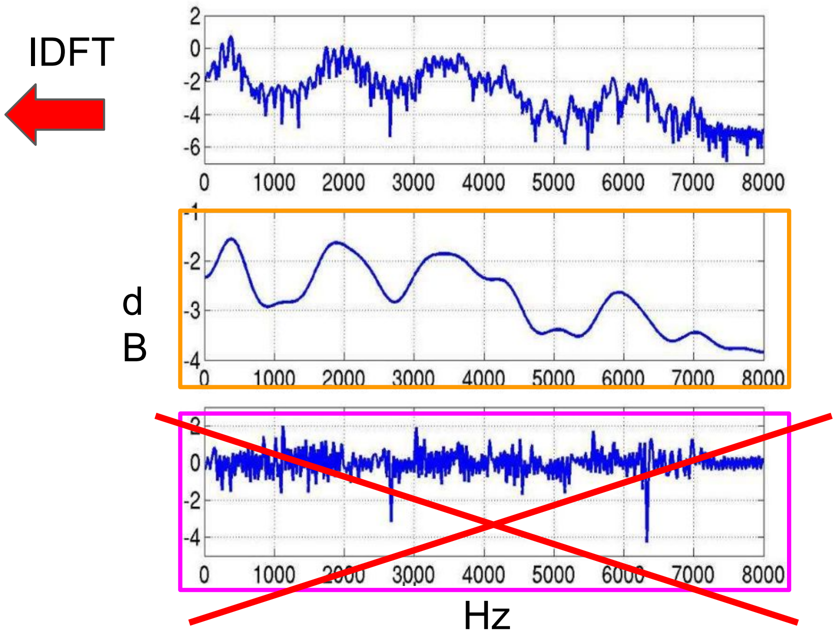
$$X(t) = \boxed{E(t)} + \boxed{H(t)}$$

Separating components

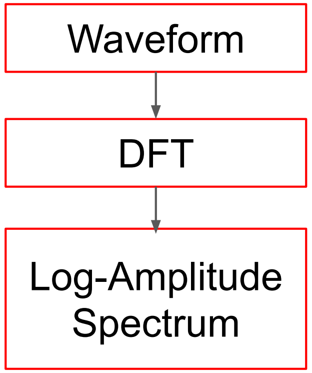


$$X(t) = \cancel{\boxed{E(t)}} + \boxed{H(t)}$$

$$log(X(t)) = log(E(t)) + log(H(t))$$



Computing Mel-Frequency Cepstral Coefficients

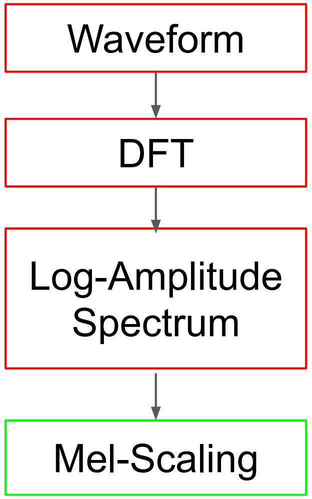


Waveform

DFT

Log-Amplitude Spectrum

Computing Mel-Frequency Cepstral Coefficients

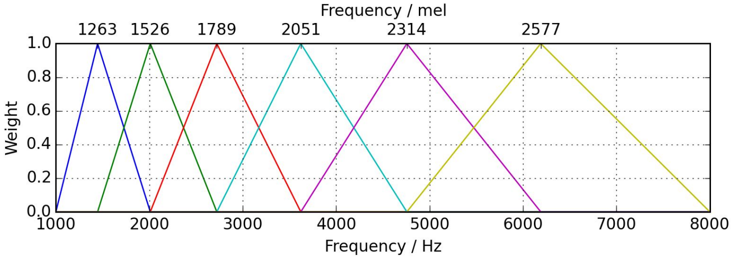


Waveform

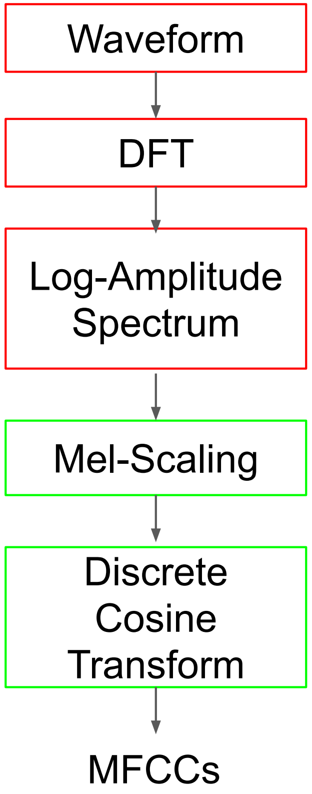
DFT

Log-Amplitude Spectrum

Mel-Scaling



Computing Mel-Frequency Cepstral Coefficients



Waveform

DFT

Log-Amplitude Spectrum

Mel-Scaling

Discrete Cosine Transform

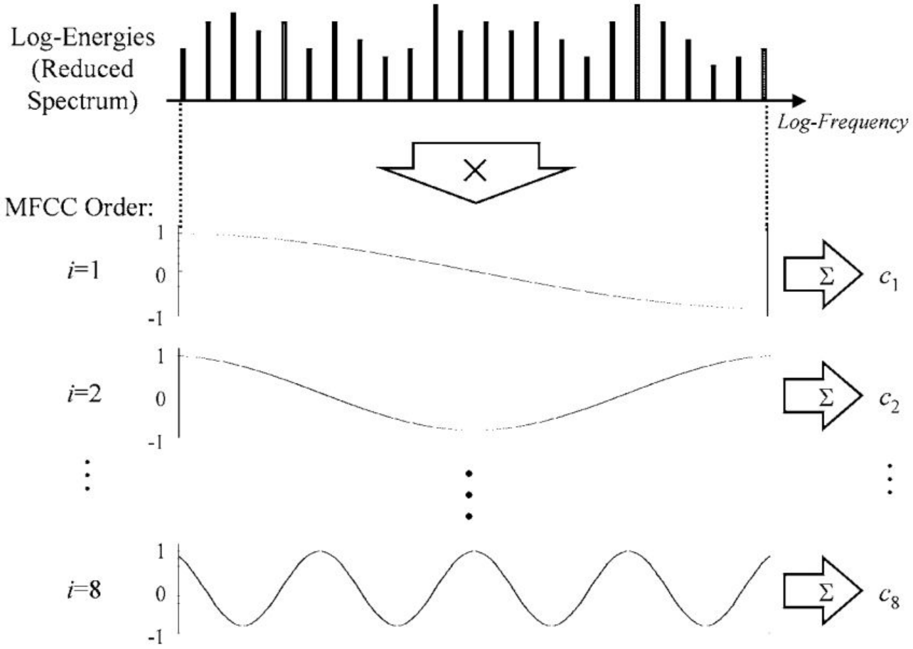
MFCCs

Why Discrete Cosine Transform?

- Simplified version of Fourier Transform
- Get real-valued coefficient

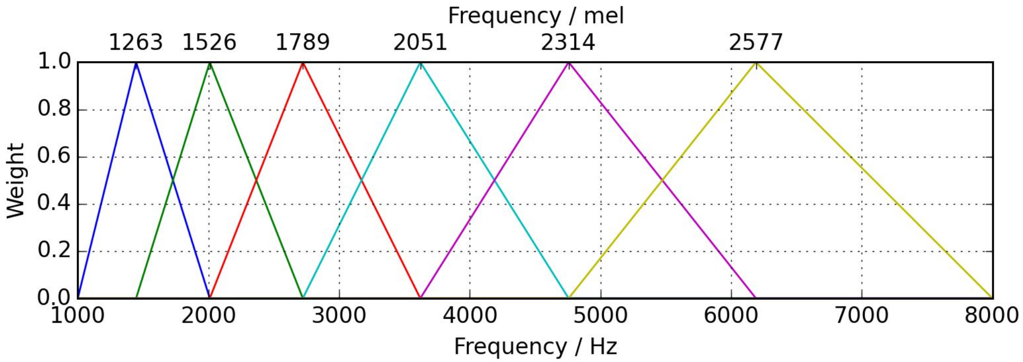
Why Discrete Cosine Transform?

- Simplified version of Fourier Transform
- Get real-valued coefficient



Why Discrete Cosine Transform?

- Simplified version of Fourier Transform
- Get real-valued coefficient
- Decorrelate energy in different mel bands



Why Discrete Cosine Transform?

- Simplified version of Fourier Transform
- Get real-valued coefficient
- Decorrelate energy in different mel bands
- Reduce # dimensions to represent spectrum

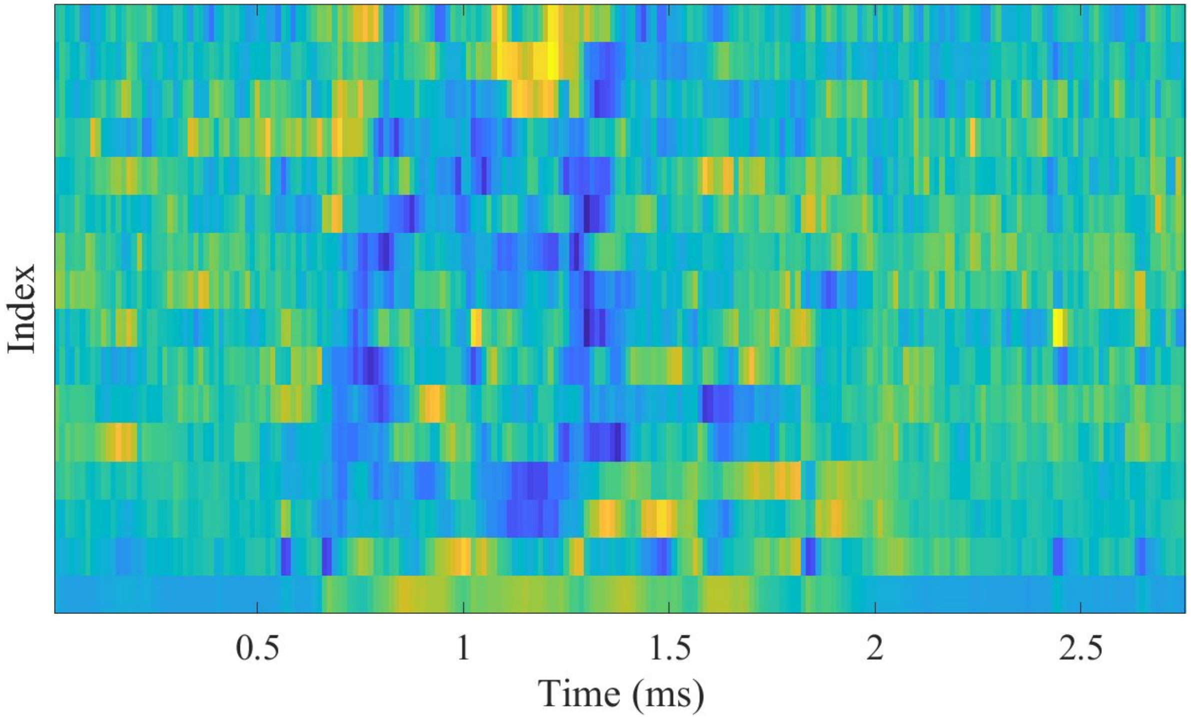
How many coefficients?

- Traditionally: first 12 - 13 coefficients
- First coefficients keep most information (e.g., formants, spectral envelope)
- Use Δ and ΔΔ MFCCs
- Total 39 coefficients per frame

Visualising MFCCs

# frames

# coefficients



MFCCs advantages

- Describe the “large” structures of the spectrum
- Ignore fine spectral structures
- Work well in speech and music processing

MFCCs disadvantages

- Not robust to noise
- Extensive knowledge engineering
- Not efficient for synthesis

MFCCs applications

- Speech processing
  - Speech recognition
  - Speaker recognition
  - ...
- Music processing
  - Music genre classification
  - Mood classification
  - Automatic tagging
  - ...

What’s up next?

- Extract MFCCs with Python and Librosa
- Visualise MFCCs