SGN-14006 Audio and Speech Processing

Cepstrum

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

$$\text{Log} \ |S(f)| = \text{log} \ |H(f)| + \text{log} \ |G(f)|$$
Spectral periodicity

- The spectrum of a periodic sound involves frequency components at the fundamental frequency $F_0$ and its whole-number multiples (harmonics) $n \cdot F_0$
- The $F_0$ + 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. It is coined as ’quefrecy’ – pseudo time.

• The $|G(f)|$ can now in principle be removed by ”liftering” the cepstrum of high-frequency components.
Definition of the cepstrum

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

where \( S(n) \) is the Fourier spectrum of a speech frame, \( n \) = discrete frequency index.

• In other words, the cepstrum is defined as 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

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

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

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

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

DCT II cosine transform is defined as:

\[ c_x[k] = \frac{1}{N} \sum_{n=0}^{N-1} x(n) \cos\left(\frac{\pi k}{N} (n + \frac{1}{2})\right) \]

\[ c_{\log|S(n)|}[k] = \frac{1}{N} \sum_{n=0}^{N-1} \log|S(n)| \cos\left(\frac{\pi k}{N} (n + \frac{1}{2})\right) \]
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.
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 quefrecencies) represent the glottal excitation.**
Spectral envelope extraction

- By removing high quefrency components (liftering), and taking a DFT results in spectral envelope of the signal. (by removing low quefrencies and taking DFT gets the spectrum of glottal excitation)
Example

- Figure on the following page shows how 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.
Separating glottal excitation (upper cepstral coeffs.) and vocal tract response (lower coeffs.)

![Image of spectrograms showing frequency and time relationships for speech signals.](image)

- **Spectrogram**: Displays frequency and time information, showing the distribution of energy across frequencies over time.
- **Cepstrogram**: Shows the cepstral coefficients, which are derived from the Fourier transform of the log of the short-time Fourier transform of the signal, highlighting the vocal tract response.
- **Pitch Spectrogram**: Focuses on the pitch information within the speech signal, with the window type set to 'hamming', window size of 256, and window shift of 64.