Cepstrum

Notes on Speech and Audio Processing

Chia-Ping Chen

Department of Computer Science and Engineering
National Sun Yat-Sen University
Kaohsiung, Taiwan ROC

Introduction

- A basic model for speech production
  - An excitation (source) drives a system of resonators (filters).
  - The output speech signal is the convolution of the excitation and the system's impulse response.
- Excitation = vibration or air flow at vocal fold
- Resonating system = vocal tract configuration

Deconvolution

- For speech recognition, the resonators part contains more information than the excitation part.
- To analyze speech, it is natural to try a separation of the excitation (source) and the resonators (filters).
- This is called deconvolution, the inverse of convolution.
- Cepstral analysis performs deconvolution.

The Log Spectrum

- Let X be the spectrum of speech, E be excitation component and V be the vocal tract's frequency response. From convolution theorem,

$$|X(\omega)| = |E(\omega)|\,|V(\omega)|$$
$$\Rightarrow \log|X(\omega)| = \log|E(\omega)| + \log|V(\omega)|.$$

- $E(\omega)$ tends to vary rapidly with ω, while $V(\omega)$ tends to vary slowly with ω.
- Therefore, E and V can be separated by separating the fast-changing and slow-changing components. See Figure 20.1 for an example.

The Real Cepstrum

- The cepstrum is the inverse Fourier transform of $\log|X(\omega)|$,

$$c(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log|X(\omega)|e^{j\omega n}d\omega.$$

- $X(\omega)$ often comes from DFT. In this case, we apply IDFT to $\log|X(\omega)|$ to compute the cepstrum.
- Lower-order $c(n)$ (small n) correspond to vocal tract filters, while higher-order ones correspond to excitations.

The Complex Cepstrum

- The complex cepstrum is the inverse Fourier transform of $\log X(\omega)$, (instead of $\log|X(\omega)|$)

$$\hat{x}[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log X(\omega) e^{j\omega n} d\omega.$$

- $\hat{x}[n]$ provides the phase information in addition to the magnitude information.

Cepstrum of a Rational System

- Suppose the impulse response is a rational fraction

$$X(z) = \frac{A \prod_{k=1}^{M_i}(1 - a_k z^{-1}) \prod_{k=1}^{M_o}(1 - b_k z)}{\prod_{k=1}^{N}(1 - c_k z^{-1})}$$

a_k, b_k^{-1}'s are zeros, while c_k's are poles.

- $\hat{x}[n]$ is the sequence whose z-transform is $\log X(z)$,

$$\hat{x}[n] = \begin{cases} \log A, & n = 0 \\ \sum_{k=1}^{N} \frac{c_k^n}{n} - \sum_{k=1}^{M_i} \frac{a_k^n}{n}, & n > 0 \\ \sum_{k=1}^{M_o} \frac{b_k^{-n}}{n}, & n < 0 \end{cases}$$

Cepstrum of a Rational System

Note if $x[n] \overset{Z}{\Longrightarrow} X(z)$ then $-nx[n] \overset{Z}{\Longrightarrow} z\frac{dX(z)}{dz}$.

Let $X(z) = \log(1 - a_k z^{-1})$, then

$$-nx[n] \overset{Z}{\Longrightarrow} z\frac{a_k z^{-2}}{1 - a_k z^{-1}} = \frac{a_k z^{-1}}{1 - a_k z^{-1}} = \sum_{n=1}^{\infty} a_k^n z^{-n}$$

So $x[n] = -\frac{a_k^n}{n}$.

$\hat{x}[n]$ in the previous slide is obtained by collecting the contributions from all terms.

Cepstrum Liftering

- Figure 20.2 shows the creation and difference of cepstrum and complex cepstrum.
- Figure 20.3 shows the complex cepstrum of a voiced speech segment.
- Figure 20.4 shows how cepstrum liftering (also called homomorphic filtering) performs deconvolution of Figure 20.3 (a).