Speech Representation

Notes on Spoken Language Processing

Chia-Ping Chen

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

Introduction

- We'll describe several signal representation methods for speech coding, synthesis and recognition.
- They're based on a source-filter model of speech production
  - time-varying signals and filters
- They're inspired by speech production and/or perception.

Short-Time Analysis

- Given a speech signal $x[n]$, the short-time signal is

$$x_m[n] = x[n]w_m[n],$$

where $w_m[n]$ is a *window function* focused on a small region in time (short-time).

- For example,

$$w_m[n] = \text{rect}_N[n - m],$$

is non-zero only for $m \leq n < m + N$.

Spectrum

- Each window of signal is called a frame.
- We extend the samples of $x[n]$ in a frame to a periodic signal with period M. The spectrum is

$$\tilde{X}_m(e^{j\omega}) = \frac{2\pi}{M} \sum_k X_m[k] \delta(\omega - 2\pi k/M).$$

- If a window function $w_m[n] = w[m-n]$ is multiplied, the spectrum is the convolution of $\tilde{X}_m(e^{j\omega})$ and $W(e^{-j\omega})e^{-j\omega m}$,

$$X_m(e^{j\omega}) = \sum_k X_m[k] W(e^{-j(\omega - 2\pi k/N)}) e^{-j(\omega - 2\pi k/N)m}.$$

Spectrogram

- Spectrogram represents the log energy of short-time Fourier transform.
- The phase of FT is not represented.
- It is a 2-D representation, time and frequency.
- Wide-band spectrogram uses short windows (≤ 10 ms), with fine time resolution and coarse frequency resolution.
- Narrow-band spectrogram uses longer windows (> 20 ms).

Pitch-Synchronous Analysis

- We apply periodic extention to a window of signal. It makes good sense if the window contains exactly one pitch period.
- A fixed-size window may contain several, non-integral, or less than one pitch periods.
- A pitch-synchronous analysis matches window size and pitch period dynamically.
- Challenge: local pitch period estimation is not an easy task.

Models for Speech Production

- volume velocity at glottis: $u_G[n]$
- volume velocity at lips: $u_L[n]$
- vocal tract transfer function

$$V(z) = \frac{U_L(z)}{U_G(z)}.$$

- pressure at lips

$$P_L(z) = U_L(z)\, Z_L(z)$$

- overall

$$P_L(z) = U_G(z) V(z) Z_L(z)$$

Source-Filter Models

- For voiced sound, we model $u_G[n]$ as a convolution of $g[n]$, the glottal pulse, and an impulse train $e[n]$.
- For unvoiced sound, we model $u_G[n]$ as a random noise.
- We can lump $G(z), V(z)$ and $Z_L(z)$ as $H(z)$ for voiced speech. Then

$$X(z) = E(z)H(z).$$

  This is a source-filter model.
- Similarly we can lump $V(z)$ and $Z_L(z)$ as $H(z)$ for unvoiced speech.

Linear Predictive Coding

- An all-pole filter with sufficient poles is a good approximation for speech signals.

$$\frac{X(z)}{E(z)} = H(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}}.$$

- This is equivalent to linear prediction in the time domain

$$x[n] = \sum_{k=1}^{p} a_k x[n-k] + e[n].$$

- The first term on the rhs is a linear prediction.
- The second term is the prediction error.

Orthogonality Principle

- Let $x_m[n]$ be a frame of speech around sample m. Specifically,

$$x_m[n] = x[m+n]$$

- Define the short-term prediction (squared) error

$$E_m = \sum_n e_m^2[n] = \sum_n (x_m[n] - \sum_j a_j x_m[n-j])^2.$$

- Setting the derivative with respect to $a_j's$ to 0,

$$< \mathbf{e_m}, \mathbf{x_m^i} >= \sum_n e_m[n] x_m[n-i] = 0, \;\; 1 \leq i \leq p.$$

- The error vector is orthogonal to past vectors.

Yule-Walker Equations

- The orthogonality principle is a set of p linear equations, called the Yule-Walker equations.
- Specifically, they are

$$\sum_{j=1}^{p} a_j \phi_m[i,j] = \phi_m[i,0],$$

where

$$\phi_m[i,j] = \sum_n x_m[n-i] x_m[n-j].$$

- They can be solved by covariance, autocorrelation (Levinson-Durbin method), or lattice formulation.

Cepstral Processing

- A homomorphic transformation $\hat{x}[n] = D(x[n])$ converts a convolution into a sum

$$x[n] = e[n] * h[n] \;\Rightarrow\; \hat{x}[n] = \hat{e}[n] + \hat{h}[n].$$

- We want to find a value n_0 such that

$$\hat{e}[n] \sim 0 \text{ for } n < n_0, \text{ and } \hat{h}[n] \sim 0 \text{ for } n > n_0.$$

- In a sense, the source and filter are separated. This is also known as deconvolution.
- Cepstrum is one such homomorphic transformation.

The Real and Complex Cepstrum

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

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

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

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

- Note

$$\log X(e^{j\omega}) = \log|X(\omega)| + j\arg[X(e^{j\omega})].$$

A Rational System

- Suppose the transfer function is rational

$$H(z) = \frac{Az^r \prod_{k=1}^{M_i}(1 - a_k z^{-1}) \prod_{k=1}^{M_o}(1 - u_k z)}{\prod_{k=1}^{N_i}(1 - b_k z^{-1}) \prod_{k=1}^{N_o}(1 - v_k z)}$$

  with the magnitudes of a_k, b_k, u_k, v_k's are less than 1.

- a_k's are zeros inside the unit circle, u_k^{-1}'s are zeros outside; similarly, b_k's are poles inside, v_k^{-1}'s are poles outside the unit circle.
- z^r is time shift.

Cepstrum of a Rational System

- Taking the logarithm, we have

$$\hat{H}(z) = \log A + \log z^r + \sum_{k=1}^{M_i} \log(1 - a_k z^{-1}) + \sum_{k=1}^{M_o} \log(1 - u_k z) - \sum_{k=1}^{N_i} \log(1 - b_k z^{-1}) - \sum_{k=1}^{N_o} \log(1 - v_k z)$$

- $\hat{h}(n)$ is the inverse z-transform of $\hat{H}(z)$

$$\hat{h}[n] = \begin{cases} \log A, & n = 0 \\ \sum_{k=1}^{N_i} \frac{b_k^n}{n} - \sum_{k=1}^{M_i} \frac{a_k^n}{n}, & n > 0 \\ \sum_{k=1}^{M_o} \frac{u_k^n}{n} - \sum_{k=1}^{N_o} \frac{v_k^n}{n}, & n < 0 \end{cases}$$

Cepstrum of an All-Pole System

- Recall that the transfer function is

$$H(z) = \frac{G}{1 - \sum_{j=1}^{p} a_j z^{-j}}$$

- Taking logarithm, we have

$$\hat{H}(z) = \log G - \log(1 - \sum_{j=1}^{p} a_j z^{-j}) = \sum_n \hat{h}[n] z^{-n}.$$

- Taking the derivative with respect to z, we have

$$\frac{-\sum_{j=1}^{p} j a_j z^{-j-1}}{1 - \sum_{j=1}^{p} a_j z^{-j}} = -\sum_n n \hat{h}[n] z^{-n-1}.$$

LPC Cepstrum

- With LPC coefficients a_j, the complex cepstrum can be obtained recursively as follows.

$$\hat{h}[n] = \begin{cases} 0, & n < 0 \\ \log G, & n = 0 \\ a_n + \sum_{k=1}^{n-1} \frac{k}{n} \hat{h}[k] a_{n-k}, & 0 < n \leq p \\ \sum_{k=n-p}^{n-1} \frac{k}{n} \hat{h}[k] a_{n-k}, & n > p \end{cases}$$

Cepstrum of Speech Signal

- DFT of a window with length N

$$X_a[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi nk/N},\ 0 \leq k < N$$

- followed by logarithm

$$\hat{X}_a[k] = \log X_a[k]$$

- followed by IDFT

$$\hat{x}_a[n] = \frac{1}{N} \sum_{k=0}^{N-1} \hat{X}_a[k] e^{j2\pi nk/N}$$

MFCC

- The mel-frequency cepstral coefficients is a representation defined as the real cepstrum of a windowed short-time signal.
- We define a filterbank of M overlapping triangular filters centering around frequencies based on the mel-scale.
- Each filter computes the energy in the signal passing it.
- The DCT of the logarithm of filter outputs is the MFCC.

Formant Frequencies

- Formant frequencies are the resonances of the vocal tract.
- Trained professional can identify phones by the formant positions in a spectrogram.

Pitch Determination

- Pitch determination is important in synthesis, tone recognition (e.g. Chinese), and others.
- The determination of pitch periods uses short-time analysis. We look for the optimal candidate in frame m that

$$T_m = \arg\max_T f(T|\mathbf{x_m}).$$

- The autocorrelation method finds the peak of autocorrelation function as the pitch period.
- In unvoiced region the pitch period as determined by such a method is essentially random.