Speech Features

Notes on Speech and Audio Processing

Chia-Ping Chen

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

Introduction

- (Speech) feature extraction = signal processing
- Different applications require different features for representation. In the case of speech recognition, features are designed such that different linguistic units can be distinguished.
- The basic speech feature extraction archetypes are filter bank, cepstral analysis and linear prediction. We start with the filter bank.

Short-Term Spectrum

- One of the key measurements used in speech processing is the short-term spectrum. It is a local spectral estimate, typically over $20 - 30$ ms.
- It has been shown to be useful in speech recognition and speech coding.
- The basic idea is to represent the time-varying spectral envelope for the speech. Each short-time estimate is an envelope.
- The filter-bank approach use power estimate from a bank of band-pass filters.

Critical-Band Experiments

- The listener is presented with a tone and a wide-band noise. For a given bandwidth,
  - Initially, the tone has a low enough intensity that it is not perceived.
  - Then the tone intensity is increased gradually until it is barely perceived. The intensity is recorded and called the threshold intensity.
- The threshold intensity remains unchanged until the bandwidth is reduced beyond a critical value. The band is called the critical band.
- This suggests the existence of an auditory filter in the vicinity of the tone.

Critical Band Properties

- For the perception of a tone, it is the SNR within the critical band that counts.
- The width of a critical band increases with the tone frequency. But the rate of increasing is not constant: it increases slower with the center frequency in low-frequency range, and faster in high-frequency range.
- In practice, we often design filters to have constant bandwidth at low frequencies and increasing bandwidth at high frequencies.

Filter Shapes I

- The determination of the filter shape is more difficult. It can be done with psychoacoustic tests.
- In the test, there are two non-overlapping bands of noise, one low-pass (600 Hz) and one high-pass (1200 Hz). The listener is asked to detect a tone as its frequency varies from 400 to 1400 Hz. For each tone, the SPL that it is barely audible is recorded.
- The results is mapped against curves obtained by hypothetical filter shapes, such as rectangular, resonant and symmetric.
- This is shown in Figure 19.3. It appears that the rectangular is the worst.

Filter Shapes II

- In the previous experiment, the noise band is fixed and the tone frequency is varied. It can also be done by varying the noise band while keeping the tone frequency fixed, as shown in Figure 19.4.
- For each choice of noise band W, the power P for the tone to be barely heard is recorded.

$$P = K \int_0^W N(f)|H(f)|^2 df.$$

- In the case that $N(f) = N_0$ (flat noise spectrum), $|H(W)|^2 = \frac{1}{KN_0}\frac{dP}{dW}$.

Skirts of Auditory Filters

- Apply low-pass or high-pass noise bands.
- The signal level that 75% of the subjects discern the tone is recorded. This is shown in Figure 19.5.
- The spread of the skirts clearly shows the filter bandwidth increases with the tone frequency.

Symmetric Filter

- If the filter is off-center, computing $H(f)$ becomes difficult.
- The noise band from one side can be replaced by notched wideband noise, i.e. noise bands from both sides of the tone, to avoid the problem of off-center filter shape.
- Patterson showed that experimental results could be quite accurately represented by the symmetric filter:

$$|H(f)|^2 = \frac{1}{[(\Delta f/\alpha)^2 + 1]^2}.$$

Filter Bank Design

- In designing filters, we want a reasonable emulation of the auditory filter bank.
- How many filters do we need?
- How are the center frequencies distributed?
- Are the filters narrow-band or wide-band?
- Do we apply temporal filter to the output?
- How frequent is the spectrum estimate updated?

FFT

- We can use FFT as a spectral analyzer. A 1024-point FFT is equivalent to 1024 bandpass filters.
- The 1024 points do not have to come entirely from speech samples. If the analysis window is 25 ms, then we may use just 25 ms of data and pad the remaining by zeroes. For 8000 Hz speech sampling rate, this is 200 samples plus 824 zeroes.
- In addition, one may want to multiply the samples by Hamming or Kaiser window function to remove the artifacts of rectangular windows.

Introduction to Cepstrum

- 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,

$$\begin{aligned} |X(\omega)| &= |E(\omega)|\,|V(\omega)| \\ \Rightarrow \log|X(\omega)| &= \log|E(\omega)| + \log|V(\omega)|. \end{aligned}$$

- $E(\omega)$ 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).

Linear Prediction

- Speech can be modeled as been produced by a periodic or noise-like source that drives an non-uniform acoustic tube.
- Based on this model of speech production, we will describe the feature set of the linear prediction coefficients. They are succinct and smooth.

Resonances and Formants

- A tube driven by periodic pulses has a set of resonant frequencies.
- Resonance corresponds to peaks in the frequency response. The resonant frequencies of vocal tract are called the formants.
- There are $4 - 5$ formants in the frequency range we are normally interested in, i.e., approximately one formant per kHz.

Predictive Model

- (Sec 6.8) The following transfer function represents a resonance,

$$H(z) = \frac{1}{1 - bz^{-1} - cz^{-2}}.$$

- Multiple formants can be modeled by a cascade of such resonators,

$$H(z) = \frac{1}{1 - \sum_{i=1}^{P} a_i z^{-i}},$$

$$\text{or} \;\; y[n] = x[n] + \sum_{i=1}^{P} a_i y[n-i].$$

Comments on Predictive Model

- The second term on the rhs $\tilde{y}[n] = \sum_{i=1}^{P} a_i y[n-i]$ is a linear predictor of $y[n]$.
- The difference between $\tilde{y}[n]$ and $y[n]$ is called the residual (or prediction error) signal.
- The energy of the error signal is minimized to compute a_i, called the linear prediction coefficients.
- The transfer function has only poles, so it is also called the all-pole model.

Getting the Coefficients

- Define the error signal

$$e[n] = y[n] - \tilde{y}[n]$$

- Let $D = \sum_{n=0}^{N-1} e^2[n]$ and we want to find the $a_i's$ that minimizes D. At the minimum value, the first-order partial derivative must vanish, and

$$\sum_{i=1}^{P} a_i \phi(j,i) = \phi(j,0), \text{ for } j = 1, \ldots, P$$

- a_i's can be solved using the Levinson-Durbin recursion.