Time-Domain Methods for Speech Processing

Introduction

Figure 1 illustrates the speech production model universally used in speech signal processing.

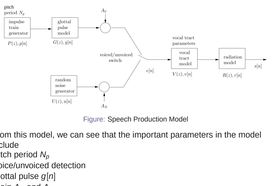

Figure: Speech Production Model

From this model, we can see that the important parameters in the model include

- pitch period N_p
- voice/unvoiced detection
- glottal pulse $g[n]$
- Gain A_V and A_N
- vocal tract model $v[n]$
- radiation model $r[n]$

Short-Time Speech Analysis

Since speech is time-varying, the model parameters are also time-varying and we need a short-time analysis to estimate them. Furthermore, from speech samples to model parameters, alternative short-time representations are often required.

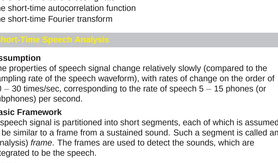

Figure: Short-time Speech Analysis for Parameter Estimation

Here we introduce the representations of

- the short-time energy
- the short-time zero-crossing rate
- the short-time autocorrelation function
- the short-time Fourier transform

Short-Time Speech Analysis

Assumption

The properties of speech signal change relatively slowly (compared to the sampling rate of the speech waveform), with rates of change on the order of $10-30$ times/sec, corresponding to the rate of speech $5-15$ phones (or subphones) per second.

Basic Framework

A speech signal is partitioned into short segments, each of which is assumed to be similar to a frame from a sustained sound. Such a segment is called an (analysis) *frame*. The frames are used to detect the sounds, which are integrated to be the speech.

Frame

- window function $w[n]$: the (finite-duration) function used to extract a frame from the speech waveform
  - rectangular window vs. Hamming window

$$w_R[n] = \begin{cases} 1, & 0 \le n \le L-1 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

$$w_H[n] = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{L}\right) & 0 \le n \le L-1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

- frame length (size) L: the length (seconds or samples) of a frame
  - short frames ($5-20$ ms), medium frames ($20-100$ ms), long frames ($100-500$ ms)
- frame shift R: the time between frames

General Representation of Short-time Processing

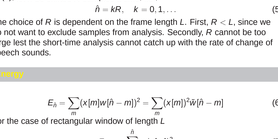

Figure: General Representation of Short-time Analysis

All the short-time processing we will discuss can be represented mathematically as

$$Q_{\hat{n}} = \sum_{m} T(x[m])\hat{w}[\hat{n}-m] \tag{3}$$

In (3), $T(\cdot)$ is meant to extract certain feature(s) of the speech signal. The feature(s) is then summed over a window $\hat{w}[\hat{n}-m]$ anchored at $\hat{n}$. The result is a short-time feature near $\hat{n}$. For example, for a rectangular window with length L,

$$Q_{\hat{n}} = \sum_{m=\hat{n}-L+1}^{\hat{n}} T(x[m]) \tag{4}$$

Frame Shift

In (3), let $\hat{n}$ be shifted R samples a time, i.e.,

$$\hat{n} = kR, \quad k = 0, 1, \ldots \tag{5}$$

The choice of R is dependent on the frame length L. First, $R < L$, since we do not want to exclude samples from analysis. Secondly, R cannot be too large lest the short-time analysis cannot catch up with the rate of change of speech sounds.

Energy

$$E_{\hat{n}} = \sum_{m} (x[m]w[\hat{n}-m])^2 = \sum_{m} (x[m])^2\hat{w}[\hat{n}-m] \tag{6}$$

For the case of rectangular window of length L

$$E_{\hat{n}} = \sum_{m=\hat{n}-L+1}^{\hat{n}} (x[m])^2 \tag{7}$$

Without additive noises, $E_{\hat{n}}$ is useful in detecting voiced segments of speech. It is also useful to detect silence segments.

Magnitude

The short-time magnitude is similar to the short-time energy. Only $(\cdot)^2$ is replaced by $|\cdot|$.

Zero-Crossing Rate

Zero-Crossing

For discrete-time signals, a zero-crossing occurs if two adjacent samples are of different signs.

Consider a sinusoidal wave $x_c(t)$ of frequency F_0. If the sampling period T is not longer than half the period $1/2F_0$, then the number of zero-crossing per second of $x[n] = x_c(nT)$ is $2F_0$, since there are two zero-crossing in $x[n]$ per cycle of $x_c(t)$ and there are F_0 cycles per second. The number of zero-crossings per sample is

$$Z^{(1)} = \frac{2F_0}{F_S} \tag{8}$$

where $F_S \triangleq \frac{1}{T}$ is the sampling frequency. The number of zero-crossings per M samples is

$$Z^{(M)} = M\left(\frac{2F_0}{F_S}\right) \tag{9}$$

Reversing (8), the equivalent sinusoidal frequency can be defined as

$$F_e = \frac{1}{2}F_S Z^{(1)} \tag{10}$$

Examples

Suppose $F_S = 10000$ samples/sec. For

$$F_0 = 100, 1000, 5000 \quad \text{Hz}$$

of sinusoidal signals, the zero-crossing rates are

$$0.02, 0.2, 1$$

crossings/sample respectively. (Note the typo in text)

Practical Considerations

It is vulnerable to noise and DC offset.

General Representation

In terms of (3), we can write

$$Z_{\hat{n}} = \frac{1}{2L} \sum_{m=\hat{n}-L+1}^{\hat{n}} |\text{sgn}\, x[m] - \text{sgn}\, x[m]| \tag{11}$$

Usage

Together with energy, Z or $Z^{(M)}$ can be used in the classification of voiced/unvoiced segments of speech signal. Specifically, a voiced segment is high in E and low in Z.

Autocorrelation Function

The deterministic autocorrelation function of a discrete-time signal $x[n]$ is defined as

$$\phi[k] = \sum_{m=-\infty}^{\infty} x[m]x[m+k] \tag{12}$$

If $x[n]$ is random or periodic, the useful definition is

$$\phi[k] = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{m=-N}^{N} x[m]x[m+k] \tag{13}$$

Properties

$\phi[k] = \phi[-k]$; $\phi[k] \le \phi[0]$; $\phi[0]$ is the signal energy or power.

Short-time Autocorrelation

At analysis time $\hat{n}$, the short-time autocorrelation is defined as the autocorrelation function of the windowed segment, i.e.,

$$R_{\hat{n}}[k] = \sum_{m=-\infty}^{\infty} (x[m]w[\hat{n}-m])(x[m+k]w[\hat{n}-k-m]) \tag{14}$$

The properties holding for $\phi[k]$ also hold for $R_{\hat{n}}[k]$.

General Representation

In terms of (3), we can write

$$\begin{aligned} R_{\hat{n}}[k] &= R_{\hat{n}}[-k] \\ &= \sum_{m=-\infty}^{\infty} (x[\hat{n}-k])(w[\hat{n}-m]w[\hat{n}+k-m]) \\ &= \sum_{m=-\infty}^{\infty} (x[m]x[m-k])\hat{w}_k[\hat{n}-m], \end{aligned} \tag{15}$$

where $\hat{w}_k[\hat{n}-m]$ is the effective window for lag k defined as $\hat{w}_k[\hat{n}] = w[\hat{n}]w[\hat{n}+k]$.

Computation

(14) can be re-written as

$$R_{\hat{n}}[k] = \sum_{m=-\infty}^{\infty} (x[\hat{n}+m]w'[m])(x[\hat{n}+m+k]w'[k+m]), \tag{16}$$

where $w'[m] = w[-m]$, and the change of variable $m \to \hat{n}+m'$ has been made. Considering $w'[m]$, we can see there are only certain non-zero terms in (17), which becomes

$$R_{\hat{n}}[k] = \sum_{m=\hat{n}}^{\hat{n}-L+k} (x[\hat{n}+m]w'[m])(x[\hat{n}+m+k]w'[k+m]) \tag{17}$$

Practical Considerations

$R_{\hat{n}}[k]$ decreases with k as the number of terms in the summation is $L-k$. Some modification can be made such that there are L terms for each k.

Usage

For voiced segments, the autocorrelation function shows periodicity. Furthermore, the rectangular window works better than the Hamming window.