AI2651

Intelligent Speech ~~Distinguish~~ Recognition

T.H.E. Note

YBiuR

魔法少女HMM

Preface

Hey Siri.

Chapter 1

Foundamentals of Speech Signal Processing

The basic concepts of signal processing have been covered in EI015 Signals and Systems, and therefore will not be mentioned in this note.
In other words, I am lazy.

1.1 Discrete Fourier Transform

Similar to the continuous case, given a periodic discrete signal $\tilde{x}[n]$[1] with period N, it can be represented by *a discrete sum of sinusoids*, rather than an integral (Recall the DTFT Synthesis Formula. Thank you, EI015!).

$\tilde{x}[n]$ can be represented as a sum of complex exponentials with radian frequency $(2\pi k/N)$, where $k = 0, 1, \ldots, N-1$.

$$\tilde{X}[k] = \sum_{n=0}^{N-1} \tilde{x}[n] e^{-j\frac{2\pi}{N}kn}$$

And the corresponding synthesis expression is

$$\tilde{x}[n] = \frac{1}{N} \sum_{k=0}^{N-1} \tilde{X}[k] e^{j\frac{2\pi}{N}kn}$$

This representation of a periodic discrete signal is *exact*. However, the DFT is generally used in another case, where $x[n]$ is a *finite* sequence of signal. Since performing DFT only need $\tilde{x}[n]$ in a period $0 \leq n \leq N-1$, and whatever is out of this range does not matter, we may extend $x[n]$ and assume an "implicit periodic sequence" $\tilde{x}[n]$:

$$\tilde{x}[n] = \sum_{r=-\infty}^{+\infty} x[n + rN]$$

And here comes our DFT.

Definition 1.1.1 (Discrete Fourier Transform)**.**

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn} \quad k = 1, 2, \ldots, N-1$$

$$x[n] = \sum_{k=0}^{N-1} X[k] e^{j\frac{2\pi}{N}kn} \quad n = 1, 2, \ldots, N-1$$

Remark. Bear in mind that when using DFT representations, all signals behave as if they were implicitly *periodic*, as the DFT is originally defined on periodic signals.

[1] The tilde sign here means that $\tilde{x}[n]$ is periodic.

1.2 Voiced, Unvoiced and Silence

The speech waveform can be classified into basically 3 stages.

Unvoiced Produced by creating a constriction somewhere in the vocal tract tube and forcing air through that constriction, thereby creating turbulent air flow, which acts as *a broad-spectrum noise excitation of the vocal tract tube.*

Voiced Produced by forcing air through the glottis with the tension of the vocal cords adjusted so that they vibrate in a relaxation oscillation, leading to *quasi-periodic* waveforms.

Silence Usually occurs at the beginning or the end of speech, lacks characteristics of either voiced sounds or unvoiced sounds.

Chapter 2

Short-Time Speech Processing

Whoever translated the name of this course should be sent back to high school to re-learn English.

2.1 Segmentations/Frames

An underlying assumption of speech signal processing is that the speech signal is slowly-time-varying, i.e. it changes slowly with time. Based on this assuption, we can perform short-time signal processing by splitting speech signals into isolated segmentations or frames.

2.1.1 Mathematical Framework of Short-Time Processing

Short-time analysis is represented in a general form of

$$Q_{\hat{n}} = \sum_{m=-\infty}^{+\infty} T(x[m]w[\hat{n}-m]) = \sum_{m=-\infty}^{+\infty} T(x[m])\tilde{w}[\hat{n}-m]$$

where $\tilde{w}[\hat{n}-m]$ is a sliding analysis window and $T()$ is the operation on input signal. $Q_{\hat{n}}$ represents the short-time representation of signal $\tilde{x}$ at time $\hat{n}$.

Remark. It's actually a discrete-time convolution of $T(x[m])$ with $\tilde{w}[n]$

2.1.2 Length of Segments

The shorter the segement, the less likely a signal will vary significantly over the segment duration (due to its slowly-changing nature), and thus tracking abrupt waveform changes is best for shorter segements. However, parameters estimated from short may be highly variable because the data available for processing is small.

2.1.3 Stepsize

Typically, the window is moved in jumps of $R > 1$ samples, which corresponds to downsampling the output of the signal by a factor of R. If the window is of length L, then we should choose $R < L$ so that each sample is included in at least one segement.
Typically, the analysis windows overlap by more than 50% of the window length.

2.1.4 Commonly Used Windows

- Rectangular Window

$$w_R[n] = \begin{cases} 1 & (0 \le n \le L-1) \\ 0 & (\text{o.w.}) \end{cases}$$

- Hamming Window

$$w_H[n] = \begin{cases} 0.54 - 0.46\cos(2\pi n/(L-1)) & (0 \le n \le L-1) \\ 0 & (\text{o.w.}) \end{cases}$$

2.2 Short-Time Energy and Short-Time Magnitude

2.2.1 Energy of Signal

Definition 2.2.1 (Energy)**.** The energy of a discrete-time signal is

$$E = \sum_{m=-\infty}^{+\infty} (x[n])^2$$

Useless because it does not give any time-dependent properties of a speech signal.

2.2.2 Short-Time Energy

Definition 2.2.2 (Short-Time Energy)**.**

$$E_{\hat{n}} = \sum_{m=-\infty}^{+\infty} (x[m]w[\hat{n}-m])^2 = \sum_{m=-\infty}^{+\infty} (x[m])^2 \tilde{w}[\hat{n}-m]$$

where $w[n]$ is the window applied to $x[n]$ before squaring, and $\tilde{w}[n]$ is the corresponding window that can be applied equivalently after squaring.
For an L-point window,

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

$E_{\hat{n}}$ provides a basis for distinguishing voiced speech segments from unvoiced speech segments. Unvoiced segments have significantly smaller short-time energy compared with Voiced segments. For high-quality speech signal (high signal-to-noise ratio), $E_{\hat{n}}$ can also be used to distinguish speech from silence.

2.2.3 Short-Time Magnitude

Short-time energy can be very sensitive to large signal levels. This can be addressed by taking square roots, or using the short-time magnitude.

Definition 2.2.3 (Short-Time Magnitude)**.**

$$M_{\hat{n}} = \sum_{m=-\infty}^{+\infty} |x[m]w[\hat{n}-m]| = \sum_{m=-\infty}^{+\infty} |x[m]|\tilde{w}[\hat{n}-m]$$

2.3 Short-Time Zero-Crossing Rate

A **zero-crossing** is said to occur if successive waveform samples have different algebraic signs.

Definition 2.3.1 (Short-Time Zeor-Crossing Rate)**.**

$$Z_{\hat{n}} = \frac{1}{2L_{eff}} \sum_{m=\hat{n}-L+1}^{\hat{n}} |\text{sgn}(x[m])| - \text{sgn}(x[m-1])|\tilde{w}[\hat{n}-m]$$

Typically, the window used here is a rectangular window, so

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

Voiced speech will have relatively low zero-crossing rate, and Unvoiced speech will have relatively high zero-crossing rate.

2.4 Short-Time AutoCorrelation

2.4.1 AutoCorrelation

Definition 2.4.1 (AutoCorrelation)**.** For determinisitc or aperiodic signals

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

For stationary random or periodic signals

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

Remark. AutoCorrelation highlights the period of signals because a local maximum is achieved at samples $0, \pm N, \pm 2N, \ldots$, regardless of the time origin of the periodic signal.

Proposition 2.4.1. Properties of autocorrelation:

- For periodic signals:
$$\phi[k] = \phi[k+N]$$
- $\phi[k] = \phi[-k]$
- $\phi[0] \leq |\phi[k]|$ for all k
- $\phi[0]$ equals to the total energy for determinisitc signals and average power of random signals.

2.4.2 Short-Time AutoCorrelation

Definition 2.4.2 (Short-Time AutoCorrelation)**.**

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

Remark. Short-Time AutoCorrelation preserves the properties of autocorrelation in the previous section, and $R_{\hat{n}}[0]$ is equivalent to the short-time energy.

2.5 Short-Time Average Magnitude Difference Function

The computation of autocorrelation involves considerable arithmetic.
Note that for a truly periodic signal,

$$d[k] = x[n] - x[n-k]$$

would be 0 whenever $k = 0, \pm N, \pm 2N, \ldots$. So it is reasonable to expect that $d[n]$ will be small at multiples of period for short segements of voiced speech.

Definition 2.5.1 (Short-Time AMDF)**.**

$$\gamma_{\hat{n}}[k] = \sum_{m=-\infty}^{+\infty} |x[\hat{n}+m]w_1[m] - x[\hat{n}+m-k]w_2[m-k]|$$

Remark. $\gamma_{\hat{n}}[k]$ would drop sharply near multiples of period.

AutoCorrelation and AMDF are used to find the **pitch** (aka **foundamental frequency**) of the speech.

2.6 Short-Time Fourier Transform

Definition 2.6.1 (Continuous Time STFT)**.**

$$X(\hat{t}, \Omega) = \int_{-\infty}^{+\infty} w(\hat{t} - t)x(t)e^{-j\Omega t}\mathrm{d}t$$

Definition 2.6.2 (Discrete Time STFT)**.**

$$X_{\hat{n}}(e^{j\omega}) = \sum_{m=-\infty}^{+\infty} w[\hat{n} - m]x[m]e^{-j\omega m}$$

Definition 2.6.3 (Spectrogram)**.** A spectrogram is a gray-scale image whose x-axis is time and y-axis is frequency, and its colormap indicates the log amplitude.

Remark. If the band pass filter has wide bandwidth(300-900Hz), then the spectrogram has good temporal resolution but poor frequency resolution. On the other hand, if the band pass filter has narrow bandwidth(30-90Hz), then the spectrogram has poor temporal resolution but good frequency resolution.