Lecture 6: Speech Signal Processing

Kai Yu

Speech Lab
Department of Computer Science & Engineering
Shanghai Jiao Tong University

July 2013
Table of Content

- Digital speech waveforms
- Basic waveform processing
- Spectral analysis
**Types of signal**

**Perodicity**

**Deterministic signal:** evolve according to a formulae

**Periodic** signal: repeats exactly with period \( \tau \)

**Aperiodic** signals: any non-periodic signal

**Stochastic signal:** signal at time \( t \) is a function of random variable: by definition not exactly predictable
Types of Signal

Discrete and continuous: 4 possibilities

Discrete-time/value signal:

Continuous-time/value signal:
Speech sounds are produced by vibratory activity in the human vocal tract.

Speech is normally transmitted to a listener’s ears or to a microphone through the air, where speech and other sounds take on the form of radiating waves of variation in air pressure.

These type of waves are known as longitudinal waves.
Digital Speech Waveforms

When we speak into a microphone, these changes in pressure are converted to proportional variations in electrical voltage. Computers equipped with the proper hardware can convert the analog voltage variations into digital sound waveforms by a process called analog-to-digital conversion (ADC), which involves:
Digital Speech Waveforms (Cont’d)

- **Sampling:** taking a fixed number of pressure value readings at equal time intervals from the continuously varying speech signal. Typical rate is 16,000 times per second, i.e. 16K Hz for PC and 8,000 for telephone.

- **Quantization:** represent the sampled waveform amplitudes as discrete values (rounded to the nearest value which is expressible in a given number of bits). For example, 8 bits and 16 bits can represent a total of 256 and 65536 possible quantization levels respectively.

- **Compression:** represent the quantized values in a more compact form to save space for transmission or storage. Typical compression includes $\mu$-law, A-law etc.

- Digital speech waveform is **stochastic discrete signal**
By sampling, the analog "continuous time" signals can be converted to digital "discrete time" signals.
▶ Speech waveform consists of a long sequence of sampled values. Useful to break the long sequence into blocks/frames which are quasi-stationary
▶ **Frame-size** is a compromise between:
  ▶ having sufficient sample points for accurate analysis
  ▶ ensuring that the quasi-stationary assumption is valid
▶ **Frame shift** number of samples (seconds) between start of successive frames. use overlapped frames to better represent the signal dynamics
Short-time energy is defined as the sum of squares of the samples in a frame:

$$E = \sum_{i=0}^{N-1} s_i^2$$
Zero-crossing rate (ZCR) is defined as the number of times the zero axis is crossed per frame. ZCR is large for unvoiced speech.

```c
int ZCR(float s[]) {
    int count = 0;
    for (int i=1; i<s.length; i++)
        if (s[i-1]*s[i] <= 0)
            count++;
    return count;
}
```
Basic waveform processing

Zero-crossing rate (II)
Auto-correlation emphasizes periodicity of waveforms. Normally calculated in a wider window.

\[ r_k = \sum_{i=0}^{N-k-1} s_i s_{i+k} \]

where \( k \) is the correlation period.

```c
float AutoCorr(float s[], int k) {
    float sum = 0;
    for (int i=0; i<s.length-k; i++)
        sum += s[i] * s[i+k];
    return count;
}
```
Basic waveform processing

Auto-correlation
Pitch or fundamental frequency $F_0$ is the lowest frequency of human speech which forms higher frequency components.

- Voiced region with strong periodicity while unvoiced region with just turbulence
- Human $F_0$ ranges from 60Hz to 300Hz
Pitch Detector

1. Compute \( r_0 \) to \( r_{max} \), where \( r_0 \) is the energy
2. Find peak \( r_p \) in the range \( r_0 \) to \( r_{max} \)
3. if \( r_p > 0.3r_0 \) then speech is voiced with period \( p \)
4. otherwise the frame is unvoiced
Any periodic signal of frequency $f_0$ can be constructed exactly by adding together sinusoids with frequencies $f_0, 2f_0, 3f_0, 4f_0, 5f_0, \ldots$ each with the appropriate amplitude and phase. $f_0$ is called the **fundamental frequency** and $2f_0, 3f_0$ etc are the **harmonics**.

$$s_n = s(nT) = \sum_{p=0}^{N-1} A_p \cos(\omega pnT + \phi_p)$$

where $A_p$ and $\phi_p$ are known as the **amplitude** and **phase** of the $p^{th}$ harmonic.
Fourier analysis
Example: a square wave (odd harmonics only)
Any periodic function can be characterised by the amplitude and phase of its sinusoidal components. This characterisation is called the spectrum.
Fourier Analysis

Aperiodic signal

- Periodic signal only have spectral components at integer multiples of the fundamental
- Aperiodic and stochastic signals have spectra that are continuous functions of frequency i.e. at all possible frequencies
Assume $f(t)$ is a continuous time signal, if $\int_{-\infty}^{+\infty} |f(t)| \, dt < \infty$, the \textbf{fourier transform} exists and is defined as

$$F(jw) = \int_{-\infty}^{+\infty} f(t) e^{-jwt} \, dt$$

where $e^{-jwt} = \cos(\omega t) + j \sin(\omega t)$.

The \textbf{inverse fourier transform} is defined as:

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F(jw) e^{j\omega t} \, dw$$

Note: \textbf{Periodic signal} can be expanded to fourier series and its fourier transform results in pulse trains.
Discrete-time fourier transform

\[ F(e^{jw}) = \sum_{n=-\infty}^{+\infty} f[n] e^{-jwn} \]

Inverse discrete-time fourier transform

\[ f[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} F(e^{jw}) e^{jwn} \, dw \]

Note: \( F(e^{jw}) \) is periodic

\[ F(e^{j(w+2\pi)}) = F(e^{jw}) \]
Relationship Between Time domain and Frequency domain

<table>
<thead>
<tr>
<th>Time</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous + Aperiodic</td>
<td>Continuous + Aperiodic</td>
</tr>
<tr>
<td>Continuous + Periodic</td>
<td>Discrete + Aperiodic</td>
</tr>
<tr>
<td>Discrete + Aperiodic</td>
<td>Continuous + Periodic</td>
</tr>
<tr>
<td>Discrete + Periodic</td>
<td>Discrete + Periodic</td>
</tr>
</tbody>
</table>
Discrete Fourier Transform

Discrete time and frequency signal with finite length
Discrete Fourier Transform

Calculation

DFT is a **linear transform** of the waveform signal.

\[ F[k] = \sum_{n=0}^{N-1} f[n] W_N^{kn} \]

\[ k = 0, 1, \ldots, N - 1 \]

\[ f[n] = \frac{1}{N} \sum_{k=0}^{N-1} F[k] W_N^{-kn} \]

\[ n = 0, 1, \ldots, N - 1 \]

\[ W_N = e^{-j \frac{2\pi}{N}} \]
DFT Property

- **Symmetric:**

  \[
  |F[k]| = |F(N - k)| \quad \text{arg}(F[k]) = - \text{arg}(F[N - k])
  \]

- **Law of conservation of energy (Parseval Theorem)**

  \[
  \sum_{n=0}^{N-1} |x(n)|^2 = \frac{1}{N} \sum_{n=0}^{N-1} |F[k]|^2
  \]
If the sampling rate is $f_s = \frac{1}{\Delta T}$, then time resolution (duration) is $N\Delta T$, frequency resolution is $\frac{f_s}{N} = \frac{1}{N\Delta T}$.

As the length of the analysis window (N) increases:

- **Time resolution** is poorer as sudden changes can not be effectively modelled.
- **Frequency resolution** is more accurate.
- **Zero-padding**: add trailing zeros to the signal to increase $N$.
  - yields more frequency points but doesn’t increase real resolution as no info. is added.
Implicit Periodicity of DFT

- DFT evaluates spectrum at $N$ evenly split discrete frequencies
- Only periodic signals have discrete spectrum
- DFT assumes periodicity outside the analysis window
- Such assumption may
  - give rise of boundary effect (edge effect)
  - distort high-frequency components
Implicit Periodicity Example
In signal processing, a *window function* is a function that is zero-valued outside the chosen region. Signal is multiplied by the window function to reduce the boundary effect.

- **Hamming window:**
  \[
  w(n) = 0.53836 - 0.46164 \cos\left(\frac{2\pi n}{N-1}\right), \quad n = 0, \cdots, N-1
  \]

- **Hanning window:**
  \[
  w(n) = 0.5 \left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right)
  \]
Implicit Periodicity with Windowing

Graphs showing periodic functions with and without windowing effects.
**Motivation:** within a band of frequencies, to increase the magnitude of some frequencies (usually higher) w.r.t. the magnitude of other (usually lower) frequencies to increase the overall signal-to-noise ratio.

**Calculation:**

\[ x_n = x_n - \alpha x_{n-1} \]

where \( \alpha \) is the pre-emphasis factor (typically 0.97). The boundary condition \( x_{-1} = x_0 \) is assumed.
Spectral Analysis with and without Pre-emphasis

Figure: Graphs showing spectral analysis with and without pre-emphasis.
Speech Signal Processing - Short Time Fourier Transform

- Assume speech signal is quasi-stationary
- Divide speech into short-time segments, e.g. 10ms
- Apply DFT to each segment
Trade-off between time and frequency resolution has to be made
- Adjust **window length** and **window overlap** to make trade-off
  - **Middle**: 256 points window and 50% overlap (better frequency resolution)
  - **Bottom**: 64 points window and 50% overlap (better time resolution)