Digital Speech Processing—Lectures 7-8

Time Domain Methods in Speech Processing

General Synthesis Model

\[ R(z) = 1 - az^{-1} \]

Log Areas, Reflection Coefficients, Formants, Vocal Tract Polynomial, Articulatory Parameters, ...

Pitch Detection, Voiced/Unvoiced/Silence Detection, Gain Estimation, Vocal Tract Parameter Estimation, Glottal Pulse Shape, Radiation Model

General Analysis Model

- All analysis parameters are time-varying at rates commensurate with information in the parameters;
- We need algorithms for estimating the analysis parameters and their variations over time

Overview

- time domain processing \( \Rightarrow \) direct operations on the speech waveform
- frequency domain processing \( \Rightarrow \) direct operations on a spectral representation of the signal

- simple processing
- enables various types of feature estimation

Basics

- 8 kHz sampled speech (bandwidth < 4 kHz)
- properties of speech change with time
- excitation goes from voiced to unvoiced
- peak amplitude varies with the sound being produced
- pitch varies within and across voiced sounds
- periods of silence where background signals are seen
- the key issue is whether we can create simple time-domain processing methods that enable us to measure/estimate speech representations reliably and accurately

Fundamental Assumptions

- properties of the speech signal change relatively slowly with time (5-10 sounds per second)
  - over very short (5-20 msec) intervals \( \Rightarrow \) uncertainty due to small amount of data, varying pitch, varying amplitude
  - over medium length (20-100 msec) intervals \( \Rightarrow \) uncertainty due to changes in sound quality, transitions between sounds, rapid transients in speech
  - over long (100-500 msec) intervals \( \Rightarrow \) uncertainty due to large amount of sound changes
- there is always uncertainty in short time measurements and estimates from speech signals
Compromise Solution

• “short-time” processing methods => short segments of the speech signal are "isolated" and "processed" as if they were short segments from a "sustained" sound with fixed (non-time-varying) properties
  – this short-time processing is periodically repeated for the duration of the waveform
  – these short analysis segments, or "analysis frames" almost always overlap one another
  – the results of short-time processing can be a single number (e.g., an estimate of the pitch period within the frame), or a set of numbers (an estimate of the formant frequencies for the analysis frame)
  – the end result of the processing is a new, time-varying sequence that serves as a new representation of the speech signal

Frame-by-Frame Processing in Successive Windows

Frame 1: samples 0,1,...,L-1
Frame 2: samples R,R+1,...,R+L-1
Frame 3: samples 2R,2R+1,...,2R+L-1
Frame 4: samples 3R,3R+1,...,3R+L-1

Frames and Windows

Short-Time Processing

\[ x[n] = \text{samples at 8000/sec rate; (e.g. 2 seconds of 4 kHz bandlimited speech, } x[n], 0 \leq n \leq 16000) \]
\[ f[m] = \text{vectors at 100/sec rate, } 1 \leq m \leq 200, \]
\[ L \text{ is the size of the analysis vector (e.g., 1 for pitch period estimate, 12 for autocorrelation estimates, etc) } \]
Generic Short-Time Processing

\[ Q_n = \left( \sum_{m=-\infty}^{\infty} T(x[m]) \tilde{w}[n - m] \right)_{n=0}^{\infty} \]

- \( Q_n \) is a sequence of local weighted average values of the sequence \( T(x[n]) \) at time \( n = \tilde{n} \)

Short-Time Energy

\[ E = \sum_{n=-\infty}^{\infty} x'[m] \]

- this is the long term definition of signal energy
- there is little or no utility of this definition for time-varying signals

\[ E_n = \sum_{m=-\infty}^{\infty} x'[m] = x'[n-L+1] + ... + x'[n] \]
- short-time energy in vicinity of time \( n \)

\[ \tilde{w}[n] = 1 \quad 0 \leq n \leq L - 1 \]
\[ = 0 \quad \text{otherwise} \]

Effects of Window

\[ Q_n = T(x[n]) * \tilde{w}[n] \]
\[ = x'[n] * \tilde{w}[n] \]

- \( \tilde{w}[n] \) serves as a lowpass filter on \( T(x[n]) \) which often has a lot of high frequencies (most non-linearities introduce significant high frequency energy—think of what \( (\hat{x}[n] \cdot \hat{x}[n]) \) does in frequency)
- often we extend the definition of \( Q_n \) to include a pre-filtering term so that \( \hat{t}[n] \) itself is filtered to a region of interest

Short-Time Energy Properties

- depends on choice of \( h[n] \), or equivalently, window \( \tilde{w}[n] \)
  - if \( \tilde{w}[n] \) duration very long and constant amplitude \( (\tilde{w}[n]=1, n=0,1,...,L-1) \), \( E_n \) would not change much over time, and would not reflect the short-time amplitudes of the sounds of the speech => this is the essential conflict in all speech processing, namely we need short duration window to be responsive to rapid sound changes, but short windows will not provide sufficient averaging to give smooth and reliable energy function

Computation of Short-Time Energy

Fig. 4.1 Illustration of the computation of short-time energy

- window jumps/slides across sequence of squared values, selecting interval for processing
- what happens to \( E_n \) as sequence jumps by 2,4,8,... samples (\( E_n \) is a lowpass function—so it can be decimated without loss of information; why is \( E_n \) lowpass?)
- effects of decimation depend on \( L \); if \( L \) is small, then \( E_n \) is a lot more variable than if \( L \) is large (window bandwidth changes with \( L \! \! \! \! \! \! \! \! \) )

Short-Time Energy

- serves to differentiate voiced and unvoiced sounds in speech from silence (background signal)
- natural definition of energy of weighted signal is:
  \[ E_n = \sum_{m=-\infty}^{\infty} [x[m] \tilde{w}[n-m]]^2 \]
  (sum or squares of portion of signal)
- concentrates measurement at sample \( n \), using weighting \( \tilde{w}[n-m] \)
  \[ E_n = \sum_{m=-\infty}^{\infty} x'[m] \tilde{w}'[n-m] = \sum_{m=-\infty}^{\infty} x'[m] h[n-m] \]
  (short time energy)

Linear or non-linear transformation

window sequence (usually finite length)
Windows

- consider two windows, \( \tilde{w}[n] \)
  - rectangular window:
    - \( \tilde{w}[n]=1 \), \( 0\leq n\leq L-1 \) and 0 otherwise
  - Hamming window (raised cosine window):
    - \( \tilde{w}[n]=0.54-0.46 \cos(2\pi n/(L-1)) \), \( 0\leq n\leq L-1 \) and 0 otherwise
- rectangular window gives equal weight to all \( L \) samples in the window \((n,...,n-L+1)\)
- Hamming window gives most weight to middle samples and tapers off strongly at the beginning and the end of the window

Window Frequency Responses

- rectangular window
  \[ H(\omega) = \frac{\sin(\Omega L/2)}{\sin(\Omega/2)} e^{-j\omega(L-1)/2} \]
- first zero occurs at \( f=F_s/L = 1/(LT) \) (or \( \Omega=(2\pi)/LT \)) => nominal cutoff frequency of the equivalent “lowpass” filter
- Hamming window
  \[ \tilde{w}_h[n] = 0.54e^{-j\pi n/(L-1)} - 0.46 \cos(2\pi n/(L-1))\tilde{w}_r[n] \]
- can decompose Hamming Window FR into combination of three terms

Rectangular and Hamming Windows

- log magnitude response of RW and HW
- bandwidth of HW is approximately twice the bandwidth of RW
- attenuation of more than 40 dB for HW outside passband, versus 14 dB for RW
- stopband attenuation is essentially independent of \( L \), the window duration => increasing \( L \) simply decreases window bandwidth
- \( L \) needs to be larger than a pitch period (or severe fluctuations will occur in \( E_0 \)) but smaller than a sound duration (or \( E_0 \) will not adequately reflect the changes in the speech signal)

There is no perfect value of \( L \), since a pitch period can be as short as 20 samples (200 Hz at a 10 kHz sampling rate) for a high pitch child or female, and up to 255 samples (40 Hz pitch at a 10 kHz sampling rate) for a low pitch male; a compromise value of \( L \) on the order of 100-200 samples for a 10 kHz sampling rate is often used in practice.

Short-Time Energy

\[ E_s = \frac{1}{R} \sum_{m=0}^{R-1} (x[m])^2 \]

For \( R \)-point rectangular window,
\[ s[m]=1, \quad m=0,1,...,L-1 \]

giving
\[ E_s = \frac{1}{R} \sum_{m=0}^{R-1} (x[m])^2 \]
### Short-Time Energy using RW/HW

- As L increases, the plots tend to converge (however you are smoothing sound energies).
- Short-time energy provides the basis for distinguishing voiced from unvoiced speech regions, and for medium-to-high SNR recordings, can even be used to find regions of silence/background signal.

### Short-Time Energy for AGC

Can use an IIR filter to define short-time energy, e.g.,

- Time-dependent energy definition

\[
\sigma^2[n] = \sum_{m=n}^{\infty} x^*[m]h[n-m]\sum_{m=0}^{\infty} h[m]
\]

- Consider impulse response of filter of form

\[
h[n] = a^n u[n-1] = a^{n-1} \quad n \geq 1
\]

\[
= 0 \quad n < 1
\]

\[
\sigma^2[n] = \sum_{m=0}^{n} (1-a)x^*[m]a^{n-m}u[n-m-1]
\]

### Recursive Short-Time Energy

\[
x[n] \quad \left(\begin{array}{c} (1) \end{array}\right) x^*[n] z^{-1} (1-a) \]  
\[
\sigma^2[n] \quad \left(\begin{array}{c} (1) \end{array}\right) \sigma^2[n-1] z^{-1} \]

\[
\sigma^2[n] = \alpha \cdot \sigma^2[n-1] + x^*[n-1](1-a)
\]
Use of Short-Time Energy for AGC

Short-Time Magnitude

- short-time energy is very sensitive to large signal levels due to $x^2[n]$ terms
  - consider a new definition of ‘pseudo-energy’ based on average signal magnitude (rather than energy)
    \[ M_t = \sum_{m} |x[m]| \bar{u}[\hat{t} - m] \]
  - weighted sum of magnitudes, rather than weighted sum of squares
    \[
    \hat{M}_t = \frac{1}{L} \sum_{n=0}^{L-1} |x[n]| \\
    \hat{S}_t = \frac{1}{L} \sum_{n=0}^{L-1} |x[n]|^2 \\
    \hat{E}_t = \frac{1}{L} \sum_{n=0}^{L-1} |x[n]|^2
    \]
  - computation avoids multiplications of signal with itself (the squared term)

Short Time Energy and Magnitude—Rectangular Window

Short Time Energy and Magnitude—Hamming Window

Use of Short-Time Energy for AGC

- differences between $E_t$ and $M_t$ noticeable in unvoiced regions
  - dynamic range of $M_t$ ~ square root (dynamic range of $E_t$) => level differences between voiced and unvoiced segments are smaller
  - $E_t$ and $M_t$ can be sampled at a rate of 100/sec for window durations of 20 msec or so => efficient representation of signal energy/magnitude
Other Lowpass Windows

- can replace RW or HW with any lowpass filer
- window should be positive since this guarantees $E_n$ and $M_n$ will be positive
- FIR windows are efficient computationally since they can slide by $R$ samples for efficiency with no loss of information (what should $R$ be?)
- can even use an infinite duration window if its $z$-transform is a rational function, i.e.,

$$h[n] = a^n, \quad n \geq 0, \quad 0 < a < 1$$

$$H(z) = \frac{1}{1 - az^{-1}}, \quad |z| > |a|$$

Short-Time Average ZC Rate

- zero crossing rate is a simple measure of the 'frequency content' of a signal—especially true for narrowband signals (e.g., sinusoids)
- sinusoid at frequency $F_0$ with sampling rate $F_s$ has $F_s / F_0$ samples per cycle with two zero crossings per cycle, giving an average zero crossing rate of

$$z = (2F_s / F_0) \text{ crossings/sample}$$

$$z = M (2F_s / F_0) \text{ crossings/(M samples)}$$

Sinusoid Zero Crossing Rates

Assume the sampling rate is $F_s = 10,000$ Hz

1. $F_s = 100$ Hz sinusoid has $F_s / F_s = 10,000 / 100 = 100$ samples/cycle;
   or $z_1 = 2 / 100 \text{ crossings/sample}, \text{ or } z_1 = 2 / 100 * 100 = 2 \text{ crossings/10 msec interval}$

2. $F_s = 1000$ Hz sinusoid has $F_s / F_s = 10,000 / 1000 = 10$ samples/cycle;
   or $z_2 = 2 / 10 \text{ crossings/sample, or } z_2 = 2 / 10 * 100 = 20 \text{ crossings/10 msec interval}$

3. $F_s = 5000$ Hz sinusoid has $F_s / F_s = 10,000 / 5000 = 2$ samples/cycle;
   or $z_3 = 2 / 2 \text{ crossings/sample, or } z_3 = 2 / 2 * 100 = 100 \text{ crossings/10 msec interval}$

Zero Crossings for Noise

- random Gaussian noise
- random Gaussian noise with dc offset

Zero Crossing for Sinusoids

- 100 Hz sinusoid
- 100 Hz sinusoid with dc offset

Offset=0.75
ZC Rate Definitions

\[ Z_n = \frac{1}{M} \sum_{m=0}^{M-1} |\text{sgn}(x[n+m]) - \text{sgn}(x[n])| \]

\[ \text{sgn}(x[n]) = \begin{cases} 1 & \text{if } x[n] > 0 \\ -1 & \text{if } x[n] < 0 \\ 0 & \text{otherwise} \end{cases} \]

- simple rectangular window:
  \[ w[n] = 1 \quad 0 \leq n < L - 1 \]
  \[ w[n] = 0 \] otherwise

\[
\hat{Z}_n = -\sum_{m=0}^{M-1} |\text{sgn}(x[n+m]) - \text{sgn}(x[n])| \\
\hat{Z}_n = \frac{1}{M} \sum_{m=0}^{M-1} |\text{sgn}(x[n+m]) - \text{sgn}(x[n])| \\
\]

The formal definition of \( Z_n \) is:

\[ Z_n = \frac{1}{M} \sum_{m=0}^{M-1} |\text{sgn}(x[n+m]) - \text{sgn}(x[n])| \]

is interpreted as the number of zero crossings per sample.

For most practical applications, we need the rate of zero crossings per fixed interval of \( M \) samples, which is

\[ z_n = \frac{z}{M} = \text{rate of zero crossings per } M \text{ sample interval} \]

Thus, for an interval of \( T \) sec., corresponding to \( M \) samples we get

\[ z_n = \frac{z}{M} = \frac{T}{M} \]

For a 1000 Hz sinewave as input, using a 40 msec window length \( L \), with various values of sampling rate \( F_s \), we get the following:

\[
\begin{array}{cccc|cc}
F_s & L & z_n & M & z_n \\
8000 & 320 & 0.14 & 80 & 20 \\
10000 & 400 & 0.15 & 100 & 20 \\
16000 & 640 & 0.18 & 160 & 20 \\
\end{array}
\]

Thus we see that the normalized (per interval) zero crossing rate, \( z_n \), is independent of the sampling rate and can be used as a measure of the dominant energy in a band.

ZC Normalization

Zero crossings/10 msec interval as a function of sampling rate

ZC Rates for Speech

• for voiced speech, energy is mainly below 1.5 kHz
• for unvoiced speech, energy is mainly above 1.5 kHz
• mean ZC rate for unvoiced speech is 49 per 10 msec interval
• mean ZC rate for voiced speech is 14 per 10 msec interval
Issues in ZC Rate Computation

• for zero crossing rate to be accurate, need zero DC in signal => need to remove offsets, hum, noise => use bandpass filter to eliminate DC and hum
• can quantize the signal to 1-bit for computation of ZC rate
• can apply the concept of ZC rate to bandpass filtered speech to give a ‘crude’ spectral estimate in narrow bands of speech (kind of gives an estimate of the strongest frequency in each narrow band of speech)

Summary of Simple Time Domain Measures

1. Energy:
   \[ E = \sum_{n} x[n] x[n] \]
   can downsample at rate commensurate with window bandwidth

2. Magnitude:
   \[ M = \sum_{n} |x[n]| |x[n] - |x[n]| | \]

3. Zero Crossing Rate:
   \[ Z = \sum_{n} \text{sgn}(x[n]) - \text{sgn}(x[n] - |x[n]|) \]
   where \( \text{sgn}(x[n]) = 1 \) for \( x[n] > 0 \)
   \( = -1 \) for \( x[n] < 0 \)

Periodic Signals

• for a periodic signal we have (at least in theory \( \Phi[P] = \Phi[0] \) so the period of a periodic signal can be estimated as the first non-zero maximum of \( \Phi[k] \)
  - this means that the autocorrelation function is a good candidate for speech pitch detection algorithms
  - it also means that we need a good way of measuring the short-time autocorrelation function for speech signals

Short-Time Autocorrelation

• for a deterministic signal, the autocorrelation function is defined as:
  \[ \Phi[k] = \sum_{n} x[n] x[n+k] \]

• for a random or periodic signal, the autocorrelation function is:
  \[ \Phi[k] = \lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} x[n] x[n+k] \]
  - if \( x[n] = x[n+P] \) then \( \Phi[k] = \Phi[k+P] \) => the autocorrelation function preserves periodicity
  - properties of \( \Phi[k] \):
    1. \( \Phi[0] \) is even, \( \Phi[k] = \Phi[-k] \)
    2. \( \Phi[k] \) is maximum at \( k = 0 \), \( |\Phi[k]| < |\Phi[0]|, \forall k \)
    3. \( \Phi[0] \) is the signal energy or power (for random signals)
### Short-Time Autocorrelation

\[ R_k[k] = \sum_{n=-L}^{L} [\tilde{x}(m)|\tilde{x}(n-m)^*|] \tilde{x}(m+n) \tilde{w}(n) \]

\( L \) points used to compute \( R_k[k] \);
\( L-k \) points used to compute \( \hat{R_k} \).

### Examples of Autocorrelations

- Autocorrelation peaks occur at \( k=12, 144, \ldots \Rightarrow 120 \) Hz pitch.
- \( \Phi(P) < \Phi(0) \) since windowed speech is not perfectly periodic.
- Over a 401 sample window (40 msec of signal), pitch period changes occur, so \( P \) is not perfectly defined.
- Much less clear estimates of periodicity since window tapers signal so strongly, making it look like a non-periodic signal.
- No strong peak for unvoiced speech.

### Voiced (female) \( L=401 \) (magnitude)

- \( T_0 = N_0 T \)

### Voiced (female) \( L=401 \) (log mag)

\( T_0 \)

### Voiced (male) \( L=401 \)

\( T_0 \)
Effects of Window Size

- choice of $L$, window duration
- small $L$ so pitch period almost constant in window
- large $L$ so clear periodicity seen in window
- as $k$ increases, the number of window points decrease, reducing the accuracy and size of $R_k$ for large $k$ to have a taper of the type $R(k) = \frac{1}{L}$ for $|k| < L$
- allow $L$ to vary with detected pitch periods (so that at least 2 full periods are included)

Modified Autocorrelation

- want to solve problem of differing number of samples for each different $k$ term in $R_k$, so modify definition as follows:

$$R_k = \sum_{n=-\infty}^{\infty} x[n]x[n-k]$$
- where $x[n]$ is a standard $L$-point window, and $x[n]$ is extended window of duration $L + K$ samples, where $K$ is the largest lag of interest
- we can rewrite modified autocorrelation as:

$$R_k = \sum_{n=-\infty}^{\infty} x[n]x[n-k]$$
- where $x[n] = x[n+K]$ and $x[n] = x[n-K]$ for rectangular windows we choose the following:

- $x[n] = \frac{1}{L}$, $0 \leq n \leq L - 1$
- $x[n] = \frac{1}{L}$, $0 \leq n \leq L - 1$
- giving

$$R_k = \sum_{n=-\infty}^{\infty} x[n]x[n-k]$$
- always use $L$ samples in computation of $R_k$
Examples of Modified AC

- Modified Autocorrelations – fixed value of \( L = 401 \)
- Modified Autocorrelations – values of \( L = 401, 251, 125 \)

Modified Autocorrelations –

- Waterfall Examples

Short-Time AMDF

- belief that for periodic signals of period \( P \), the difference function
  \[ d[n] = x[n] - x[n - k] \]
  - will be approximately zero for \( k = 0, \pm P, \ldots \) For realistic speech signals, \( d[n] \) will be small at \( k = P \) but not zero. Based on this reasoning, the short-time Average Magnitude Difference Function (AMDF) is defined as:
  \[ \hat{\gamma}_y[k] = \sum |x[n + m]w[m] - x[n + m - k]w[m - k]|^2 \]
  - with \( w[n] \) and \( \hat{w}[n] \) being rectangular windows. If both are the same length, then \( \hat{\gamma}_y[k] \) is similar to the short-time autocorrelation, whereas if \( \hat{w}[n] \) is longer than \( w[n] \), then \( \hat{\gamma}_y[k] \) is similar to the modified short-time autocorrelation (or covariance) function. In fact it can be shown that
  \[ \hat{\gamma}_y[k] = \beta \hat{\gamma}_y[k] \]
  - where \( \beta[k] \) varies between 0.6 and 1.0 for different segments of speech.

AMDF for Speech Segments

- Waterfall Examples

- Waterfall Examples

- Waterfall Examples

- Waterfall Examples
Summary

• Short-time parameters in the time domain:
  – short-time energy
  – short-time average magnitude
  – short-time zero crossing rate
  – short-time autocorrelation
  – modified short-time autocorrelation
  – Short-time average magnitude difference function