Automatic Speech Recognition
handout (1)
Jan - Mar 2012
Revision : 1.1

Speech Signal Processing and Feature Extraction

Hiroshi Shimodaira (h.shimodaira@ed.ac.uk)
Speech Communication

- Intention
- Language
- Motion Control
- Articulate organ (vocal tract)
- Signal source (vocal cords)

Speaker

- Understanding
- Language
- Auditory processing
- Auditory organs

Listener

speech sound
Spectrogram

Waveform

Spectrogram

Cross-section of spectrogram
Speech Production Model

Vocal Organs & Vocal Tract

Time domain: \( x(t) = h(t) \ast v(t) = \int_{0}^{\infty} h(\tau)v(t - \tau) \, d\tau \)

\[ \downarrow \] Fourier transform

Frequency domain: \( X(\Omega) = H(\Omega)V(\Omega) \)

\( \Omega \): angular frequency \( (= 2\pi F) \)

\( F \): frequency
Automatic Speech Recognition

Find the word sequence $W$ such that

$$\max_W P(W|X) = \max_W \frac{P(X|W)P(W)}{P(X)}$$

**Signal Analysis**
- LPC, FFT etc

**Speech signal**

**Feature vector sequence** $X$

**Search for**

$$\max_W p(X|W)P(W)$$

**Acoustic likelihood** $p(X|W)$

**Search constraints** $P(W)$

**Recognition results** $W^*$

**Optimal candidates**

**Acoustic model**
- HMM, NN, adaptation

**Language model**
- word n-gram, grammar...

Signal Analysis for ASR

Front-end analysis

Convert acoustic signal into a sequence of **feature vectors**

* e.g. MFCCs, PLP cepstral coefficients

\[ x_c(t) \xrightarrow{\text{LPF}} A/D \xrightarrow{\text{Pre-emphasis}} x[n] \xrightarrow{\text{Spectral analysis}} c_m[k] \]

- **LPF**: Low-pass filter
- **A/D conversion**: Analog-to-Digital conversion
- **Pre-emphasis**: Pre-emphasizing the signal
- **Spectral analysis**: Analysis of the signal spectrum
- **Feature extraction**: Extraction of features

**Sampling frequency** \( F_s \)

**Analysis window**

**Frame-shift**

\( m: \) frame number
\( k: \) feature index
Feature parameters for ASR

Features should

- contain sufficient information to distinguish phonemes / phones
  - good time-resolutions [e.g. 10ms]
  - good frequency-resolutions [e.g. 20 channels/Bark-scale]
- not contain (or be separated from) $F_0$ and its harmonics
- be robust against speaker variation
- be robust against noise / channel distortions
- have good characteristics in terms of pattern recognition
  - The number of features is as few as possible
  - Features are independent of each other
Converting analogue signals to machine readable form

- Discretisation (sampling) \( x_c(t) \rightarrow x[n] \)
  - continuous time \( \Rightarrow \) discrete time
  - continuous amplitude \( \Rightarrow \) discrete amplitude

Problem: information can be lost by sampling
Sampling of continuous-time signals

- **Continuous-time signal:** \( x_c(t) \)

- **Modulated signal by a periodic impulse train:**

\[
x_s(t) = x_c(t) \sum_{n=\infty}^{\infty} \delta(t - nT_s) = \sum_{n=\infty}^{\infty} x_c(nT_s)\delta(t - nT_s)
\]

- **Sampled signal:** \( x[n] = x_s(nT_s) \) \( \cdots \) discrete-time signal

\( T_s \): Sampling interval
Is the C/D conversion invertible?

\[ x_c(t) \xrightarrow{\text{C/D}} x[n] \xrightarrow{\text{D/C}} x_c(t)? \]
Q: Is the C/D conversion invertible?

\[ x_c(t) \xrightarrow{C/D} x[n] \xrightarrow{D/C} x_c(t) \]

A: “No” in general, but “Yes” under a special condition: “Nyquist sampling theorem”

If \( x_c(t) \) is band-limited (i.e. no frequency components \( > F_s/2 \)), then \( x_c(t) \) can be fully reconstructed by \( x[n] \).

\[
x_c(t) = h_{T_s}(t) * \sum_{k=-\infty}^{\infty} x[k] \delta(t - kT_s) = \sum_{k=-\infty}^{\infty} x[k] h_{T_s}(t - kT_s)
\]

\[
h_{T_s}(t) = \text{sinc}(t/T_s) = \frac{\sin(\pi t/T_s)}{\pi t/T_s}
\]

\( F_s/2 : \text{Nyquist Frequency} \), \( F_s = 1/T_s : \text{Sampling Frequency} \)
Sampling of continuous-time signals

Interpretation in frequency domain:

\[ X_s(\Omega) = \frac{1}{T_s} \sum_{k=-\infty}^{\infty} X_c(\Omega - k\Omega_s) \]

- Spectrum of \( x_s(t) \)
- Spectrum of \( x_c(t) \)
Questions

1. What sampling frequencies \( (F_s) \) are used for ASR?
   - microphone voice: \( 12 \text{kHz} \sim 20 \text{kHz} \)
   - telephone voice: \( \sim 8 \text{kHz} \)

2. What are the advantages / disadvantages of using higher \( F_s \)?

3. Why is pre-emphasis (+6dB/oct.) employed?
   \[
x[n] = x_0[n] - ax_0[n - 1], \quad a = 0.95 \sim 0.97
\]
Spectral analysis: Fourier Transform

- **FT for continuous-time signals (& continuous-frequency)**

\[ X_c(\Omega) = \int_{-\infty}^{\infty} x_c(t) e^{-j\Omega t} dt \quad \text{(time domain \rightarrow freq. domain)} \]

\[ x_c(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X_c(\Omega) e^{j\Omega t} d\Omega \quad \text{(freq. domain \rightarrow time domain)} \]

- **FT for discrete-time signals (& continuous-frequency)**

\[ X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x[n] e^{-j\omega n} \]

\[ x[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega}) e^{j\omega n} d\omega \]

\[
\begin{align*}
|X(e^{j\omega})|^2 & \quad \cdots \quad \text{Power spectrum} \\
\log |X(e^{j\omega})|^2 & \quad \cdots \quad \text{Log power spectrum}
\end{align*}
\]

where \( \omega = T_s \Omega = 2\pi f \),

\( e^{-j\omega n} = \cos(\omega n) + j \sin(\omega n), \quad j : \text{the imaginary unit} \)
An interpretation of FT

Inner product between two vectors (Linear Algebra)

■ 2-dimensional case
\[ a = (a_1, a_2)^t \]
\[ b = (b_1, b_2)^t \]
\[ a \cdot b = a^t b = a_1 b_1 + a_2 b_2 = \|a\| \|b\| \cos \theta \]

■ Infinite-dimensional case
\[ x \triangleq \{x[n]\}_{-\infty}^{\infty} \]
\[ e_\omega \triangleq \{e^{jwn}\}_{-\infty}^{\infty} = \{\cos(\omega n) + j \sin(\omega n)\}_{-\infty}^{\infty} \]
\[ X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x[n] e^{-jwn} = x \cdot e^{j\omega n} = x \cdot \cos \omega + jx \cdot \sin \omega \]

\( x \cdot \cos \omega \) : proportion of how much \( \cos \omega \) component is contained in \( x \)
Short-time Spectrum Analysis

Problem with FT

- Assuming signals are stationary: signal properties do not change over time
- If signals are non-stationary ⇒ loses information on time varying features

⇒ Short-time Fourier transform (STFT) (Time-dependent Fourier transform)

Divide the signal $x[n]$ into short-time segments (frames) $x_k[m]$ and apply FT to each segment.

<table>
<thead>
<tr>
<th>$x[n]$</th>
<th>$x_1[m]$, $x_2[m]$, ..., $x_k[m]$, ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X(\omega)$</td>
<td>$X_1(\omega)$, $X_2(\omega)$, ..., $X_k(\omega)$, ...</td>
</tr>
</tbody>
</table>
Short-time Spectrum Analysis (cont. 2)

- **Windowing**
- **Shift**
- **Frame**

Discrete - Fourier Transform

Short-time power spectrum

ASR (H. Shimodaira)
Short-time Spectrum Analysis (cont. 3)

- Trade-off problem of short time spectrum analysis

<table>
<thead>
<tr>
<th></th>
<th>window width</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>short → long</td>
</tr>
<tr>
<td>frequency resolution</td>
<td></td>
</tr>
<tr>
<td>time resolution</td>
<td></td>
</tr>
</tbody>
</table>

⇒ a compromise for ASR:

window width (frame width): 20 ~ 30 ms
window shift (frame shift): 5 ~ 15 ms
The Effect of Windowing in STFT

Time domain:

\[ y_k[n] = w_k[n]x[n], \quad w_k[n] : \text{time-window for } k\text{-th frame} \]

Simply cutting out a short segment (frame) from \( x[n] \) implies applying a rectangular window on to \( x[n] \).

⇒ causes discontinuities at the edges of the segment.

Instead, a tapered window is usually used. e.g. Hamming (\( \alpha = 0.46164 \)) or Hanning (\( \alpha = 0.5 \)) window)

\[ w[\ell] = (1 - \alpha) - \alpha \cos \left( \frac{2\pi \ell}{N - 1} \right) \]

\( N \) : window width
The Effect of Windowing in STFT (cont. 2)

Frequency domain:

\[ Y_k(e^{j\omega}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} W_k(e^{j\theta}) X(e^{j(\omega-\theta)}) d\theta \quad \cdots \quad \text{Periodic convolution} \]

- Power spectrum of the frame is given as a periodic convolution between the power spectra of \( x[n] \) and \( w_k[n] \).

- If we want \( Y_k(e^{j\omega}) = X(e^{j\omega}) \), the necessary and sufficient condition for this is \( W_k(e^{j\omega}) = \delta(\omega) \), i.e. \( w_k[n] = \mathcal{F}^{-1}\delta(\omega) = 1 \), which means the length of \( w_k[n] \) is infinite.
  \[ \Rightarrow \text{there is no window function of finite length that causes no distortion.} \]

NB: hereafter \( x[n] \) will be also used to denote a segmented signal for simplicity.
Spectral analysis of two sine signals of close frequencies

- **ASR (H. Shimodaira)**

---

**The Effect of Windowing in STFT**

---

ASR (H. Shimodaira)
Problems with STFT

- The estimated power spectrum contains harmonics of $F_0$, which makes it difficult to estimate the envelope of the spectrum.

- Frequency bins of STFT are highly correlated each other, i.e. power spectrum representation is highly redundant.
Cepstrum Analysis

Idea: split (deconvolve) the power spectrum into spectrum envelope and $F_0$ harmonics.

Log-spectrum [freq. domain]

\[ \downarrow \text{Inverse Fourier Transform} \]

Cepstrum [time domain] (quefrency)

\[ \downarrow \text{Liftering to get low/high part} \]
\[ \text{(lifter: filter used in cepstral domain)} \]

\[ \downarrow \text{Fourier Transform} \]

Smoothed-spectrum [freq. domain]

(low-part of cepstrum)

Log-spectrum of high-part of cepstrum
Cepstrum Analysis (cont. 2)

\[ x[n] = h[n] \ast v[n] \]

\[ h[n] : \text{vocal tract} \]
\[ v[n] : \text{glottal sounds} \]

\[ \downarrow \mathcal{F} \quad (\text{Fourier transform}) \]

\[ X(e^{j\omega}) = H(e^{j\omega})V(e^{j\omega}) \]

Log spectrum

\[ \downarrow \log \]

\[ \log |X(e^{j\omega})| = \underbrace{\log |H(e^{j\omega})|} + \underbrace{\log |V(e^{j\omega})|} \]

(spectral envelope) (spectral fine structure)

Cepstrum

\[ \downarrow \mathcal{F}^{-1} \]

\[ c(\tau) = \mathcal{F}^{-1} \left\{ \log |X(e^{j\omega})| \right\} \]

\[ = \mathcal{F}^{-1} \left\{ \log |H(e^{j\omega})| \right\} + \mathcal{F}^{-1} \left\{ \log |V(e^{j\omega})| \right\} \]
LPC Analysis

Linear Predictive Coding (LPC):
  a model-based / parametric spectrum estimation

Assume a “linear system” for human speech production

- sound source \( v[n] \) \( \Rightarrow \) vocal tract \( \Rightarrow \) speech \( x[n] \)

\[
v[n] \rightarrow [h[n]] \rightarrow x[n] \quad h[n] : \text{impulse response}
\]

\[
x[n] = h[n] * v[n] = \sum_{k=0}^{\infty} h[k] v[n - k]
\]

Using a model enables us to
- estimate a spectrum of vocal tract from small amount of observations
- represent the spectrum with a small number of parameters
- synthesise speech with the parameters
Predict \( x[n] \) from \( x[n-1], x[n-2], \cdots \)

\[
\hat{x}[n] = \sum_{k=1}^{N} a_k x[n-k]
\]

\[
e[n] = x[n] - \hat{x}[n] = x[n] - \sum_{k=1}^{N} a_k x[n-k] \quad \cdots \text{prediction error}
\]

**Optimisation problem**

Find \( \{a_k\} \) that minimises the mean square (MS) error:

\[
P_e = E \{e^2[n]\} = E \left\{ \left( x[n] - \sum_{k=1}^{N} a_k x[n-k] \right)^2 \right\}
\]

\( \{a_k\} : \) LPC coefficients
Spectrums estimated by FT & LPC

ASR (H. Shimodaira)
LPC summary

- Spectrum can be modelled/coded with around $14 \text{LPCs}$. 

- LPC family
  - PARCOR (Partial Auto-Correlation Coefficient)
  - LSP (Line Spectral Pairs) / LSF (Line Spectrum Frequencies)
  - CSM (Composite Sinusoidal Model)

- LPC can be used to predict log-area ratio coefficients lossless tube model

- LPC-(Mel)Cepstrum: LPC based cepstrum.

- Drawback:
  - LPC assumes AR model which does not suit to model nasal sounds that have zeros in spectrum.
  - Difficult to determine the prediction order $N$. 
## Taking into Perceptual Attributes

<table>
<thead>
<tr>
<th>Physical quality</th>
<th>Perceptual quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intensity</td>
<td>Loudness</td>
</tr>
<tr>
<td>Fundamental frequency</td>
<td>Pitch</td>
</tr>
<tr>
<td>Spectral shape</td>
<td>Timbre</td>
</tr>
<tr>
<td>Onset/offset time</td>
<td>Timing</td>
</tr>
<tr>
<td>Phase difference in binaural hearing</td>
<td>Location</td>
</tr>
</tbody>
</table>

### Technical terms
- equal-loudness contours
- masking
- auditory filters (critical-band filters)
- critical bandwidth
Taking into Perceptual Attributes (cont. 2)
Non-linear frequency scale

- **Bark scale**
  \[ b(f) = 13 \arctan(0.00076f) + 3.5 \arctan\left(\frac{f}{7500}\right)^2 \]  
  [Bark]

- **Mel scale**
  \[ B(f) = 1127 \ln(1 + \frac{f}{700}) \]
Filter Bank Analysis

Speech $x[n]$ is filtered by Bandpass Filter 1 and Bandpass Filter K, resulting in $x_1[n]$ and $x_K[n]$ respectively.

The impulse response of Bandpass filter $i$ is given by:

$$x_i[n] = h_i[n] * x[n] = \sum_{k=0}^{M_i-1} h_i[k] x[n-k]$$

$h_i[n]$: Impulse response of Bandpass filter $i$
Filter Bank Analysis (cont. 2)

Speech

\[ x[n] \]

\[ x_i[n] \]

\[ x_K[n] \]

\[ v[n] \]

\[ v_i[n] \]

\[ v_K[n] \]

\[ y[n] \]

\[ y_i[n] \]

\[ y_K[n] \]

Frequency resolution

number of filters

length of filter

Time resolution

Trade-off problem

ASR (H. Shimodaira)
Another implementation of filter banks:
apply a mel-scale filter bank to STFT power spectrum to obtain mel-scale power spectrum
MFCC: Mel-frequency Cepstral Coefficients  $c[n]$

$$x[n] \xrightarrow{\text{DFT}} X[k] \rightarrow |X[k]|^2 \xrightarrow{\text{Mel-frequency filterbank}} \log |S[m]| \xrightarrow{\text{DCT}} c[n]$$

DCT: $c[n] = \sqrt{\frac{2}{N}} \sum_{i=1}^{N} s[i] \cos \left( \frac{\pi n(i - 0.5)}{N} \right)$, where $s[i] = \log |S[i]|$

DFT: discrete Fourier transform, DCT: discrete cosine transform

- MFCCs are widely used in HMM-based ASR systems.
- The first 12 MFCCs ($c[1] \sim c[12]$) are generally used.
MFCCs are less correlated each other than DCT/Filter-bank based spectrum.

Good compression rate.

<table>
<thead>
<tr>
<th>Feature</th>
<th>dimensionality / frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech wave</td>
<td>400</td>
</tr>
<tr>
<td>DCT Spectrum</td>
<td>64 ~ 256</td>
</tr>
<tr>
<td>Filter-bank</td>
<td>10 ~ 20</td>
</tr>
<tr>
<td>MFCC</td>
<td>12</td>
</tr>
</tbody>
</table>

where $F_s = 16 kHz$, frame-width = 25 ms, frame-shift = 10 ms are assumed.

MFCCs show better ASR performance than filter-bank features, but MFCCs are not robust against noises.
Perceptually-based Linear Prediction (PLP)

[Perceptually-based Linear Prediction (PLP)]

PLP had been shown experimentally to be
- more noise robust
- more speaker independent
than MFCCs
Other features with low dimensionality

- Formants ($F_1, F_2, F_3, \cdots$)

They are not used in modern ASR systems, but why?
Using temporal features: **dynamic features**

In SP lab-sessions on speech recognition using HTK,
- MFCCs, and energy
- \( \Delta \) MFCCs, \( \Delta \) energy
- \( \Delta^2 \) MFCCs, \( \Delta^2 \) energy

\( \Rightarrow \Delta^* , \Delta^{2*} : \) **delta features**

(dynamic features / time derivatives) [Furui, 1986]

<table>
<thead>
<tr>
<th>Continuous time</th>
<th>Discrete time</th>
</tr>
</thead>
<tbody>
<tr>
<td>( c(t) )</td>
<td>( c[n] )</td>
</tr>
<tr>
<td>( c'(t) = \frac{dc(t)}{dt} )</td>
<td>( \Delta c[n] )</td>
</tr>
<tr>
<td>( c''(t) = \frac{d^2c(t)}{dt^2} )</td>
<td>( \Delta^2 c[n] )</td>
</tr>
</tbody>
</table>

\[ c_0(t) = \frac{dc(t)}{dt} \]
\[ c[n] = c[n+1] - c[n-1] \]

\[ \sum_{i=-M}^{M} w_i c[n + i] \]

**e.g.** \( \Delta c[n] = \frac{c[n+1] - c[n-1]}{2} \)
Using temporal features: **dynamic features** (cont. 2)

\[ c(t) \]

\[ c'(t_0) \]

\[ t_0 \]
An acoustic feature vector, eg MFCCs, representing part of a speech signal is highly correlated with its neighbours.

HMM based acoustic models assume there is no dependency between the observations.

Those correlations can be captured to some extent by augmenting the original set of static acoustic features, eg. MFCCs, with dynamic features.
General Feature Transformation

- Orthogonal transformation (orthogonal bases)
  - DCT (discrete cosine transform)
  - PCA (principal component analysis)

- Transformation based on the bases that maximises the separability between classes.
  - LDA (linear discriminant analysis) / Fisher’s linear discriminant
  - HLDA (heteroscedastic linear discriminant analysis)
A comparison of speech features


<table>
<thead>
<tr>
<th>Feature</th>
<th>WER(%)</th>
<th>SER(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SBC (16)</td>
<td>6.2</td>
<td>21.3</td>
</tr>
<tr>
<td>WPSR125 (16)</td>
<td>6.3</td>
<td>21.8</td>
</tr>
<tr>
<td>OWPF (16)</td>
<td>6.4</td>
<td>22.1</td>
</tr>
<tr>
<td>LFCC-FB40</td>
<td>6.9</td>
<td>23.5</td>
</tr>
<tr>
<td>HFCC-FB23</td>
<td>8.2</td>
<td>27.3</td>
</tr>
<tr>
<td>HFCC-FB40</td>
<td>8.7</td>
<td>28.2</td>
</tr>
<tr>
<td>PLP-FB19</td>
<td>9.0</td>
<td>29.4</td>
</tr>
<tr>
<td>MFCC-FB40</td>
<td>9.0</td>
<td>29.9</td>
</tr>
</tbody>
</table>

NB The above result was obtained for TIMIT speech corpus. Results might change a lot under different conditions (e.g. noise, tasks, ASR systems)
Further topics on feature extraction

- Feature normalisation/enhancement in terms of
  - noise / environments
  - speakers / speaking styles
  - speech recognition

- Pitch ($F_0$) adapted feature extraction
SUMMARY

- Nyquist Sampling theory
- Short-time Spectrum Analysis
  - Non-parametric method
    - Short-time Fourier Transform
    - Cepstrum, MFCC
    - Filter bank
  - Parametric methods
    - LPC, PLP
  - Windowing effect: trade-off between time and frequency resolutions
- Dynamic features (delta features)
- There is no best feature that can be used for any purposes, but MFCC is widely used for ASR and TTS.
Front-end analysis has a great influence on ASR performance.

For robust ASR in real environments, various techniques for front-end processing have been proposed. e.g. spectral subtraction (SS), cepstral mean normalisation (CMN)

Spectrum analysis and feature extraction involve information loss and non-linear distortions. There is always a trade-off between accuracy and efficiency. (e.g. spatial resolution vs. temporal resolution)
References


- http://mi.eng.cam.ac.uk/comp.speech/

- http://mi.eng.cam.ac.uk/~ajr/SpeechAnalysis/

- http://cslu.cse.ogi.edu/HLTsurvey/


- “Spoken language processing: a guide to theory, algorithm, and system development”, Xuedong Huang, Alex Acero and Hsiao-Wuen Hon, Prentice Hall (2001). isbn: 0130226165