6. Feature Extraction from Speech and other kinds of Audio
Overview

• Learn about the most established feature extraction from speech

• Mel Frequency Cepstral Coefficients: MFCC
Quantization

• Uniform quantization:
  – 10-12 bit are sufficient to code speech

• Improvement:
  – Use distribution of amplitude values
  – $\mu$-law:

  $$f_n^{(\mu)} = f_{\text{max}} \text{sgn}(f_n) \frac{\log(1 + \mu |f_n|)}{\log(1 + \mu)} \quad \mu \approx 200$$

  $$\propto \log(1 + \mu' |f_n|)$$
Features in the Time Domain: Short-time Energy

Definition: \[ E^{(n)} = \sum_{m=0}^{M-1} | f_{m+n} |^2 \]

Example:

From: Schukat-Talamazzini
Pre-emphasis

- Correct for filtering of the lips
- Iterative scheme:
  \[ f_n' = f_n + \alpha f_{n-1} \]
- Typical values: \( \alpha = 0.95 \)
From Signal to Spectrum: Fourier Transform

• Definition

\[ F^{(m)}(e^{i\omega}) = \sum_{n} f_n w_{m-n} e^{-i\omega n} \]

\( w_n \): window function
\( \omega \): frequency times 2\( \pi \)
Example: putting a rectangular on a speech signal

Frame shift typ.: 10ms
Frame width typ.: 25ms
A Simple Example for Fourier Transform

a Maple script
Fourier Transform in Practice

- Use “Fast Fourier Transform” (FFT)
- Requires number of samples N to be power of 2 (e.g. N=256)
- Code available
- Complexity \( N \log(N) \)
Established Window Functions

- Use to get sharper peaks
- Rectangular window: \( w_n^R = 1 \)
- Generalized Hamming Window:
  \[
  w_n^H = (1 - \alpha) \cos\left(\frac{2\pi n}{N-1}\right)
  \]
  \( (\alpha=0.46 : \text{standard Hamming window}) \)
- Gauss window:
  \[
  w_n^G = e^{-\left(\frac{n-N/2}{3N/2}\right)^2}
  \]
- Parabola window:
  \[
  w_n^P = 4 \left(1 - \frac{n}{N}\right)
  \]
  \( n=0\ldots N-1 \)
- Window functions vanish outside this interval
Rewrite of Fourier Transform

• Definition:

\[ F^{(m)}(e^{i\omega}) = \sum_{n=-\infty}^{\infty} f_n w_{m-n} e^{-i\omega n} \]

• Window functions vanish outside the interval \(n=0...N-1\)

• Define

\[ \omega = \frac{2\pi v}{N} \]

\[ F_v^{(m)} = \sum_{n=0}^{N-1} f_m w_n e^{-i\frac{2\pi v n}{N}} \]

Note: for further processing, we take the absolute value of the Fourier Transform
Example for ö

Short time spectrum

Smoothed spectrum

Frequency (Hz)
Spectrogram

- Calculate a spectrum for any point in time
- Code the local intensity: color/grey scale
"To return to the main menu, press the star key".
Use praat to generate a Spectrogram

• **Praat**: software for doing phonetics by computer
• **Written by**: Paul Boersma and David Weenink
• quite powerful: spectrograms, formants, pitch, …
• **Download**: [http://www.fon.hum.uva.nl/praat/](http://www.fon.hum.uva.nl/praat/)
Use praat to generate a Spectrogram
Smoothing the Spectrogram: Filterbank

- Idea: imitate ear
  - Do an average over neighboring frequencies
  - Scale the frequencies according to the mel or the Bark scale
- Reduction from 256 Fourier coefficients to 24 outputs of a filterbank
Example of a Filterbank

MELSPEC

Energy in Each Band

freq

m_1 \quad ... \quad m_j \quad ... \quad m_p
Filterbank

• Spacing of center frequency:
  – According to mel scale:
    \[ Mel(f) = 259 \log_{10}(1 + \frac{f}{700}) \]

• Low frequency cut off:
  – E.g. 300 Hz (for telephone speech)

• High frequency cut off:
  – E.g. 3400 Hz (for telephone speech)

• Different settings for e.g. head set connected PC
Vocal Tract Length Normalization

- **Idea:**
  - Average position of formants depends on length of vocal tract
  - A varying position of frequencies of filter bank
  - A kind of speaker adaptation
Vocal Tract Length Normalization: Frequency Warping
Learning the Warp Factor $\alpha$

• Issue: how to scale for a specific speaker
• Slow version:
  • Use 11 different warping factors
  • Do speech recognition with all of them
  • Pick the best one
• Oldest approach
• Not very efficient
• Improvement: 10% less recognition errors
How does a telephone change your voice?
From Spectrum to Cepstrum

• Name: swapping of letters
• Idea: separate out the convolutional contribution
• Useful as a preparation to remove channel distortions (e.g. telephone)
• Cepstral mean subtraction (CMS)
Definition “Cepstrum”

Signal

Spectrum

Cepstrum

Fourier Transform

log

Discrete Cosine Transform
Math for Cepstrum

- en: original signal (e.g. excitation from glotis)
- fn: measured signal
- hn: impulse response of channel (e.g. vocal tract)

\[ f_n = \sum_{n=-\infty}^{\infty} h_{m-n} e_n \]
Math for Cepstrum

• Apply Fourier transform \( F \)

\[
F \{ f_n \} = \mathcal{F} \left\{ \sum_{n=\infty}^{\infty} h_{m-n} e_n \right\}
\]

• Use convolution theorem

\[
F \{ f_n \} = \mathcal{F} \{ h_n \} F \{ e_n \}
\]
Math for Cepstrum

- Apply logarithm

\[
\log(F \{ f_n \}) = \log(\mathcal{F}\{ h_n \}) + \log(F \{ e_n \})
\]

- Impulse response and excitation now separated
Cepstrum: do discrete cosine transform after log

• Discrete cosine transform:

\[
c_0^{(m)} = \sqrt{2/N} \sum_{\nu=0}^{N/2-1} \log(F_{\nu}^{(m)})
\]

\[
c_q^{(m)} = \sqrt{4/N} \sum_{\nu=0}^{N/2-1} \log(F_{\nu}^{(m)}) \cos\left(\frac{\pi q (2\nu + 1)}{N}\right)
\]
Dynamic Features

- Cepstrum captures local aspects of speech
- Window size 25 ms
- Capture slow changes in spectrum
- Other name: delta features
Dynamic Features

• Capture slow changes in spectrum
Dynamic Features

- Calculate first and second derivatives
- Naïve approach to first derivative
  - Continuous function
    \[
    \frac{df(t)}{dt} \approx \frac{f(t + \Delta t) - f(t - \Delta t)}{2\Delta t}
    \]
  - Time discrete sampling
    \[
    \frac{df(t_m)}{dt} \approx \frac{f(t_m + \Delta) - f(t_m - \Delta)}{2\Delta + 1}
    \]
Difference/Regression

i-th component of feature vector

Line through extremes

Regression curve

Sample
Regression Formula

\[
\frac{df(t)}{dt} = \frac{\sum_{i=1}^{M} i(f(t_{m+i}) - f(t_{m-i}))}{\sum_{i=1}^{M} i^2}
\]

• Check M=1

How could you derive this formula?
Dynamic Features

• Invented by Furui 1981
• Standard in any modern ASR system

• Alternative:
  • Linear mapping of neighboring feature vectors

• Issue:
  • Dimension of feature vectors
Linear Discriminant Analysis

- Method to decrease size of feature vector
- Maximize separability of class regions
- Linear transform of feature vectors
- More: later in the lecture
Complete Pipeline for Mel-Frequency Cepstral Coefficients (MFCC)

Signal
- Sampling
- Pre-emphasis
- Windowing
- Fast Fourier Transform
- Absolute Value
- Mel-scaled Filterbank
- log
- Discrete Cosine Transform
- Dynamic Features (1. and 2. derivative)
- Linear Discriminant Analysis

Feature Vectors

Typical values:
- 16 kHz; 16 Bit quantization
- Window size: 25 ms
- 512 Fourier Coefficients
- 24 filterbank values
- keep only 20 lowest cepstra
- 60 dimensional vector
Alternative Feature Extraction Methods

• LP-Cepstrum (LP=linear prediction)
  • Derived from speech coding
  • No longer much in use

• PLP (=Perceptual linear prediction)
  • For certain applications popular
  • Claim: more noise robust than MFCCs
  • Main change: us |.|1/3 instead of log in MFCC
Summary

• Classical “plain vanilla” feature extraction: Mel-Frequency Cepstral Coefficients
• Main deficiency: not very noise robust
• Used in
  • Speech Recognition
  • Speaker Recognition
  • Music genre classification