Apprentissage statistique et applications

Audio Signal Processing

TP2

Outline

Outline

1. Introduction
2. Extraction of features - MFCC

Speech Recognition

- What is Speech Recognition ?

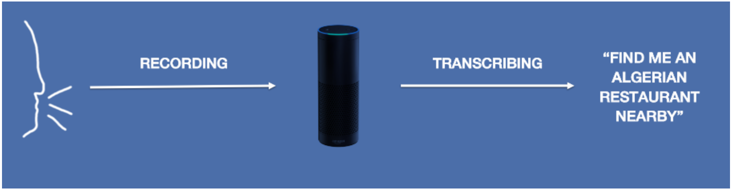

FIGURE – Speech Recognition

Speech Recognition

- From transcribing to understanding humour ?

FIGURE – Messing with Siri

Anatomy of a speech recognition system

- The anatomy of a speech recognition system :
  - Speech Features extraction
  - Acoustic Model
  - Language Modelling

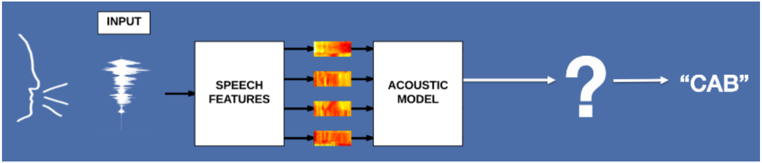

FIGURE – Steps of Speech Recognition

THE WAVE SIGNAL

- What does a microphone record ?
  - A microphone measures variations in air pressure.
  - It collects a discretized signal.
  - Can record at different samplerates (8kHz, 16kHz, etc.)

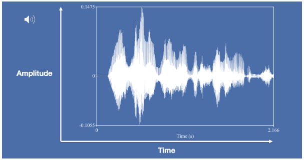

FIGURE – The wave signal

THE WAVE FORM

- The wave signal 3 seconds $*$ 16000 Hz $=$ vector of length 48000

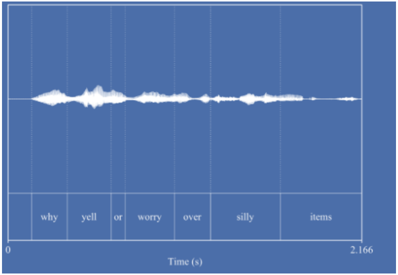

FIGURE – the wave signal

Desired properties of speech features

- The wave signal is a highly non stationary signal.
  $\rightarrow$ We need **local** features.
- Phonemes are characterized by their spectral signature.
  $\rightarrow$ We need **spectral** features
- The speech signal is high dimensional.
  $\rightarrow$ We need **compact** features
- Two types of speech features have these properties :
  mel-filterbanks and **MFCC**

The MFCC (Mel-frequency Cepstral Coefficients) Pipeline

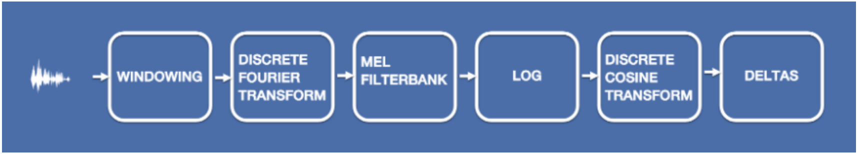

FIGURE – MFCC pipeline

Outline

Introduction

The MFCCs can be viewed as a sinusoidal decomposition of the Mel spectrum and allows representing its global shape with only a few coefficients (usually 12-13 coefficients). We will detail the steps of calculating MFCCs.

- First, a signal goes through a pre-emphasis filter.
- Then, it gets sliced into (overlapping) frames and a window function is applied to each frame.
- Afterwards, we do a Fourier transform on each frame (or more specifically a Short-Time Fourier Transform) and calculate the power spectrum, and subsequently compute the filter banks.
- To obtain MFCCs, a Discrete Cosine Transform (DCT) is applied to the filter banks retaining a number of the resulting coefficients while the rest are discarded.
- A final step in both cases, is mean normalization.

PRE-EMPHASIS

- The first step is to apply a **pre-emphasis filter** on the signal to amplify the high frequencies.
- The pre-emphasis filter can be applied to a signal x using the first order filter in the following equation :

$$y(t) = x(t) - \alpha x(t-1)$$

- Typical values for α are 0.96 - 0.97

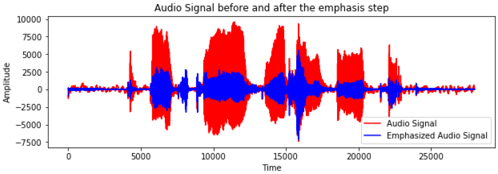

FIGURE – Preemphasis step

FRAMING

- After pre-emphasis, we need to split the signal into short-time frames.
- Typical frame sizes in speech processing range from 20 ms to 40 ms with $50\%(+/-10\%)$ overlap between consecutive frames. Popular settings are 25 ms for the frame size,and a 10 ms stride (15 ms overlap).

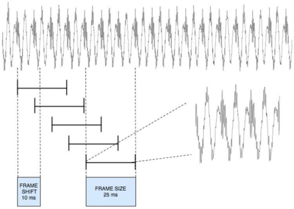

FIGURE – Framing

WINDOWING

- There are several reasons why we need to apply a window function to the frames, notably to counteract the assumption made by the FFT that the data is infinite and to reduce spectral leakage.
- After slicing the signal into frames, we apply a window function such as the Hamming window to each frame.
- A Hamming window has the following form :

$$w[n] = 0.54 - 0.46\cos(\frac{2\pi n}{N-1})$$

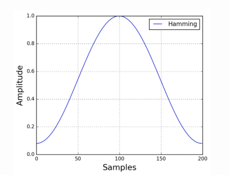

FIGURE – Hamming Window

WINDOWING

- The Hamming window shrinks the values of the signal toward zero at the window boundaries, avoiding discontinuities.

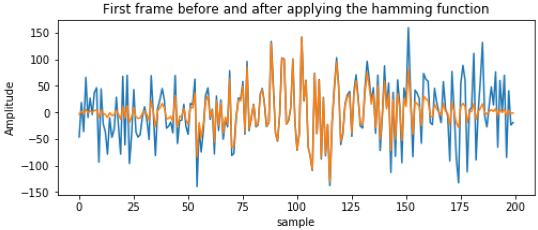

FIGURE – Hamming applied to the first frame of Wave Signal

Fourier Transform and Power Spectrum

- Let's call our time domain signal $s(n)$.
- Once it is framed we have $s_i(n)$ where n ranges over $\{1, \ldots, N\}$ and i ranges over the number of frames.
- When we calculate the complex DFT, we get $S_i(k)$ where the i denotes the frame number corresponding to the time-domain frame.
- To take the Discrete Fourier Transform of the frame, perform the following :

$$S_i(k) = \sum_{n=1}^{N} s_i(n)w(n)e^{-j\frac{2\pi kn}{N}} \quad 1 \leq k \leq K \quad \text{K is the length of the DFT}$$

FOURIER TRANSFORM AND POWER SPECTRUM

- The periodogram-based power spectral estimate for the speech frame $s_i(n)$ is given by :

$$P_i(k) = \frac{1}{N}|S_i(k)|^2$$

- This is called the **Periodogram estimate of the power spectrum**.
- We would generally perform a 512 point FFT and keep only the first 257 coefficents.

Filter banks

- The filter banks step consists in applying triangular filters, typically 40 filters on a Mel-scale to the power spectrum to extract frequency bands.
- The Mel-scale aims to mimic the non-linear human ear perception of sound, by being more discriminative at lower frequencies and less discriminative at higher frequencies
- We can convert between Hertz (f) and Mel (m) using the following equations :

$$m = 2595 \log_{10}(1 + \frac{f}{700}) \quad \text{and} \quad f = 700(10^{m/2595} - 1)$$

FILTER BANKS

- Each filter in the filter bank is triangular having a response of 1 at the center frequency and decrease linearly towards 0 till it reaches the center frequencies of the two adjacent filters where the response is 0, as shown in this figure :

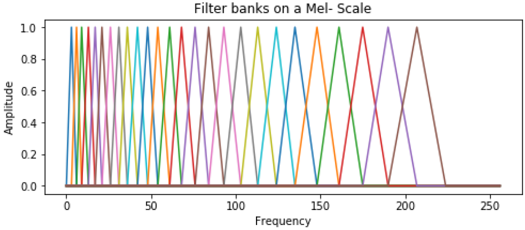

FIGURE – Filter bank on a Mel-Scale

FILTER BANKS

- After applying the filter bank to the power spectrum (periodogram) of the signal, we obtain the following spectrogram :

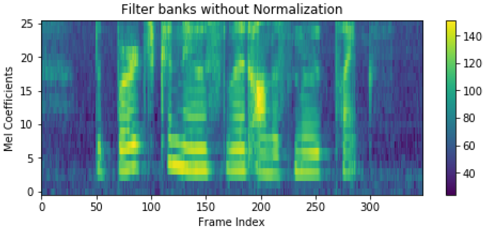

FIGURE – Mel Coefficients

MFCCs

- It turns out that filter bank coefficients computed in the previous step are highly correlated, which could be problematic in some machine learning algorithms.
- Therefore, we can apply **Discrete Cosine Transform** (DCT) to decorrelate the filter bank coefficients and yield a compressed representation of the filter banks.
- The resulting MFCCs :

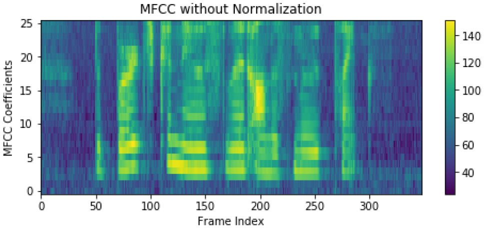

FIGURE – MFCC

Thanks for your attention