1 Introduction

Speech is probably the most crucial tool for communication in our daily lives. Therefore constructing a speech recognition system is desirable at all times.

Basically, speech recognition is the process of converting an acoustic signal to a set of words. The recognized words can be the final results, as for applications such as commands and control, data entry, and document preparation. They can also serve as the input to further processing in order to achieve speech understanding.

Feature extraction refers to procedure of transforming the speech signal into a number of parameters, while pattern matching is a task of obtaining parameter sets from memory which closely matches the parameter set extracted from the input speech signal. In simple words, the speech recognizer is to provide a powerful and accurate mechanism to transcribe speech into text.

2 Mel Frequency Cepstral Coefficients (MFCC)

Feature extraction is a crucial step of the speech recognition process. The best presented algorithm in feature extraction is Mel Frequency Cepstral Coefficients (MFCC) and the perceptual linear predictive (PLP) feature. Between them MFCC features are, the more commonly used, most popular, and robust technique for feature extraction in currently available speech recognition systems especially in clean speech or clean environment.

On the other hand the overall performance of MFCC features is not a superior in noisy environment. In real world applications the performance of MFCC degrades rapidly because of the noise. since 1980, notable efforts have been carried out to enhance MFCC feature in noisy environments.

MFCC Feature Extraction

The first stage of speech recognition is to compress a speech signal into streams of acoustic feature vectors, referred to as speech feature vectors. The extracted vectors are assumed to have sufficient information and to be compact enough for efficient recognition.

Feature extraction is actually divided into two parts: first is transforming the speech signal into feature vectors; secondly is to choose the useful features which are insensitive to changes of environmental conditions and speech variation. However, changes of environmental conditions and speech variations are crucial in speech recognition systems where accuracy has degraded massively in the case of their existence.

Some examples of speech variations include accent differences, and male-female vocal tract difference. For developing robust speech recognition, speech features are required to be insensitive to those changes and variations.

The most commonly used speech feature is definitely the Mel Frequency Cepstral Coefficients (MFCC) features, which is the most popular, and robust due to its accurate estimate of the speech parameters and efficient computational model of speech. Moreover, MFCC feature vectors are usually a 39 dimensional vector, composing of 13 standard features, and their first and second derivatives. The procedure of this MFCC feature extraction is shown below.

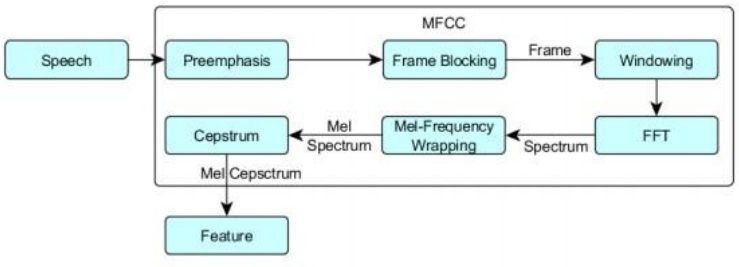


Figure 1: MFCC Processing

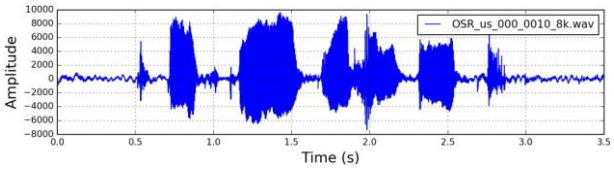


Figure 2: The raw signal has the above form in the time domain:

Pre-Emphasis

The first step is to apply a pre-emphasis filter on the signal to amplify the high frequencies. A pre-emphasis filter is useful in several ways:

(1) balance the frequency spectrum since high frequencies usually have smaller magnitudes compared to lower frequencies,

(2) avoid numerical problems during the Fourier transform operation and

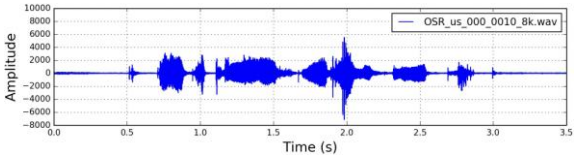
(3) may also improve the Signal-to-Noise Ratio (SNR).

The pre-emphasis filter can be applied to a signal x using the first order filter in the following equation:

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

Where typical values for the filter coefficient (α) are 0.95 or 0.97.

The signal after pre-emphasis has the following form in the time domain:



Framing

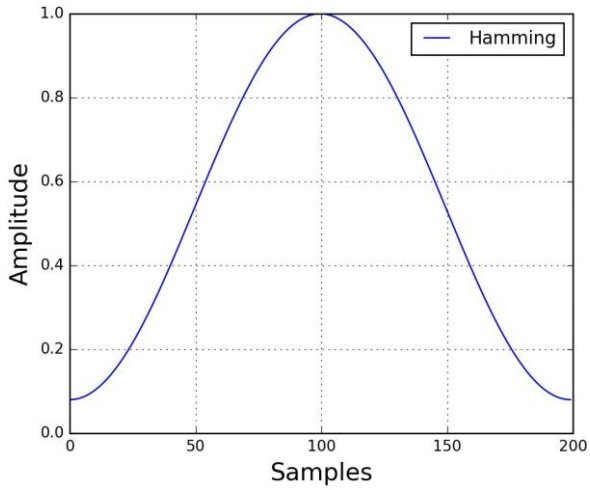
The speech signal is normally divided into small duration blocks, called frames, and the spectral analysis is carried out on these frames. This is due to the fact that the human speech signal is slowly time varying. The very popular frame length and frame shift for the speech recognition task are 20-30 ms and 10 ms respectively.

Windowing

After framing, each frame is multiplied by a window function prior to reduce the effect of discontinuity introduced by the framing process by attenuating the values of the samples at the beginning and end of each frame. The Hamming window is commonly used, it decreases the frequency resolution of the spectral analysis. A Hamming window has the following form:

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

where, 0≤n≤N-1, N is the window length. plotting the equation yields the following plot:



Spectral estimation

spectral estimation is computed for each frame by applying Discrete Fourier Transform (DFT) to produce spectral coefficients. These coefficients are complex numbers comprising the two magnitude and phase information. Phase information is usually removed and only the magnitude of the spectral coefficients are extracted. Additionally, it is common to utilize the power of the spectral coefficients.

$$X(k) = \sum_{n=0}^{N-1} y(n) \exp^{\frac{-j2\pi}{N}}, 0 \leq n, k \geq N-1$$

Where X(k) are the spectral coefficients, and y(n) the framed speech signal.

Fast Fourier Transform (FFT)

A function with limited period can be expressed in Fourier series. Fourier transform is used to convert a time series of bounded time domain signals into a frequency spectrum. The frame that has undergone the windowing process is converted into a frequency spectrum. Fourier transformation is a fast algorithm to apply Discrete Fourier Transform (DFT), on the given set of N samples shown below:

Basically the definition for FFT and DFT is same, which means that the output for the transformation will be the same; however they differ in their computational complexity.

In case of DFT, each frame with N-M samples directly will be used as a sequence for Fourier transformation. On another, in case of FFT this frame will be divided into small DFT's and then computation will be done on this divided small DFT's as individual sequence thus the computation will be more fast and easy. Thus it is in digital processing or other area instead of directly using DFT, FFT is used for applying DFT.

$$D_k = \sum_{m=0}^{N_m-1} D_m \exp^{\frac{-j\pi km}{N_m}}$$

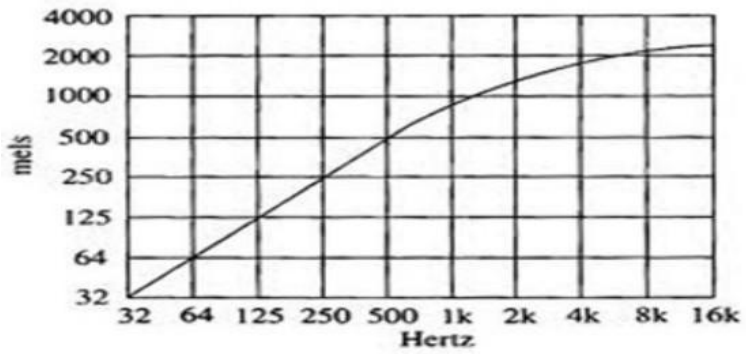
Where n = 0, 1, 2, . . . , N-1 and j = sqrt-1. X[n] is the n-frequency pattern generated from the Fourier transform, W_k. is the signal of a frame. The result of this stage is usually called Spectrum.

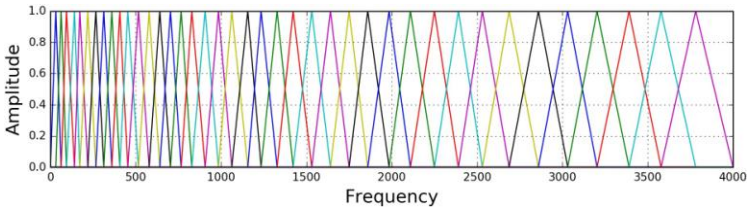
Mel scale

In this step, the above calculated spectrums are mapped on Mel scale to know the approximation about the existing energy at each spot with the help of Triangular overlapping window also known as triangular filter bank. These filter bank is a set of band pass filters having spacing along with bandwidth decided by steady Mel frequency time. Thus, Mel scale helps how to space the given filter and to calculate how much wider it should be because, as the frequency gets higher these filters are also get wider.

For Mel- scaling mapping is need to done among the given real frequency scales (Hz) and the perceived frequency scale (Mels). During the mapping, when a given frequency value is up to 1000Hz the Mel-frequency scaling is linear frequency spacing, but after 1000Hz the spacing is logarithmic as shown in the figure. The formula to convert frequency f hertz into Mel mf is given by

$$m_f = 2595 log_{10}(\frac{f}{700} + 1)$$





The above figure shows the Filter bank on a Mel-Scale. Thus, with the help of Filter bank with proper spacing done by Mel scaling it becomes easy to get the estimation about the energies at each spot and once this energies are estimated then the log of these energies also known as Mel spectrum can be used for calculating first 13 coefficients using DCT. Since, the increasing numbers of coefficients represent faster change in the estimated energies and thus have less information to be used for classifying the given images. Hence, first 13 coefficients are calculated using DCT and higher are discarded.

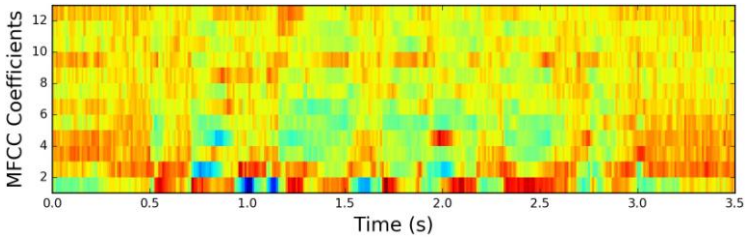
Cepstrum

Humans listen to voice information based on time domain signals. At this stage Mel-spectrum will be converted into time domain by using Discrete Cosine Transform (DCT). The result is called Mel-frequency cepstrum coefficient (MFCC).

Discrete cosine Transform (DCT)

This process of carrying out DCT is done in order to convert the log Mel spectrum back into the spatial domain. For this transformation either DFT or DCT both can be used for calculating Coefficients from the given log Mel spectrum as they divide a given sequence of finite length data into discrete vector.

However, DFT is generally used for spectral analysis where as DCT used for data compression as DCT signals have more information concentrated in a small number of coefficients and hence, it is easy and requires less storage to represent Mel spectrum in a relative small number of coefficients. This instead of using DFT DCT is desirable for the coefficients calculation as DCT outputs can contain important amounts of energy. The output after applying DCT is known as MFCC (Mel Frequency Cepstrum Coefficient).



$$C_n = \sum_{k-1}^{k} (logD_k) \cos[m(k - \tfrac{1}{2})\tfrac{\pi}{k}]$$

where m = 0, 1. . . k- 1, where C_n represents the MFCC and m is the number of the coefficients here m=13 so, total number of coefficients extracted from each frame is 13.

Log energy calculation

The energy of the speech frame is additionally computed from the time- domain signal of a frame as a feature along with the normal MFCC features. In some cases, it is replaced by C0, the 0th component of the MFCC feature, which is the sum of the log Mel filterbank coefficients.

Derivatives and accelerations calculation

The time derivatives (the first delta) and accelerations (second delta) are used to restore the information of the speech signals that have been lost in the frame-by-frame analysis. The derivative of coefficient x(n) can be calculated as

$$x(n) = \dot{x}(n) \equiv \frac{d}{dn}x(n) \approx \sum_{m=-M}^{M} m(n+m)$$

To produce the second order derivative, the same formula can be applied to the first order derivative. The final feature vectors are formed simply by adding the derived features to the original cepstral features.