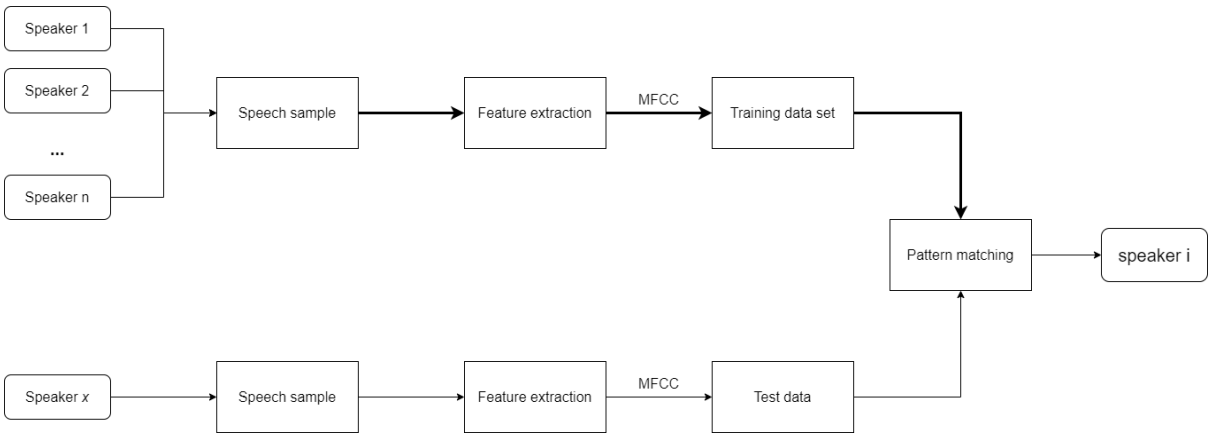
Speech Recognition

Introduction

Voices are unique because of the actual shape and size of an individual's vocal cords. They vary person to person. Not only that it is also due to the size and shape of the rest of that person's body, especially the vocal tract, and the manner in which the speech sounds are habitually formed and articulated. Speech is the most natural and efficient way of communication between humans. Lots of efforts have been made to develop a human computer interface so that one can easily interact and communicate in an unskilled way.

Speech recognition systems find their applications in our daily lives and have huge benefits for those who are suffering from some kind of disabilities. We can also use these speech-based devices for security measures to reduce cases of fraud and theft. One such speech recognition system is the main aim of this mini project which includes Speaker identification and speaker verification.

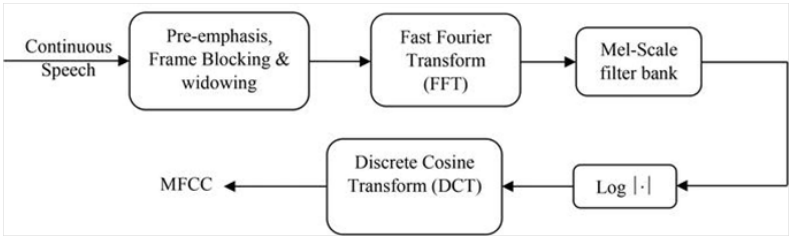
Speech recognition mainly focuses on training the system to recognize an individual's unique voice characteristics. The most popular feature extraction technique is the Mel Frequency Cepstral Coefficients called MFCC which is less complex in implementation and more effective and robust under various conditions. We can also use other feature extraction techniques such as pitch, RASTA filtering, LPC Linear predictive model etc. But what makes special about MFCC is that it is able to mimic the human auditory system i.e., by being more discriminative at lower frequencies and less discriminative at higher frequencies. This will be done by Mel scale.



This is a closed-set speaker identification: the audio of the speaker under test is compared against all the available speaker models (a finite set) and the closest match is returned. The figure above shows basic model of how speech recognition is done. First, we extract features such as the mentioned before, here we only do MFCC, and then storing the individual's feature in a model for training and then these training samples are tested. Here the coding part is entirely done on MATLAB.

Feature Extraction

The most commonly used acoustic features are mel-scale frequency cepstral coefficients (MFCC). These are the steps involved in MFCC feature extraction:



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. One use is to balance the frequency spectrum since high frequencies usually have smaller magnitudes compared to lower frequencies. The pre-emphasis filter can be applied to a signal x using the first order filter in the following equation:

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

Frame Blocking

After pre-emphasis, we need to split the signal into short-time frames. The rationale behind this step is that frequencies in a signal change over time, and so in most cases it doesn't make sense to do the Fourier transform across the entire signal in that we would lose the frequency contours of the signal over time. To avoid that, we can safely assume that frequencies in a signal are stationary over a very short period of time. Therefore, by doing a Fourier transform over this short-time frame, we can obtain a good approximation of the frequency contours of the signal by concatenating adjacent frames.

The speech signal is segmented into small duration blocks of 20-30 ms known as frames. Voice signal is divided into N samples and adjacent frames are being separated by M (M<N). Typical values for M=100 and N=256. Framing is required as speech is a time varying signal but when it is examined over a sufficiently short period of time, its properties are fairly stationary. Therefore, short time spectral analysis is done.

Hamming Windowing

Each of the above frames is multiplied with a hamming window in order to keep continuity of the signal. So, to reduce this discontinuity we apply window function. Basically, the spectral distortion is minimized by using window to taper the voice sample to zero at both beginning and end of each frame.

Fast Fourier Transform

FFT is a process of converting time domain into frequency domain. To obtain the magnitude frequency response of each frame we perform N-point FFT, which is also called Short-Time Fourier-Transform (STFT), where N is typically 2565 or 512 (here N is 256). By applying FFT, the output is a spectrum or periodogram.

Mel-Scale Filter bank

Speech generation acts like in a pipeline form. Initially something called a glottal pulse is formed. This is like a noisy signal, high pitched signal that gets generated by vocal folds. This signal passes through vocal tract and vocal tract acts as filter on the glottal pulse and by filtering the initial signal, it creates the speech signal depending on the shape of vocal tract. Glottal pulse carries information about pitch or high frequency. Vocal tract or the frequency response from the filter gives the

information of the timbre (different phonemes, consonants) and speech can be formally defined as convolution of vocal tract frequency response with glottal pulse.

Our auditory system is like a logarithmic nature i.e., equal distances on the scale have same "perceptual" distance. This is the same feature of Mel-Scale. We multiply magnitude frequency response by a set of 20 triangular band pass filters in order to get smooth magnitude spectrum. It also reduces the size of features involved. Conversion from frequency scale to Mel scale is done as follows.

$$\text{Mel (f)} = 1125 * \ln(1 + f/700)$$

The core idea of this transformation is that sounds of equal distance on the Mel Scale are perceived to be of equal distance to humans. After this, we apply logarithm to this output signal and get log mel spectrogram.

Discrete Cosine Transform

We apply DCT to the signal obtained from the triangular band pass filters to have L mel-scale cepstral coefficients. DCT formula is shown below:

$$Cm = \sum_{k=1}^{N} \cos\left[m*(k-0.5)*\pi/N\right]*E_k, m = 1,2\ldots\ldots L$$

Where N is number of triangular band pass filters, L is number of mel-scale cepstral coefficients. Usually, N=20 and L=12.

DCT transforms the frequency domain into a time-like domain called quefrency domain. These features are referred to as the mel-scale cepstral coefficients. We can use these generated MFCC alone for speech recognition.

Training and Testing

A total of 7 speakers were taken here for training. Speech samples were taken from all of them and MFCC feature extraction was done and kept these as training data.

Now, one separate sample was taken from each of them for testing. Euclidian distance metric was used in order to verify the MFCC between test and train data.

Results

SPEAKERS	Trained 1	Trained 2	Trained 3	Trained 4	Trained 5	Trained 6	Trained7
Test 1	2.138999	4.311361	5.573339	5.105004	4.955322	3.952164	4.540837
Test 2	3.973299	1.072943	5.028429	5.231422	5.408872	3.648512	4.639406
Test 3	3.827178	4.835906	3.455899	4.381159	4.567722	4.159936	3.569927
Test 4	5.543269	5.635289	4.019038	3.463777	4.666081	5.583079	4.082547
Test 5	4.526015	4.929328	3.421182	3.375858	3.179541	2.449608	3.612892
Test 6	5.114221	3.523382	3.214627	4.012112	3.615244	2.351184	3.761233
Test 7	4.906095	4.775661	0.913201	2.568464	2.79629	3.805715	1.048358