Automatic Speaker Recognition Using SVM

Umer Malik[1], P.K. Mishra[2]

[1] M.Tech Student, Computer Science & Engineering, Sharda University, Plot No. 32-34, KnowledgePark III, Greater Noida, U.P., India - 201306

[2]Assitant Professor, Department of Computer Science & Engineering, Sharda University, Plot No. 32-34, KnowledgePark III, Greater Noida, U.P., India – 201306

Abstract: ***In this paper we describe the use of Wavelet Transform (WT) and SVM in the process of recognizing a speaker. Feature extraction and Denoising is done through Wavelet Transform and SVM is used in order to serve the purpose of classifier.***

Keywords: WT, SVM, HMM, DTW, LFCC, MFC, FFT, Mean, Variance.

1. Introduction

1.1 Speaker Recognition

Many levels of information are contained by a speech signal. The spoken words basically convey a message somehow we may say that the information like emotion, gender and identity of the person is conveyed by the speech of a speaker. In order to identify a speaker, numerous speaker recognition techniques have been devised. Based on those techniques many applications like voice dialing, voice mail, telebanking, security check for confidential information areas have been built.

Speaker recognition is the process of recognizing or identifying a speaker with the help of the voice of the speaker. This can serve the purpose of highly efficient security system in order to grant or deny the access to the authentic or authentic user respectively. Extracting speakers voice with a better performance from a noisy speech signal has been a major setback or a challenge.

We can minimize the error rate if the recognition system is kept in a separate box where the is no interference with the other signals.

In this paper the use of Wavelet Transform (WT) and the Support Vector Machine (SVM) is described for recognizing the speaker.

1.2 Wavelet Transform

Mathematically, wavelet series is a representation of a square integrable function by certain orthonormal series generated by a wavelet. The basic idea of wavelet transform is that the transformation should allow the changes
in time extension only and not the shape.

1.3 Support Vector Machine (SVM)

Support vector machine (SVM) are the advanced models with integrated learning algo's in which classification and regression analysis is done by analyzing data and recognizing the patterns.

Formally we may say that the SVM constructs a hyperplane or a set of hyperplanes in a high or infinite dimensional space that can be used for regression, classification or some other sort of tasks.

2. Conventional Approaches.

The approaches that have already been used previously include Hidden Markov Model (HMM), Mel Frequency Cepstral Spectrum (MFCC), Dynamic Time Warping (DTW), Wavelet Packet Filter Bank, Linear Predictive Coding (LPC), Pure FFT, Power Spectral Analysis, Perceptual Linear Prediction. All these techniques are briefly discussed below:

2.1 Hidden Markov Model (HMM)

A Markov chain that only partially observable states, is said to be Hidden Markov Model. We may say that the observations related to the states of a system are not sufficient to determine the states properly. Markov model is a stochastic model that is used to model randomly changing systems. It is assumed that the future states are dependent only on present states and not on preceding states. Reasoning and computation are possible only due the assumptions made. Viterbi algorithm and forward algorithm are well known algorithms for Hidden Markov Model. Baum-Welch algorithm is also an example of HMM.
Sequential analysis using HMM:

- Construct an HMM model.
- Design an HMM generator for the observed sequences.
- Assign hidden states to sequence regions.
- Set up the question to be answered in terms of hidden path way.
- Train the HMM
- Supervised or Unsupervised.
- Analyze sequences
- Viterbi decoding: Compute most likely hidden path way.
- Forward/Backward: Compute likelihood of sequences.

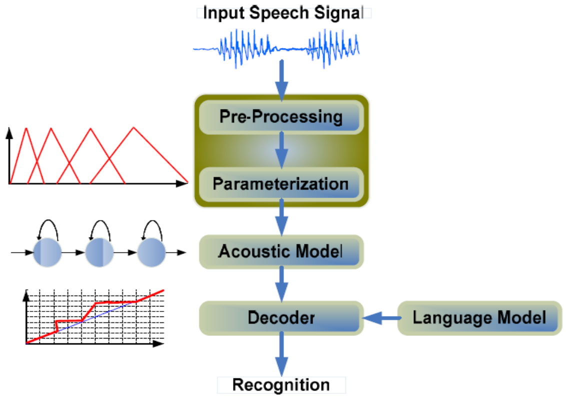


Figure 2.1: Block diagram of HMM based speech recognition process

2.2 Mel Frequency Cepstral Spectrum.

The coefficients that altogether make up Mel Frequency Cepstrum (MFC) are known as MFCC. MFC is the representation of short term power spectrum of sound. It depends on Linear Cosine Transform of a log power spectrum on a non-linear Mel Scale of frequency. The difference between MFC and the Cepstrum is that in MFC the frequency bands are equally spaced on Mel Scale and MFC can study the human auditory system much better. Noise sensitivity of MFCC is not robust and is inefficient in presence of the additional noise. Thus it allows better representation of sound. Here, there are some steps for the MFCC derivation.

a) First of all take Fourier Transform (FT) of sound signal.
b) Next, the power of spectrum obtained in above step are mapped by using Triangular overlapping window.
c) Take log of powers at each Mel frequency.
d) The amplitude of the resulting spectrum is MFCC.

The use of Mel Frequency Cepstral Coefficients can be considered as one of the suitable method for feature extraction. Use of about twenty MFCC coefficients is common in ASR, although 10-14 coefficients are often considered to be sufficient for coding speech. The most important limitation of using MFCC is its sensitivity to noise due to its dependence on the spectral form. Methods that utilize information in the periodicity of speech signals could be used to overcome this problem, although speech also contains a periodic content.

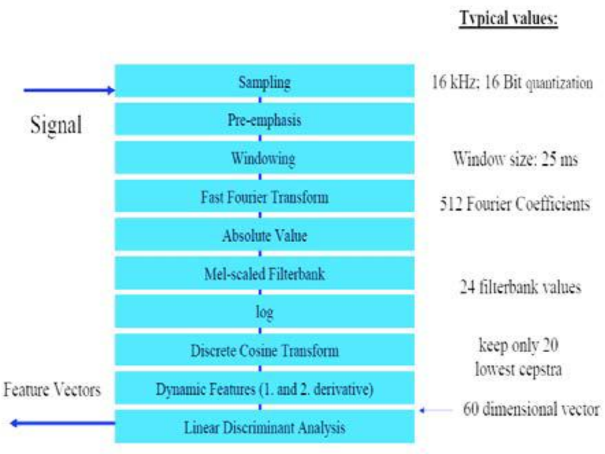


Figure 2.2: Pipelined MFCC

2.3 Dynamic Time Warping

Under some boundary condition's the regularity between any two given time dependent sequences can be estimated by a very well-known algorithm called as Dynamic Time Warping (DTW). In order to compare them, the sequences are warped in Non Linear Pattern. Dynamic Time Warping was actually used to compare the speech patterns in automatic speaker recognition. Later on it was applied to fields like information retrieval from audios and videos and data mining. DTW can analyze any data that can be turned into linear sequence. Speaker recognition is a well known example of it. DTW has some limitations like it has quadratic time and space complexity that limits its use to small time series.

The time alignment of different utterances is a serious problem for distance measures and a small shift would lead to incorrect identification. Dynamic time warping is an efficient method to solve this time alignment problem. This is the most popular method for speaking rate variability in template-based systems. The asymmetric match score β of comparison of an input frame ***y*** of M samples with the template sequence ***x*** is given as follows

$$\beta = \sum_{i=1}^{M} d(y_i, x_{j(i)})$$

The template indices *j(i)* are given by the DTW algorithm. This algorithm performs a piece wise linear mapping of the time axis to align both the signals. The variation over time in the parameters corresponding to the dynamic configuration of the articulators and the vocal tract is taken into account in this method. Figure below, shows the dynamic time warp of two energy signals. The warp path is a diagonal line for two identical signals and the warp has no effect. The accumulated deviation from the dashed diagonal warp path is the Euclidean distance between two signals and the parallelogram surrounding the warp path acts as boundary conditions for preventing excessive warping.

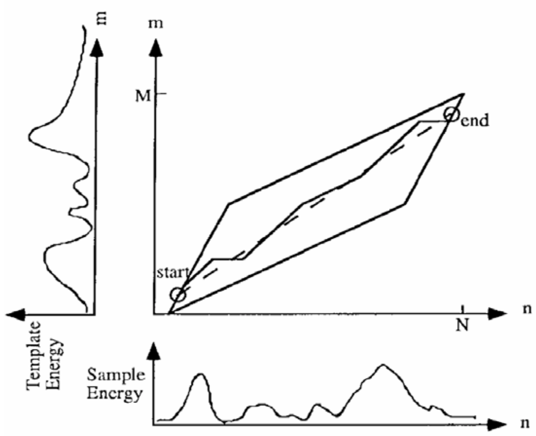


Figure 2.3: DTW of two energy signals

2.4 Wavelet Packet Filter Bank

Most of the research work going on in the field of speech recognition is to improve the performance of the recognition of the noisy speech. Human mind processes the signals using neural networks which work very fast due to parallel processing. So, neural network is better than other techniques for recognizing speech signals. The speaker independent system is improved by utilizing a different cochlea model which is designed with a high resolution Wavelet Packet Filter Bank (WPFB).

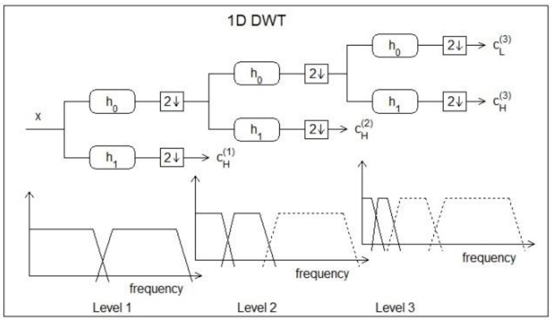


Figure 2.4: Wavelet decomposition of a sound signal.

2.5 Linear Predictive Coding (LPC)

LPC is one among the most powerful speech analysis techniques and is a very useful method for encrypting speech at a lesss bit rate. The main motive behind linear predictive analysis is that a particular speech signal at the current time can be considered as a linear combination of previous speech samples

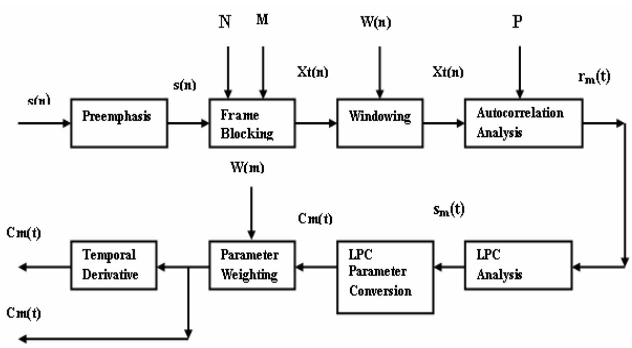


Figure 2.5: The LPC processor

2.5.1 Types of LPC

Following are the types of LPC

- Voice-excitation LPC
- Residual Excitation LPC
- Pitch Excitation LPC
- Multiple Excitation LPC(MPLPC)
- Regular Pulse Excited LPC(RPLP)
- Coded Excited LPC(CELP)

2.6 Pure FFT

Despite the popularity of MFCCs and LPC, use of vectors containing coefficients of FFT power-spectrum are also possible for feature extraction. As compared to methods retrieving knowledge about the human auditory system, the spectrum of pure FFT carries comparatively more information about the speech signal. However, most of the information is located at the comparatively higher frequency bands when using high sampling rates which are not oftenly considered to be salient in speech recognition.

2.7 Power Spectral Analysis (FFT)

One among the common techniques of studying a speech signal is through the power spectrum. The power spectrum of a speech signal determines the frequency content of the signal over time. The first step of computing the power spectrum of the speech signal is to perform a Discrete Fourier Transform. A DFT computes the frequency information of the equivalent time dependent signal. Since a speech signal contains real point values only, we can use a real-point FFT to enhance the efficiency. Output contains both the phase and magnitude information of the actual time domain signal

2.8 Perceptual Linear Prediction (PLP)

The PLP model developed by Hermansky in 1990. The aim of the original PLP model is to describe the psychophysics of human hearing more precisely in the feature extraction process. PLP is similar to LPC analysis and is based on the short-term spectrum of speech. In comparison to Pure Linear Predictive analysis of speech, PLP changes the short-term spectrum of the speech by several psychophysically based modifications.

3. Methodology

3.1 Wavelet Transform (WT)

The integral wavelet transform is the integral transform defined as

$$[W_\psi f](a,b) = \frac{1}{\sqrt{|a|}}\int_{-\infty}^{\infty}\overline{\psi\left(\frac{x-b}{a}\right)}f(x)dx$$

The wavelet coefficients c_{jk} are then given by

$cjk = [W\psi f](2-j, k2-j)$

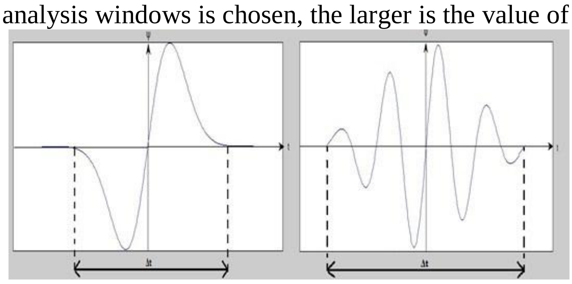
Here, $a = 2^{-j}$ is called the binary dilation or dyadic dilation, and $b = k2^{-j}$ is the binary or dyadic position

Basic idea

The fundamental idea of wavelet transforms is that the transformation should allow only changes in time extension, but not shape. This is effected by choosing suitable basis functions that allow for this. Changes in the time extension are expected to conform to the corresponding analysis frequency of the basis function. Based on the uncertainty principle of signal processing

$$\Delta t \Delta \omega \geq \frac{1}{2}$$

The higher the required resolution in time, the lower the resolution in frequency has to be. The larger the extension of the analysis windows is chosen, the larger is the value of Δt .



When Δt is large:
1. Bad time resolution
2. Good frequency resolution
3. Low frequency, large scaling factor

When Δt is small:
1. Good time resolution
2. Bad frequency resolution
3. High frequency, small scaling factor

We may conclude that the basis function Ψ can be regarded as an impulse response of a system with which the function x(t) has been filtered . Transformed signal provides information about the time and the frequency. Thus, WT contains information that resembles the short-time-Fourier-transformation, but with additional properties of the wavelets, that show up at the resolution in time at higher analysis frequencies of the basis function. The main difference in the time resolution at increasing frequencies for the FT and the WT is shown below.

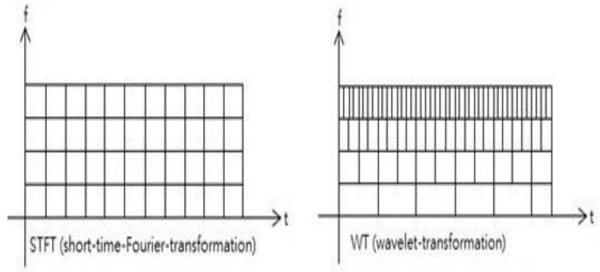


Figure 3.1: (a) FT vs WT

This shows that wavelet transformation is good in time resolution of high frequencies, while for slowly varying functions, the frequency resolution is remarkable.

Another example: The analysis of three superposed sinusoidal signals $y(t) = \sin(2\pi f_0 t) + \sin(4\pi f_0 t) + \sin(8\pi f_0 t)$ with STFT and wavelet-transformation.

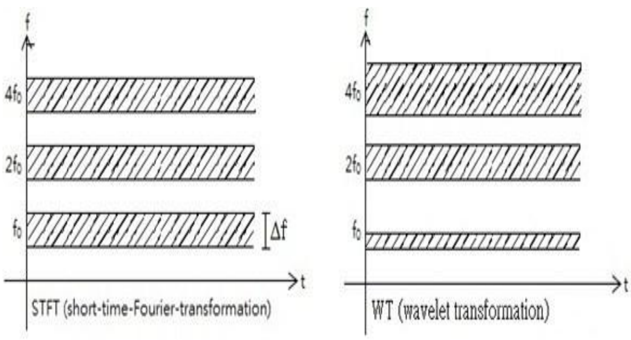
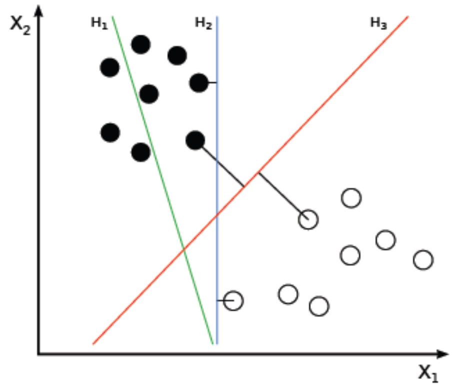


Figure 3.1: (b) FT vs WT

3.2 Support Vector Machine (SVM)

Support Vector Machines (SVM) are the learning models with integrated learning algos that analyze data and recognize patterns, used for classification and regression analysis. From the set of training examples, each marked as belonging to one of two classes, the SVM training algorithm builds a model that assigns new examples into one category or the other hence making it a non probabilistic binary linear classifier. An SVM model is a depiction of the examples as points in the space, mapped such that the examples of the separate categories are distinguished by a clear gap which is as wide as possible. More precisely, an SVM constructs a hyperplane or set of hyperplanes in an infinite dimensional space, which can be useful for classification, regression, or some other tasks.



H_1 does not separate the classes. H_2 does, but only with a small margin. H_3 separates them with the maximum margin.

Figure 3.2: SVM

Support vector machines are a very fine techniques for inference with minimal parameter choices. The translation into the popular adaptation of SVM in many application domains by non SVM experts has sufficiently increased. The popular success of previous methodologies like neural networks, genetic algorithms, and decision trees was led by the intuitive motivation of these approaches, that in some sense enhanced the end users ability to develop applications independently and have a confidence in the results obtained. There are three main ideas needed to understand SVM: maximizing margins, the dual formulation, and kernels. Most people intuitively grasp the idea that maximizing margins should help in improving generalization. Changing from the primal to dual formulation is typically black magic for those uninitiated in duality theory. Duality is the core concept usually missing in the understanding of SVM.

3.2.1 SVM Properties

Support Vector Machines belong to a family of generalized linear classifiers and can be considered as an extension of the perception. They can be considered a case of Tikhonov regularization. One of the special properties is that they simultaneously minimize the empirical classification error and maximize the geometric margin thus they are also known as maximum margin classifiers

4. Implementation

The implementation of Wavelet Transform and SVM is described through the following flow diagram.

Overall flow diagram of speaker recognition process

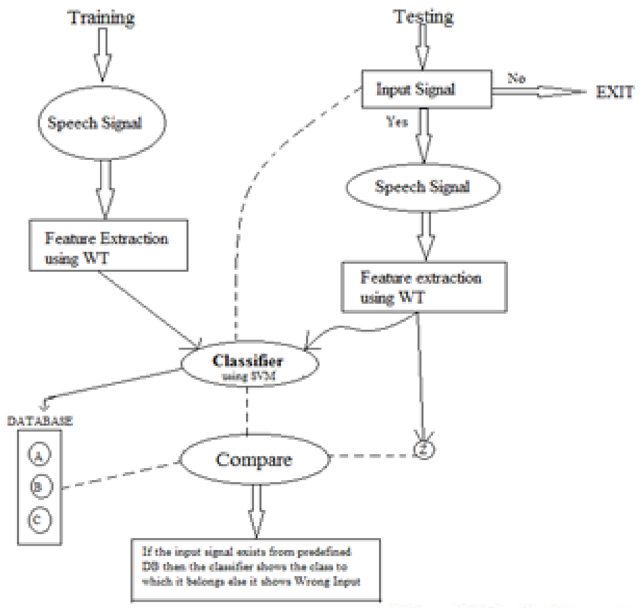


Figure 4: Overall flow diagram of speaker recognition process

4.1 Training Phase

In the training phase, firstly we should have some database for the speaker recognition. Here, we have taken 2 speakers voice which will be recognized by our system. We take 20-25 speaker voices for the database. These voices are converted into .wav format because MATLAB is taken sound in .wav format. Then using wavelet transform, we decompose the input speech signal into 4 decomposition level with different frequency description.

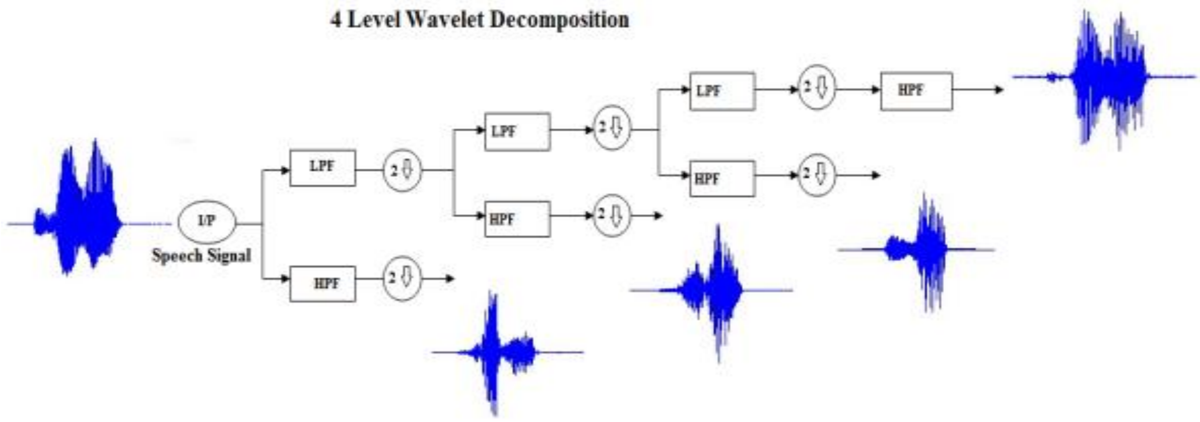


Figure 4.1: 4Level decomposition using wavelets transform

After that we calculate mean and variance of 4 different levels and make a matrix of this dataset. This is the feature extraction of input speech signal using wavelet transform.

4.2 Testing Phase

In the testing phase, we take one speech signal and applied to the wavelet filter bank for the decomposition. After that system generates 4 decomposition components and calculates its mean and variance. Then using support vector machine, we match these features with the predefined database. If it matches, then it shows which person it is. This is the speaker recognition process.

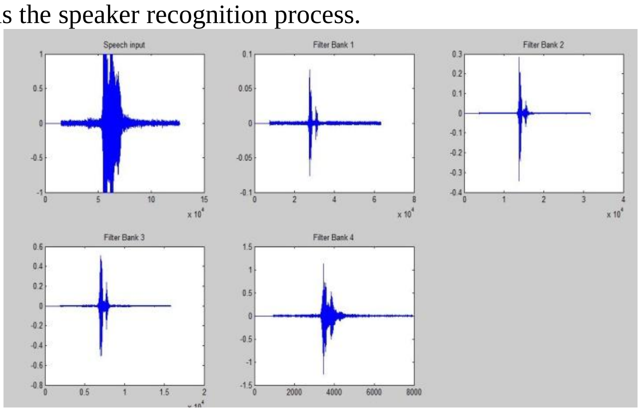


Figure 4.2: Wavelet filter bank output

5. Results Analysis

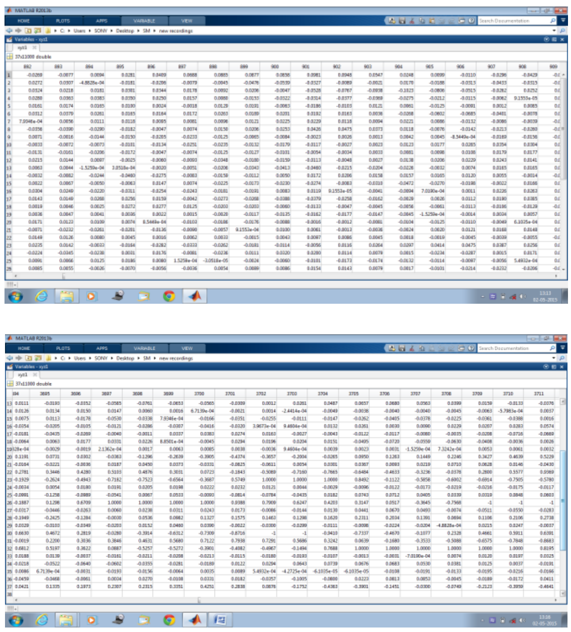


Figure 5: Information contained by the input signals at different intervals.

The graphs generated on plotting this data is given below. Only few of them are plotted

5.1 PLOT: Input Signals of Speaker 1

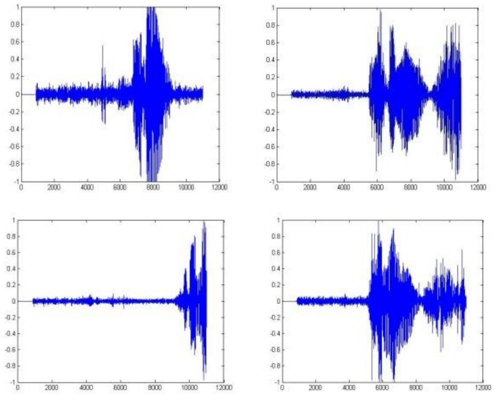


Figure 5.1: Energy band of input signal of speaker 1

5.1.1 4 Level Decomposition of one of the input signals of Speaker 1 using WT

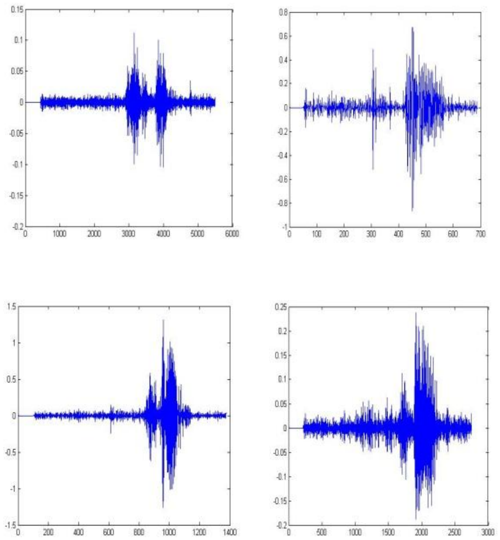


Figure 5.1.1.4: Levels of Decomposition

By the use of Wavelet Transform the input signal is decomposed upto 4th level. This process helps us to denoise the signal. Up-sampling and Down-sampling are used.

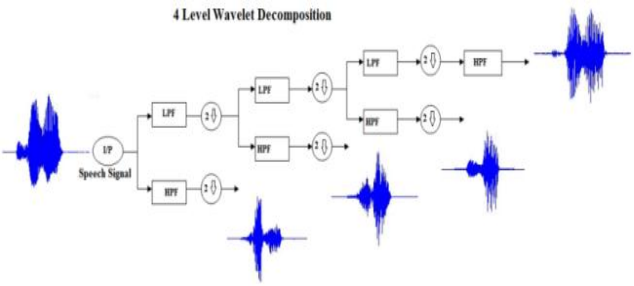
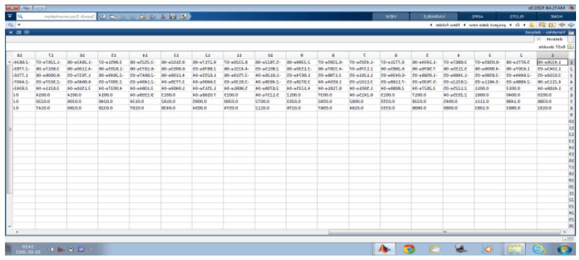


Figure 5.1.2: Example of 4 level decomposition using WT

5.2 Mean and Variance

Even though the signals obtained after the decomposition of the input signal (as shown in above example) could have been used to identify the speaker but the efficiency in that case was too low (50-60%). Thus it was very important to find a way that would increase the efficiency of our system. So out of so many features Mean and Variance are two such features which have been taken into consideration and based upon these two features the identification process of a speaker is carried out. The calculated mean and variance of some of the signals in our dataset is shown in the below given tables.



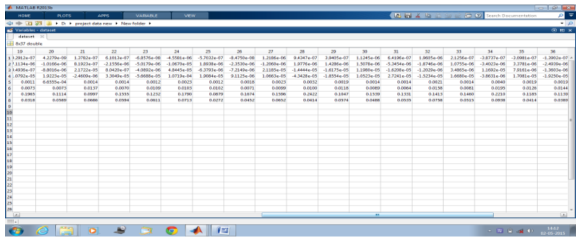


Figure 5.2: Mean and Variance of input signals

5.3 The Graphical User Interface

5.3.1 ON: This will turn on the tool for recognizing the speaker. Unless and until the ON button is clicked the tool will not function and likewise rest of the buttons will be disabled.

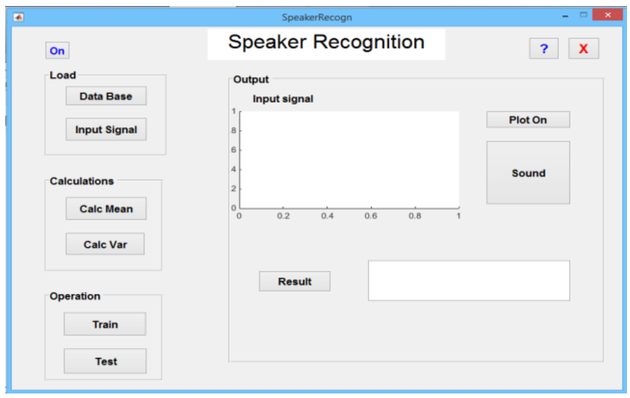


Figure 5.3.1: ON/LOADING Window

5.3.2 LOAD: This section deals with loading of Database that is already stored in our system and the Input Signal which is to be compared with the stored database.

a) DATABASE: In order to perform the process of speaker recognition we need to develop or maintain a database of all the voices recorded against the class of a particular speaker. For every speaker whose voice is stored into our database will have its name labelled that can be a name, number or anything eg. “Speaker 1”, “Speaker 2” and so on.

b) INPUT SIGNAL: This button serves the purpose of loading the new input signal which is to be compared with the database stored. The new input signal has to be in such a format which is acceptable by our tool. Input may directly ne given by recording a voice on computer system but we have to make sure that the level of noise should be minimum. Similarly the input may also be given through some other recording devices as well.

5.3.3 Calculations

Calculation of the features (based upon which the speaker recognition) is done in this step.

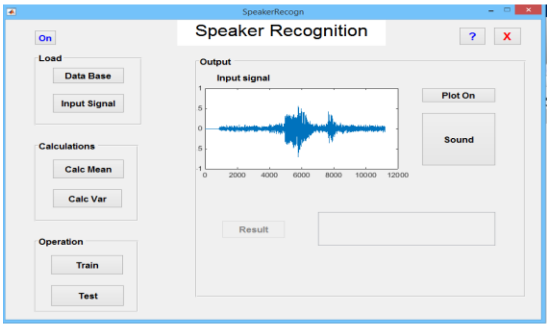


Figure 5.3.3: Calculation Window

a) Mean: By clicking on this button the mean of the input signal will be calculated.
The **mean**, denoted by **μ** (Greek *mu*), is simply the average value of a signal. It is calculated just by adding all of the samples together, and divide by *N*. Mathematically:

$$\mu = \frac{1}{N}\sum_{i=0}^{N-1} x_i$$

b) Variance: The variance is calculated by clicking on this button on the user interface. The variance of a signal is found by summing the deviations of all the individual samples, and then dividing by the number of samples, N. The variance provides a single number representing the typical distance that the samples are from the mean. The calculated Mean and Variance can then be seen in the output window by clicking on the **PLOT** button. **Sound** button will play the spoken phrase by the speaker.

5.3.4 Operation

Two operations will be carried throughout the process i.e Training and Testing.
a) TRAINING: In the training phase, firstly we should have some database for the speaker recognition. Here, we have taken 2 speakers voice which will be recognized by our system. We take 20-25 speaker voices for the database. These voices are converted into .wav format because MATLAB is taken sound in .wav format. Then using wavelet transform, we decompose the input speech signal into 4 decomposition level with different frequency description.

b) TESTING: In the testing phase, we take one speech signal and applied to the wavelet filter bank for the decomposition. After that system generates 4 decomposition components and calculates its mean and variance. Then using support vector machine, we match these features with the predefined database. If it matches, then it shows which person it is. This is the speaker recognition process.

5.3.5 Result Window

After performing all the previous steps now our machine is ready to generate the results, based upon the speaker's authenticity if the speaker is one among those whose voice is already there in our database and has been labelled then the output will be generated that will identify the speaker and vice versa for the speaker whose voice is not found in the database .As we can see in the following picture that the output shows "Person 1" as the same speaker's voice has been labelled as Person 1 and similarly for other speakers the labelling will be different.

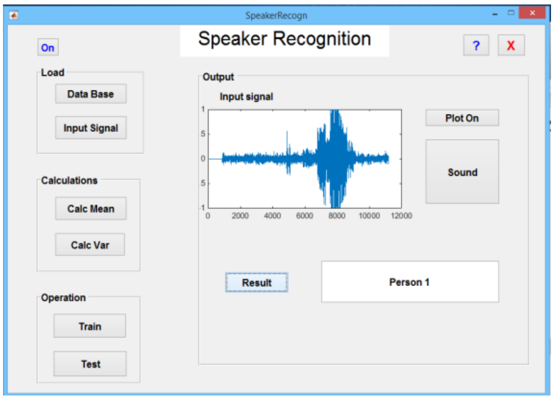


Figure 5.3.5: Result Window

In addition there are two more buttons [X] and [?].
[?] : This button will provide the information of the developer and
[X] : This button will close the window

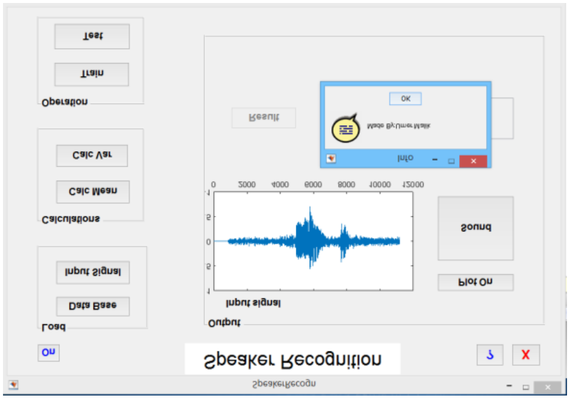


Figure 5.3.6: Info Window

6. Conclusion

Speech recognition is currently used in many real-time applications, such as cellular telephones, computers, and security systems. However, these systems are far from perfect in correctly classifying human speech into words. Speech recognizers consist of a feature extraction stage and a classification stage. The parameters from the feature extraction stage are compared in some form to parameters extracted from signals stored in a database or template. The parameters could be fed to the SVM or neural network or Hidden Markov Model as well The goal of this paper is to develop a speech recognition algorithm that uses the wavelet transform to extract and represent incoming speech signals as a basis for an accurate method of identifying and matching these signals to signals in a template. By the implementation of SVM we found that high prediction performance and high prediction accuracy has been achieved.

7. Future Work

In last few years Speaker Recognition Tool has been in great demand and the demand is increasing with the time. Thus we should ensure that the access granting and denying is highly secure. As this tool has found its application's in wide areas ranging from defence to the access of very delicate data of the large MNC's. In future an administrative control can be set on the tool which will govern the addition to the database. Only an administrator will be able to add or delete the database. This can be done by linking the tool with the administrators email and hence will enhance the security and authenticity. The number of features based upon which the speaker is recognized can be increased and as the accuracy will go on increasing the system can be made more scalable with respect to the number of users.

References

[1] ICSI Speech "FAQ: 5.1, What are features? What are their desirable properties? Answer by: dpwe", http://www.icsi.berkely.edu/local-cgibin/man-cgi-bin?ftrs-intro.html, 26 May, 2000.
[2] Bojan Imperl, "Speaker recognition techniques", Laboratory for Digital Signal Processing, Faculty of Electrical Engineering and Comp. Sci., Maribor, Slovenia, 2000.
[3] Umer Malik, "A Review of Comparatively Study of Different Speaker Recognition Techniques," IJSR volume 4 issue 4, 2015.
[4] D. O.Shaughnessy, "Speech Communications- Human and Machine", Universities Press (India) Limited, 2001.
[5] L. Rabiner and R. Schafer, "Digital Processing Speech Signals", Englewood Cliffs, NJ: Prentice Hall, Inc., 1978.
[6] Bojan Imperl, Zdravko, and Bogomir Horval, "Use of Harmonic Features in Speaker recognition", Laboratory for Digital Signal Processing, Faculty of Electrical Engineering and Comp. Sci., Maribor, Slovenia, 2000.
[7] R.Gandhiraj, Dr.P.S.Sathidevi, "Auditory-based Wavelet Packet Filter bank for Speech Recognition using Neural Network," 15th International IEEE Conference on Advanced Computing and Communications, pp. 666-671, 2000
[8] M. Kleinschmidt, J. Tchorz and B. Kollmeier, "Combining speech enhancement and auditory feature extraction for robust speech recognition", Speech Communication 34 (2001), 75-91.
[9] T.Dau, D. Puschel, and A. Kohlrausch, "A quantitative model of the effective signal processing in the auditory system: II. Simulations and measurements", J. Acoust. Soc. Am. 99 (6), 3623-3631, 1996
[10] T.H. Hwung, L.M. Lee and H.C. Wung, "Feature adaptation using deviation vector for robust speech recognition in noisy environment", IEEE International Conference on Acoustics, Speech, and Signal Processing 1997 (ICASSP-97), 1997.
[11] Meinard M¨ Uller Frank Kurth Michael Clausen, "Audio Matching Via Chroma-Based Statistical Features" Queen Mary, University Of London, 2005.