Processing of speech signal using wavelet transform and support vector machine

Jayashree B M
Student, Dept of ECE
East West institute of technology
Bangalore-560091, Karnataka,
India

Pooja Nayak S
Asst Professor, Dept of CSE
East West institute of technology
Bangalore-560091, Karnataka,
India

Dr S G Hiremath
HOD, Dept of ECE
East West institute of technology
Bangalore-560091, Karnataka,
India

Abstract-

In the growth of a strong speech recognition system, the speech signal is pre-processed and it is taken as an essential stage. In this project, the processing of the speech signal is employed on the basis of MFCCs for extracting the features from the speech and the WT of these signals. The Mel frequency approach extracts the speech signal features to get the training and testing vectors. Based on the time and frequency multi-resolution property the input speech is decayed into many frequency channels. It is called as MFCCs based speaker detection method.

The MFCCs gives a superior performance in unpolluted environments, and they are not strong enough in loud environments. Support vector machine techniques are used for matching the feature. This project reports on how support vector machine is urbanized and qualified based on the extracted features for purpose of identification.

Keywords**:** *End point detection, Mel frequency cepstral coefficient, Support vector machine, Speaker identification, Wavelet transform*

I. INTRODUCTION

Speech is considered as the most primitive, frequent and resourceful form of medium which is used to communicate with each other. It is a distinctive form of audio data and it solely depends on articulation of the spokesman and the way they speak. Speech is dissimilar from various other forms of classification such as passwords or keys, because it is the most non-intrusive as a biometric. So with this feature, a voice of the person cannot be stolen, elapsed or lost, therefore speaker recognition proceeds for a safe and sound method of authenticating speakers.

Two main operations performed in speaker identification namely feature extraction and classification. Feature extraction reduces the data that makes effort to detain the necessary characteristics of spokesman with small rate of data. Speech features usually extracted from a variety of techniques such as the linear predictive coding (LPCs), Linear prediction cepstral coefficients (LPCCs) and Mel-frequency cepstral coefficients (MFCCs). Because of the simplicity and effective nature of LPCs they are used in speaker recognition. MFCCs are used because the set of filters used has equal bandwidth with admiration to the Mel-scale frequencies and they are evaluated by using filter bank tactic. MFCCs will not ensue a linear scale even though they are based on human ear sensitive to the frequency of sounds.

Classification has two parts namely modeling of the speaker and matching of the speaker. Classifier is defined as the blend of both speaker model and technique matcher. Classification techniques used contain Gaussian mixture models (GMMs), Hidden Markov models (HMMs), Artificial neural networks (ANNs) and Support vector machine (SVMs). In speaker identification, MFCCs are considered as the popular acoustic feature hence it is used. The information carried by the MFCCs are related on the known proof i.e. the information passed by the lower frequency components speech is more superior than conceded by higher frequency components.

SVM is chosen as the classifier of speech signals. In particular, they are worn to classify the speech patterns and they control the generalization mechanically. The basic idea of SVM is to attain maximum separation between the two classes by constructing the most favorable hyper-plane, as a decision face. The creation of hyper-plane is depend highly on this available data points. In order to solve the binary classification problems, the SVMs based on kernel are used initially. The SVMs are defined on the basis principles of structural risk minimization (SRM), which is considered to be much better than the usual Empirical risk minimization (ERM) principle. Due to the several characteristics possessed by the SVMs, they are declared as the effective separating classifiers. The characteristics are as

ISSN: 2278 – 7798

follows: Maximum margin is required for solution, Dealing with very high dimensionality samples, low cost.

The general process of speaker identification involves two stages:

Training phase

Testing phase

In training phase, a fresh voice with unknown identity is stored in the database of the system. In testing phase, an unfamiliar speaker produces an input and later the system provides a identity of the speaker. Later, the resultant feature vector obtained from both training and testing phase are compared using the SVM classifier and the result is displayed.

II. Overview of algorithms

2.1 Pre-processing

Pre-processing technique is used to alter the speech signal, so that it will be more flexible for analyzing the feature. By sampling the input of the recorded speech, discrete time speech signals are obtained. To enhance the feature extraction, pre-processing techniques are used. This technique is also used to modify the discrete speech signal. Pre-processing technique consists of windowing, pre-emphasis and voice activation detection.

2.1.1Pre-emphasis

Pre-emphasis filter is used to reduce differences in power of different components of the signal and it is defined as

$$y(n) = x(n)-a\ x(n-1) \qquad (1)$$

After using pre-emphasis filter, differences in amplitude of components are dramatically decreased. They are used to reduce the effect of noise.

Pre-emphasis is used in processing of speech signal to boost high frequencies of the signal. Two main factors are used to drive for pre-emphasis. Initially, higher frequency speech signal has more detailed information than lower frequency speech signal. Secondly, pre-emphasis reduces some of the noisy sound effects from the vocal parameters.

Pre-emphasis is described as a first-order FIR filter and it is given as

$$H(z) = 1-\alpha z^{-1} \qquad (2)$$

Basically, α is selected to be between 0.9 and 0.95. We have used α=0.95.

2.1.2 Windowing

The speech signal has discontinuity in the signal at the starting and ending of each frame. So, to reduce the discontinuity at either end of the block, windowing technique is used. In blocking, windowing technique is used to reduce the distortion and to diminish the signal to level zero at both side of every frame.

If we distinct the window as w(n), $0 \le n \le N-1$, N is defined as the samples number in every frame, and the resultant signal in every frame is as follows

$$y(n) = x(n)*w(n) \qquad (3)$$

Typically, Hamming window is generally utilized for the windowing process, and it is given as

$$w(n) = 0.54- 0.46\cos\left(\frac{2\pi n}{N-1}\right) \qquad (4)$$

2.1.3 Voice activation detection (VAD)

VAD is basically employed to split speech and non-speech information from a speech signal. Pre and post pronunciation, and plain signal between the words are included in non-speech. The major problem for the speech recognizer in the speech signal is to identify the endpoints of an expression. Speech performance is decreased with the inaccurate endpoint detection. Some of the most frequently used measurements for analyzing speech are short-term of energy estimate E_{s1}, or power estimate P_{s1}, and zero crossing rate Z_{s1}.

For signal $s_1(n)$ above specified constraints are evaluated as follows:

$$E_{s1}(t) = \Sigma_{n=t-L+1\ to\ t}\ s_1^2(n) \qquad (5)$$

$$P_{s1}(t) = 1/L\ \Sigma_{n=t-L+1\ to\ t}\ s_1^2(n) \qquad (6)$$

$$Z_{s1}(t) = 1/L\ \Sigma_{n=t-L+1\ to\ t} |\mathrm{sgn}(s_1(n)) - \mathrm{sgn}(s_1(n-1))|/2 \qquad (7)$$

Where $\mathrm{sgn}(s_1(n)) = +1$ if $s_1(n) > 0$

$\mathrm{sgn}(s_1(n)) = -1$ if $s_1(n) \le 0$

Some triggers are needed to these processes for creating decision about where the utterance start and finish. In order to generate this trigger, some kind of information is needed regarding the surroundings noise. First 10 blocks are assumed as a background noise for triggering these measures.

For this, the trigger can be defined as follows:

$$t_w = \mu_w + \alpha\ \delta_w \qquad (8)$$

where, μ_w is the mean

δ_w is the variance calculated for the 10 blocks

α is a constant term, according to the characteristics of the signal they have to be fine tuned which is given as

$$\alpha = 0.2\ \delta_w^{0.8} \qquad (9)$$

The voice activation detection function VAD(t) can be calculated as follows:

$$VAD(t) = 1 \qquad .(10)$$

$W_{s1}(t) \ge t_w = 0$, $W_{s1}(m) \ge t_{w,,}$

where $W_{s1}(t) = P_{s1}(t).\ (1- Z_{s1}(t))\ S_c\ S_c = 1000$

2.2 Feature Extraction

The feature extraction is generally used to recognize the low level properties of speaker. During, the speech production, the amount of data generated is quiet huge while the essential characteristics of the speech relatively adjust gradually and henceforth less

ISSN: 2278 – 7798

data is required. The speaker discriminative information is retained even if the data is reduced this is done by feature extraction process. The extraction captures the satisfactory information produced for better discrimination of speaker in a structure and size which allows resourceful modeling.

In this, the speech is transformed into a group of feature vectors called parameters. The feature vectors are used to generate a pattern for every speaker. The speech signal representation for the speaker detection includes wide range of possibilities such as LPCC, MFCC and others. Whereas MFCC is considered as the best, known and popular coding technique.

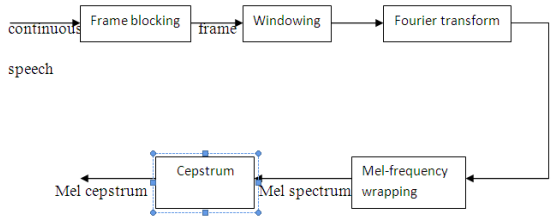
2.2.1Extraction of Mel-frequency cepstral coefficients

The MFCCs are frequently employed for identification of speaker and they are derived through cepstral analysis. As the MFCCs are evaluated using the filter banking approach, hence it has bandwidth which equals the bandwidth of Mel-scale frequencies. Hence the sounds in human perception will be in linear scale. On the basis of acknowledged variation of critical bandwidth of human ear's MFCCs is evolved and filters are spaced linearly for lower frequencies signal and logarithmically spaced at higher frequencies. And this expressed in Mel-frequency scale and it is adjusted linearly for below 1000Hz frequency and spaced logarithmically for above 1000HZ frequency. Excitation components source and vocal tract are separated using cepstral analysis tool due to this speaker related information is obtained. The redundant pitch information is separated from the information of vocal and it is done by a tool called cepstral analysis.

Speech signal is typically recorded above 10000Hz sampling rate. The aliasing effect is minimized in Ana log-to-Digital conversion by this chosen frequency. All frequencies up to 5K Hz can be captured by these sampled signals; it covers almost all energy sounds that are produced by humans.

The apparent pitch or frequency tone is measured by unit called Mel. Therefore, the original frequency scale is mapped in Hz and the apparent frequency scale is mapped in Mel's. Mapping is given by the following equation:

$$f_{Mel} = 2595 \log_{10}\left(\frac{1+f_{linear}}{700}\right) \qquad (11)$$



In this, the secluded speech is divided as frames of each N number of samples. Later the discontinuities in the signal are minimized using windowing technique Here we have chosen the Hamming windowing technique as it eliminates the ripples in the speech signal very easily. And then via fourier transform the frequency domain is obtained by converting the time domain frames.. The converted signal is then transformed next into spectrum form. Here the speech signal is broken into small sections, of N samples. And the sections are defined as frames and quasi-stationary speech inspires this framing process. The section shows stable aural characteristics and considered stationary, if signal over this section examined are of short duration.

2.2.2 Extraction of polynomial coefficients

MFCCs are sensible to time variance or to mismatches among the training and testing data. Sensitivity are decreased by adding other coefficients to MFCCS. Thus polynomial coefficients are used for this purpose. Polynomial coefficients are employed for modeling the slope and curvature of the waveform over adjacent frames for every MFCC. The cepstral coefficients of time waveform are enlarged by orthogonal polynomials to calculate the polynomial coefficients.

2.3 Wavelet transform

Wavelets are simply the functions that are used to analyze the data according to scale and resolution. The main purpose of using the one-dimensional wavelet transform is used to remove the noise from the noisy signal. Wavelet transform provides an efficient result even when the signal contains spikes and discontinuity in the signal. The characteristic features of the speaker are removed due of the telephone degradation which usually acts like a low pass filter on the speech signal. In order to overcome this degradation problem an efficient tool called discrete wavelet transform (DWT) is used. The DWT is also a very popular tool for the analyzing of non-stationary signals.

There are many types of wavelets such as Haar, Daubechies, Discerte Meyer and other which can be used to perform wavelet transform. The Daubechies wavelet is one of the popular wavelets and it has been used for speech recognition. It was named after its inventor, a Belgian physicist mathematician Ingrid Daubechies. They are considered to have the utmost count of vanishing moments for given support width which is twice the number of vanishing moments.

ISSN: 2278 – 7798

The properties of Daubechies wavelet are as follows:

- The support length of the scaling function Φ and wavelet function Ψ is 2N-1. N is the vanishing moments of Ψ.
- Some of the dbN are only symmetrical.
- The uniformity increases with the order.

When the value L becomes very huge then Ψ and Φ belongs to $C^{\mu N}$ where μ is just about equal to 0.2. Daubechies-8 wavelet is normally used for decomposition of speech signal as it needs minimum support size for the given number of vanishing points.

For the decomposition of speech signal Daubechies-8 wavelet is used as it needs minimum support size for the given number of vanishing points.

III. Support vector machine

SVM is a latest approach for classifying the pattern of speech signal. The SVMs are primarily intended for solving the problem of binary classification which are based on kernel method. The SVMs are based on the SRM principles which are better than the ERM which is produced by conventional NN. They have been considered as the very efficient classifiers due to their several excellent characteristics such as:

- Maximum margin is required in the solution.
- Dealing with samples of a very high dimensionality.
- Low cost

The key features associated with SVMs are usage of kernels, lacking of local minima, the spare solution for the problem and capacity control.

The SVMs are constructed by optimal hyper plane. A SVM is a binary classifier, and decision boundary is made that possibly divides the two classes. The optimal hyper plane is described as follows:

$$w^T \varphi (x_i) + k = 0 \qquad (12)$$

Where w is defined as vector weight, x_i is the vector from the input space and k is defined as bias that represents the distance among hyper plane and the origin.

The data labeled as $\{x_i , y_i\}$, $x_i \in R^d$, $y_i \in \{ -1, +1\}$, i=1,....N and the optimal hyper plane is chosen based on criteria of maximum margin, i.e. the Euclidean distance between the closest data points on every side of the plane are minimized by choosing the separating plane. For the case of non-separable data, after the creation of the decision surface also some error classifications are still present. The classification errors are considered by changing the equation as follows:

$$Y_i (w^T \varphi(x_i) + k) \geq 1 - \xi i \qquad (13)$$

Where ξi represents the slack variables,that are positive values.

Optimal hyper plane is obtained by minimizing the equation, the condition which has to be minimized and it is defined as:

$$\text{Min}_{w,b,\xi i} L_p = \frac{1}{2}\|w\|^2 + c \sum_{i=1}^{N} \xi i \qquad (14)$$

Where c is a parameter cost,that represents the substitution among the training errors number and generalization capacity of the SVM.

IV. The Proposed method

This system aims to illustrate the fast and accurate speech recognition system. In this a set of speech samples will undergo a filtering process. Then the noisy effects are removed by pre-emphasizing the signal and high frequency signals are enhanced by this and the discontinuities present in the words are minimized by applying to the windowing technique. And the recorded speech is described based on frames in order to differentiate as voiced or unvoiced frame. Usually hamming window is selected as the window ensuring to its helpfulness to yield good performance. The features are extracted using MFCC technique. In this the cut off speech is divided as frames of N samples and for every frame windowing technique is to reduce the noisy effect. Later the frames are converted to frequency domain via Fourier transform. The converted signal is then transformed into spectrum form and it is followed by the Mel-frequency wrapping process to attain its Mel spectrum. The best noise free speech is obtained from wavelet transform with the repeated testing. The skilled SVM is able to detect all the vocal words with training and testing data. The processing technique for the testing phase will be same as that of the training phase. SVM are global and unique, they are less prone to over fitting. The resultant feature vector of both training and testing phase are compared with each other and the output is displayed as classification is done.

The block diagram for the proposed system as shown below:

ISSN: 2278 – 7798

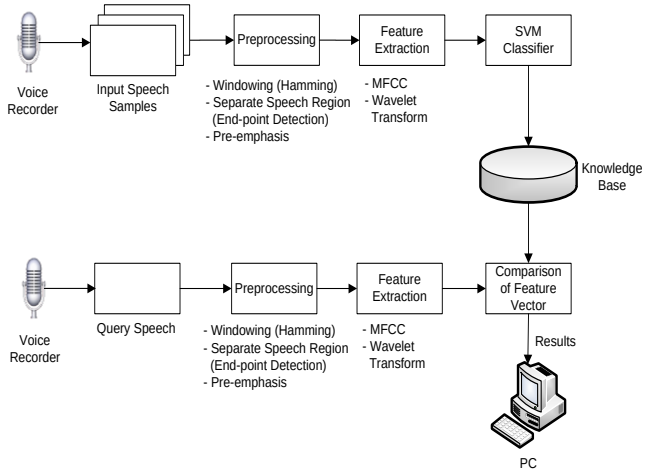


Figure2. Architecture design of the proposed system

V. Result

The results obtained after each stage for the word "GOOD" are given as follows:

1. The input speech for the word "Good" is as shown below

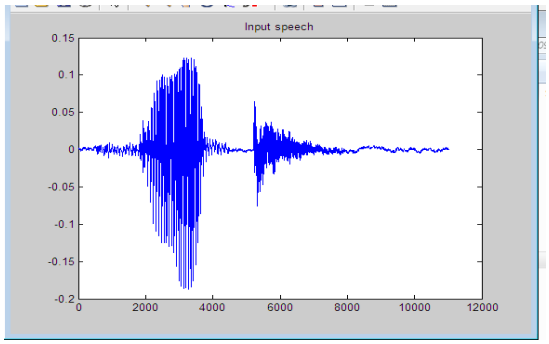


Figure3. Input speech

2. Next is pre-processing the signal, in this the first step is pre-emphazing the signal and the discontinuity present in the speech is removed using windowing technique.

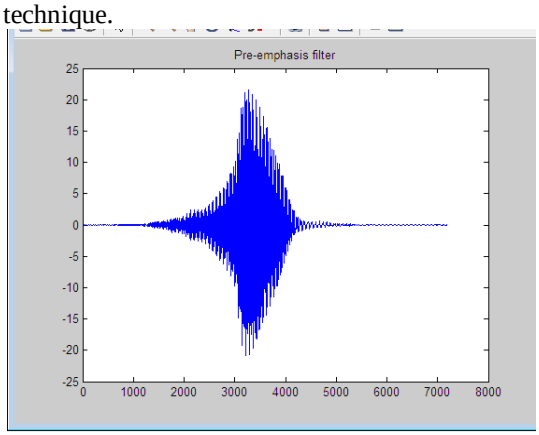


Figure4. Pre-emphasized signal

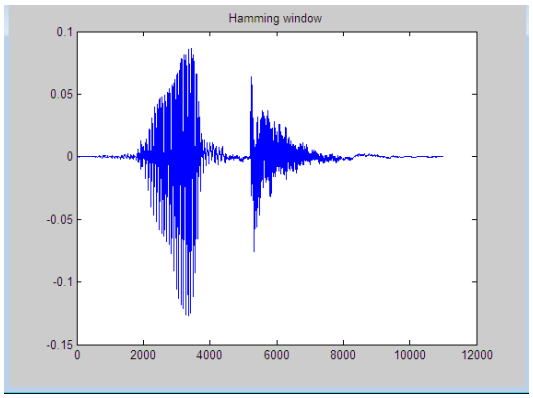


Figure5. Windowing speech signal

3. The speech region is separated and the signal is normalized and it is as shown below

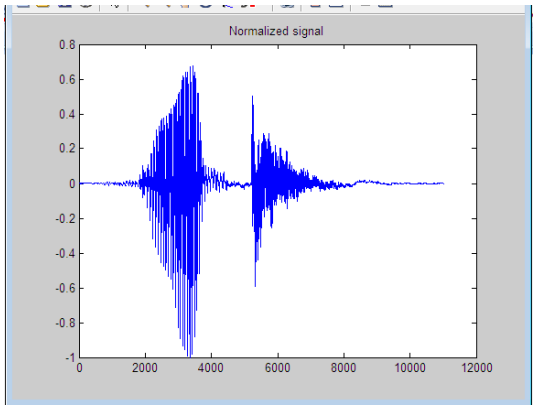


Figure6. Normalized signal

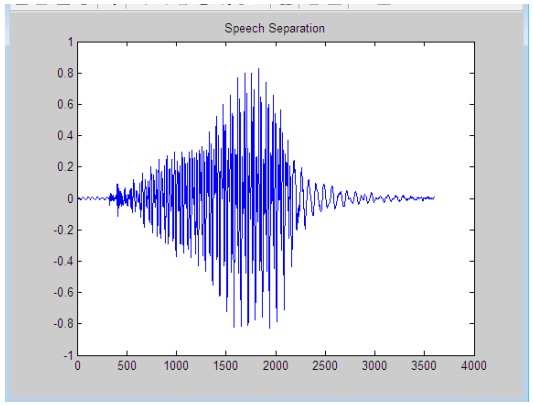


Figure7. Speech separated signal

The performance analysis of this implementation is given as follows

ISSN: 2278 – 7798

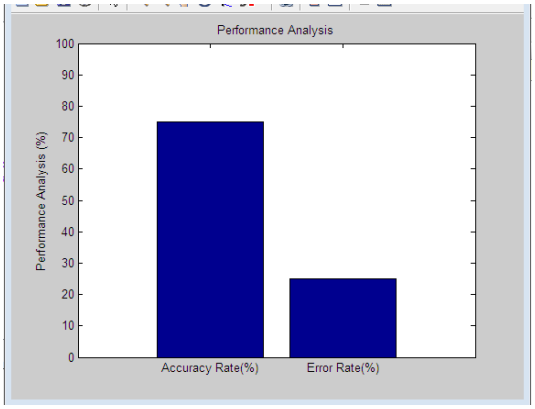


Figure8. Performance analysis based on Accuracy and Error rate.

VI. Conclusion

In this project the fast an accurate speech-recognition system is implemented. The results are tested in calm environment and would provide superior performance in noisy environment. Hence because of this feature WT has the capability to remove unnecessary voice. Hence this technique has faster process time. The result shows that the wavelet transform can be effectively used for feature extraction for speaker independent word recognition. The performances measured in this implementation are accuracy and error rate. The accuracy obtained in our project is about 95% which exceeds more than that of the previous system and clear speech signal is obtained with less distortion. The obtained result shows, it not only reduces the complexity of noise but also enhances the recognition rate. Thus the recognition rate is higher when the features are extracted accurately.

REFERENCES

[1] Balawant A sonkamble, D.D.Doye, "An overview of speech recognition system based on support vector machines",May 2008.

[2] S. Malik and Fayyaz Afsar, "Wavelet Transform Based Automatic Speaker Recognition", information of speech, IEEE, 2009.

[3] Ge zhang, Qian Liu, Chao Yang, Yin, "The Fixed point optimization of Mel frequency cepstrum coefficients for Speaker recognition", IEEE, 2011.

[4] A. Shafik, S. M. Elhalafawy, S.M. Diab, B. M. Sallam, and F. E. Abd El-samie, "A wavelet Based approach for speaker identification from Degraded speech", IJCNIS, Egypt, Vol. 1, pp. 52-58, December 2009.

[5] Tariq Abu Hilal, Hasan Abu Hilal, Riyad El Shalabi and Khalid Daqrouq, "Speaker verification system using Discrete wavelet transform and Formants Extraction Based on the Correlation Coefficient", IMECS, March 16-18, 2011, Hong Kong.

[6] Cutajar, Gatt, Greech, "A Study on pitch Variation on the use of DWT with SVM for Speaker independent Phoneme recognition", May 2012

[7] Ching-Tang Hsieh, Eugene Lai and You-Chuang wang, "Robust speaker identification System based on Wavelet transform and Gaussian mixture model", Journal of information Science and Engineering, 2003.

[8] Sadkhan, Abdul muhsen, Al-Tahan, "A proposed analog speech scrambler based on parallel structure of wavelet transform", IEEE, 2007.

[9] Daqrouq.K, Abu Hilal.T, Sherif.M, "Speaker identification system using wavelet Transform and neural network", IEEE, 2009.

[10] Gaajar.T.S, Abo Bakr.H.M, Abdalla.M, "An improvement method for speech/speaker Recognition", IEEE, May 2014.

[11] Sun Hung Liw, Ka Fei, Thang, "Development of intelligent-speech recognition system using wavelet transform and Neural network", ISBN, 2014.

[12] Nitin Trivedi, Dr. Vikesh Kumar, Saurabh Singh, Sachin Ahuja, Raman Chadha,"Speech Recognition by wavelet analysis", IJCA, volume 15-No.8, February 2011.

ISSN: 2278 – 7798