International Journal of Advance Research in Science and Engineering
Vol. No. 5, Issue No. 08 , August 2016
www.ijarse.com
IJARSE
ISSN 2319 - 8354

SPEAKER RECOGNITION USING MFCC AND DELTA-DELTA MFCC AND CLASSIFICATION USING ARTIFICIAL NEURAL NETWORK

Alka Singh[1], Surekha Ghangas[2]

[1,2]*Department of Electronics & Communication Engineering,*
ABES Engineering College Ghaziabad, U.P. (India)

ABSTRACT

Speaker recognition is one of the biometric security system that uses voice as identification. It is a process of establishing identity of an individual based on his/ her voice. Basically Speaker Recognition is achieved by two stage signal processing. One is training and other is testing. First stage that is training calculates feature parameters of speaker from the speech. Statistical models of different speakers are generated by these features. MFCC and DELTA-MFCC and DELTA-DELTA-MFCC are used to extract the features of speakers. In other stage that is testing phase, speech samples from unknown speaker are compared with the models and after then classification is done.

Keywords: Artificial-Neural-Network, Delta-Mfcc, Delta-Delta-Mfcc , Speaker Recognition , Svm.

I. INTRODUCTION

Recognition of speaker is a process which allows a secure method of authenticating speaker. It automatically recognizes who is speaking on the basis of one's information included in speech waves. It may be defined as the most popular biometric system because of easy implementation and economical hardware. Recognition of speaker basically has two categories:

1.1 Speaker Identification

1.2 Speaker Verification

Speaker Identification includes performing multiple decisions and comparing the voice of person speaking to a database reference templates on an attempt to identify the speaker whereas speaker verification includes performing binary decision and verifying their identity .This technique is very useful because it makes possible to use speaker's voice to verify their identity and control access to service such as voice mail ,confidential information areas, banking by telephone, telephone shopping, database access services etc.

II. PROCESS OF SPEAKER RECOGNITION

There are some steps which are followed by the process of speaker recognition namely as:

- Speech Pre-Processing
- Feature Extraction

- Classification

2.1 Speech Pre-Processing

It includes filtering, framing, point detection, hamming windows, speech feature, etc. of a signal before entering into the speaker recognition platform. The signal which is pre-processed is a slowly time varying signal. Basically pre-processing of signal is used to convert the speech waveform into some type of parametric representation at a considerably lower information rate.

2.1.1 Filtering

In order to reduce noise and various other external disturbances incoming digital signal needs to be filtered. To accentuate the higher frequencies pre-emphasis filter of first order is applied.

2.1.2 Framing

Framing is usually done into two types of frames: fixed size frame and dynamic size frame.

In fixed-size frame, the number of frame varies with speech speed due to different length of voice signal. We have to use dynamic size frame to obtain fixed number of frames, when we use artificial neural network .By using two methods we can get a fixed number of frames:

(1) Dynamic numbers of sample points.

(2) Dynamic overlaps rates.

Framing is a process of segmenting the speech samples into mini frames with the length in the range of (20 to 40 ms).

2.1.3 Endpoint Detection

The process which removes the silences and used to detect whether voice is present or not is called end-point detection.

2.1.4 Windowing

It is a process used to minimize the effect of spectral artifacts from the framing process. Windowing is a multiplication process which is done pointwise between the framed signal and the window function in time domain. Whereas the combination becomes the convolution between the short-term spectrum and the transfer function of the window in frequency domain . Function which has a narrow main lobe and low side lobe levels in their transfer function is considered as good window function. The windows that are commonly used during the frequency analysis of speech sounds are Hamming and Hanning window.

2.2 Feature Extraction

There are various techniques that are used for extracting the features of speech samples like LPC, MFCC, and LDB etc.

2.2.1 LPC

It stands for Linear Productive Coding and it is a process of analyzing the speech signals by calculating the formants and after then effects are removed from the speech signals and then frequency of the remaining waves

is estimated. Removal of formants is called inverse filtering and the remaining signal is considered as residue. In this type of feature extraction method each and every sample of signal is expressed as linear combination of the previous samples. That equation is termed as linear predictor and hence it is called Linear Productive Coding.

2.2.2 LDB

LDB is a feature extraction method of an audio type and regarded as multi group classification schemes which focus on identifying discriminatory time-frequency sub spaces. By using the two dissimilarity measures LDB nodes are selected and then features are extracted from them. These features which are extracted are fed to a linear discriminant analysis based classifier for a multi level hierarchical classification audio signals.

2.2.3 MFCC

MFCC performs acoustic analysis that represents the ear model which proves good result in recognition of speaker in the case when the high number of coefficients are used.

Because of the capacity of the ear model MFCC used to extract features as it operates in separated mode. By using this we can sometimes recognize a person from his voice without understanding what he is speaking about. We will use MFCC or Delta MFCC or Delta Delta MFCC to find out the features of individual speaker and then to find out the speaker recognition rate on the basis of classification.

III. MFCC

MFCC stands Mel Frequency Cepstral Coefficients. It is very useful in speaker recognition process because it works upon the human peripheral auditory system. MFCC possess human perception sensitivity with respect to frequencies and hence it is best for speaker recognition. Step by step competition of MFCC can be explained as:

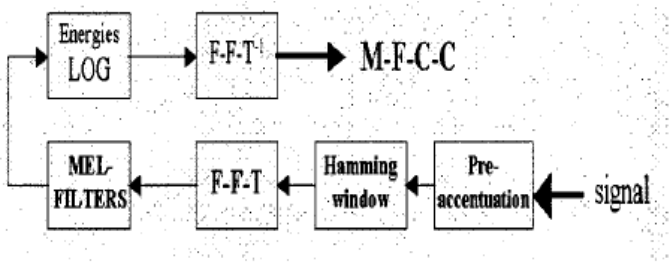
MFCC coefficients are found out by taking the log magnitude of the windowed waveform and then these waveforms are smoothened out by using triangular filters and then DCT of waveform is computed to generate the MFCC coefficients.

First of all speech signal s(n) is sent to a high pass filter:

$$S(n) = s(n) - a*s(n-1)$$

Where, S (n) is the output signal and the value is usually between 0.9 and 1.0.

Pre- emphasis is a process in which the high frequency part is compensated which was sub pressed during the sound production mechanism of humans. It is also used to amplify the importance of high frequency formants.



The next step is frame blocking which is used to divide the input speech signal into small frames. Basically the size of the frame is equal power of two in order to facilitate the use of FFT. In case, zero padding is done to the nearest length of power of two.

Now each frames has to be multiplied in order prevent the discontinuity in each and every frame and at both ends of every frame. For this hamming window will be used. Each frame will multiplied with hamming window to keep continuity. If the signal in the frame is denoted as s(n), n=0, 1….N-1, then the output signal that we get after applying hamming window is s(n)*w(n), where w(n) is the hamming window.

Now FFT will be performed to obtain the magnitude frequency response of each frame and this magnitude response of frequency is multiplied by a set of triangular filters to get the log energies of each triangular filter. Position of these filters is equally spaced along the Mel frequency:

$$mel(f) = 1125*In\ (1+f/700)$$

Finally DCT is applied on the log filter bank energies to get mel-scale cepstral coefficients. For better performance, we can add the log energy and perform Delta operation and Delta Delta operation.

3.1 Delta MFCC

By taking first derivative of MFCC features we can extract the Delta MFCC features. Delta featuresare used to represent the related Delta features to the change in the cepstral features with time. Each of the delta feature extracted as the first derivative of MFCC feature represents the change between frames. The only one benefit of Delta features over MFCC features is that they are used to represent the temporal information. One common technique allowing to differentiate crossing trajectories are delta features. This technique adds an approximation of the first time derivatives of basic features (for example MFCCs) to the feature vector.

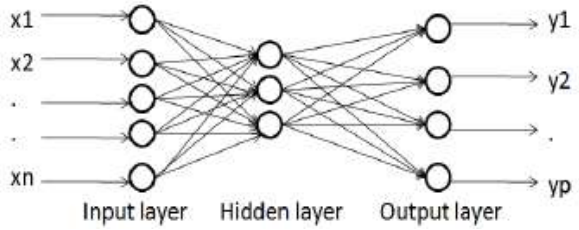
3.2 Delta Delta MFCC

By taking the derivative of Delta features,Delta-Delta features are extracted. They used to show the change between frames in the corresponding delta features. The new derivatives are called delta-delta features. These are also called as acceleration coefficients. Delta-delta features are also known to introduce even longer temporal context. If the window has also 7 frames, the temporal context is 11 frames. Delta-delta features can say whether there is a peak or a valley on the investigated part of trajectory.

IV. CLASSIFICATION

After extracting features of different voice samples classification is done. Classification with the help of Artificial neural networks proved as a promising approach for the problem of speaker recognition. Artificial Neural Network includes some particular properties for example the ability to adapt or to learn, to generalize, or to cluster or to organize data, and its operation is based on parallel processing that are needed for speech and speaker recognition.

Due to the non-linear structure of artificial neuron, it can be used to learn complex features from the data samples.



V. RESULTS

Comparisions Of Implementations Of Mfcc

- Number of Centroid :- 10
- Sampling Rate :- 8000
- Number of Filters:- 20

Training Samples :- 30 Users (Each user have 8 samples)

Testing Samples:- 30 Users (Each user have 2 samples)

TECHNIQUES	RECOGNITION RATE
MFCC	96.5%
DELTA MFCC	97%
DELTA-DELTA MFCC	97.6%

VI. CONCLUSION

The results that are obtained using MFCC and DELTA-MFCC for each speaker are computed and classified for efficient representation. Accuracy obtained using artificial neural network clearly indicates its high efficiency. Use of more number of centroids increases the performance factor but degrades the computational efficiency. Modern speaker recognition possess high accuracy at low complexity and easy calculation and produce more compatible results using MFCC and DELTA-MFCC and DELTA-DELTA-MFCC.

REFERENCES

[1] K.K. Paliwal and B.S. Atal, 'Frequency related representation of speech,' in *Proc. EUROSPEECH*, p.p.65-68 Sep. (2003).

[2] Deller, J.R., Hansen, J.H.L., Proakis, J.G. (2000). Discrete-Time . Processing of Speech Signals, Institute of Electrical and Electronics Engineers, Inc., pp-621

[3] J. P. Campbell, JR, "Speaker Recognition: A Tutorial", Proceedings of the IEEE, Vol. 85, No. 9, September 1997.

[4] A.E. Rosenberg, ―Automatic speaker verification: A review,‖ Proc IEEE, vol. 64(4), pp. 475-87, Apr. 1976.

[5] H. Gish, and M.Schmidt, ―Text-indepent speaker identification,‖ IEEE Signal Process. Mag., vol. 18, pp.18-32, Oct. 2002.

[6] Sirko Molau, Michael Pitz, Ralf Schlu¨ ter, and Hermann Ney, ― Computing Mel-Frequency Cepstral Coefficients on the Power Spectrum‖ Lehrstuhl fu¨ r Informatik VI, Computer Science Department, RWTH Aachen – University of Technology, 52056 Aachen, Germany.

[7] T. Kohonen, ―Self-organization and Associative Memory‖ Springer-Verlag, Berlin- New York, 1988a.

[8] Robinson, A. **1.** (1994). The application of recurrent

[9] nets tophone probability estimation. IEEE Transactions on Neural Net-works,vol.S no.2, March 1994

[10] D.O. Shaughnessy, ―Speaker Recognition‖, ASSP Magazine, IEEE Signal Processing Magazine, Vol. 3, No. 4, Part. 1, pp. 4-17, October 1986.