Digital Technologies, 2014, Vol. 1, No. 1, 1-3
Available online at http://pubs.sciepub.com/dt/1/1/1

DOI:10.12691/dt-1-1-1

Mel-Frequency Cepstral Coefficient (MFCC) - a Novel Method for Speaker Recognition

Asutosh das[1], Manas Ranjan Jena[2,*], Kalyan Kumar Barik[2]

[1]Department. of ETC, SIET, Odisha
[2]Department. of ELTCE, VSSUT, BURLA, ODISHA
*Corresponding author: manas.synergy@gmail.com

Received May 28, 2014; Revised June 16, 2014; Accepted August 12, 2014

Abstract The purpose of this paper is to develop a speaker recognition system which can recognize speakers from their speech. The proposed system would be text dependent speaker recognition system means the user has to speak from a set of spoken words. Mel. Frequency cepstral coefficient is used in order to extract the features of speakers from their speech signal while VQ (LBG) is used for design of codebook from extracted features. In pattern matching we derive the VQ distortion between the utterances of unknown speaker to codebooks of known speaker. We have used Euclidean distance to compute VQ distortion. The system is implemented by using TIMIT database with 630 speakers having 10 speech files each. In our project we have chosen 30 speakers as well as 100 speakers from this database. The comparison of speaker recognition performance between 30 speakers and 100 speakers are also discussed.

Keywords: *ASV, ASI, LPC, Mel, DFT, LPCC, MFCC*

Cite This Article: Asutosh das, Manas Ranjan Jena, and Kalyan Kumar Barik, "Mel-Frequency Cepstral Coefficient (MFCC) - a Novel Method for Speaker Recognition." *Digital Technologies*, vol. 1, no. 1 (2014): 1-3. doi: 10.12691/dt-1-1-1.

1. Introduction

Speech is the medium of communication between people. An acoustic speech signal contains a variety of information. It contains textual message as well as information from which we can identify whether the person is male or female, adult or child. So speech plays a vital role in speaker recognition system.

Speaker recognition is the process of recognising a speaker from their speech. Speaker recognition focused on the unique characteristic of person. So the uniqueness of individual's voice is a result of both physical feature of person's vocal tract and learned speaking habits of different individuals. From these we can discriminate between speakers [1].

In speaker recognition the system cannot measure easily the physical feature of vocal tract of an unknown person. Thus numerical values of physical feature of person have to be derived from the parameter those are extracted from speech signal.

Speaker recognition encompasses verification and identification. In automatic speaker verification (ASV) the system verifies identity of the claimed speaker from their speech. Similarly in automatic speaker identification (ASI) the system identifies the claimed speaker. Speaker recognition applications are widely used in forensic, police work, telephone banking.

The purpose of our project is to build a speaker recognition system which will identify the speaker by using their speech [2].

2. Principles of Speaker Recognition

Here our main objective is based on speaker recognition speaker recognition is a process in which it recognizes the speaker by using their speech. Speaker recognition is classified into two streams namely speaker verification and speaker identification.

Speaker recognition is often classified into closed-set recognition and open-set recognition. Just as their names suggest, the closed-set refers to the cases that the unknown voice must come from a set of known speakers; and the open-set means unknown voice may come from unregistered speakers

Speaker recognition is divided into two types according to the speech modalities: text dependent recognition and text independent recognition. For text-dependent speaker recognition system speakers are only allowed to say some specific sentences or words, which are known to the system. Where as in the text-independent speaker recognition system they could process freely spoken speech, which is either user selected phrase or conversational speech [2,3].

3. Feature Extraction

The most important thing in speaker recognition project is the feature extraction where it extracts features from speech signal.

Feature extraction is a process in which it transforms the input data into set of features is called feature extraction. In feature extraction it reduces the dimension of the input vector while retains the important discriminating feature of a speaker.

Every person has a natural sound quality due to their voice speech. So pitch is not always reliable since reliance on pitch can allow imposters to gain access by changing their own pitch. Therefore many feature extraction algorithm do not use pitch as feature instead find speaker specific information in speech [4].

The most commonly used system-based features are the cepstral coefficient. The two types of cepstral coefficient that are widely used

1. Linear predictive cepstral coefficient(LPCC).
2. Mel frequency cepstral coefficient(MFCC).

3.1. Linear Predictive Coding

Linear Predictive Coding (LPC) is a well known feature extraction technique for both speech recognition and speaker identification. LPC is based on the source-filter model of speech production. The main idea behind LPC is that a given speech sample can be approximated as a linear combination of the past speech samples. LPC models signal s (n) as a linear combination of its past values and present input (vocal cords excitation). If the signal will be represented only in terms of the linear combination of the past values then the difference between real and predicted output is called prediction error. LPC minimizes the prediction error to find out the coefficients. In practice, the prediction order is set to 12-20 coefficients, depending on the sampling rate and the number of poles in the model. Thus, the basic problem in LPC analysis is to determine prediction coefficients from the speech frame [5].

There are two main approaches to derive them, i.e., the least square method and the lattice method. The classical least-square method selects prediction coefficients to minimize the mean energy in prediction error of a speech frame. Examples of this method are autocorrelation and covariance methods. The other approach is known as lattice, permits instantaneous updating of the coefficients. In other words, LPC parameters are determined sample by sample. This method is more suitable for real-time application. In speaker recognition area the set of prediction coefficients is usually converted to the so-called Linear Predictive Cepstral Coefficients (LPCC), because the cepstrum is proved to be the most effective representation of speech signal for speaker recognition. LPC models speech signal s(n) approximately as a linear combination of previous p samples [6].

3.2. MFCC

In MFCC frequency bands are positioned logarithmically whereas in FT frequency bands are positioned linearly. As the frequency bands are positioned logarithmically in MFCC it approximates the human system response more closely than any other system [7].

In order to obtain MFCC coefficient the input speech signal is windowed and taken Discreet Fourier transform to convert into frequency domain. Here we are using bank filter to wrapping the mel frequency. And then a log magnitude of each of the mel frequency is acquired. Then the resultant signal is transformed using an inverse DFT into cepstral domain. The lower order coefficients are selected as the feature vector to avoid higher coefficients since it contains less specific information about speaker. Then the coefficients are uniformly spaced and used as output feature vector for that speech frame [8].

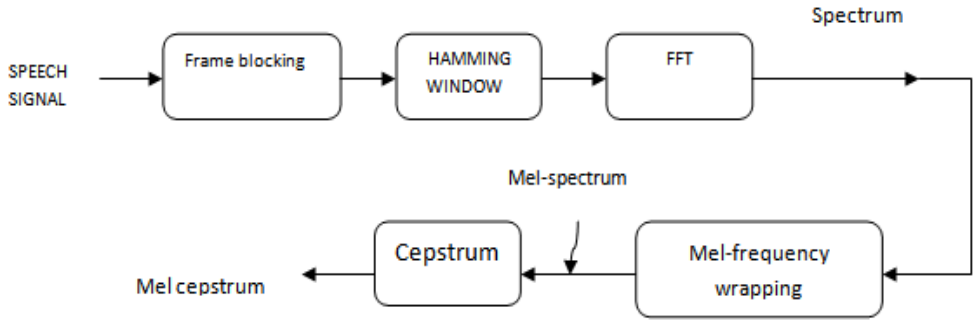

Figure 3.1. MFCC processor

A *Mel* is a unit of measure based on the human ear's perceived frequency. First 1 KHz is defined as 1000 mels as a reference. Secondly listeners are asked to change the physical frequency until they perceive it is twice of the reference, or 10 times or half or one tenth of the reference, and so on. Thirdly those frequencies are then labelled as 2000 mels, 10,000 mels, 500 mels, 100 mels, and so on.

The mel scale is approximately a linear frequency spacing below 1000 Hz, and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone 40 dB above the perceptual hearing threshold, is defined as 1000 mels [9].

Hence the approximation of Mel from frequency can be expressed as:

$$mel(f) = 2595 \times \log_{10} \log_{10}(1 + \frac{f}{700}) \qquad (3.1)$$

Where 'f' denotes the real frequency, and mel(*f*) denotes the perceived frequency.

4. Simulation Results & Analysis

We have performed the simulation using MATLAB. Performance of speaker recognition system improves.

4.1. Speaker Recognition Performance for 100 Speakers

When MFCC algorithm is being employed and respective speaker recognition performance for different code book size is given in the Table 1. From the Table 1, we can notice our performance of system improves further and further with increment of code book size.

Table 1. Speaker Recognition Performance for 100 Speakers at different codebook size.

Codebook size	16	32	64	128
Performance (%)	93	97.5	98.5	98.5

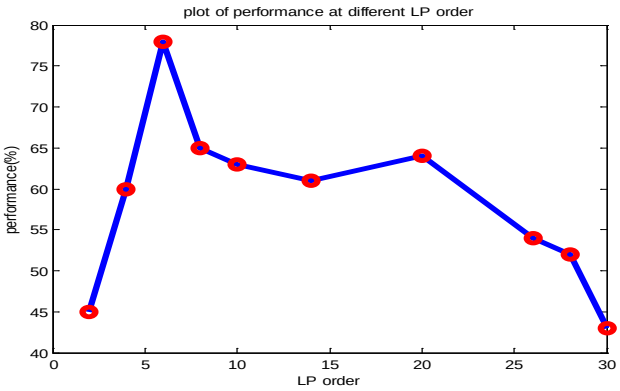

Figure 4.1. plot of performance at different LP order for set of 50 speakers

4.2. Speaker Recognition Performance for 50 Speakers Using LP Residual

LP is used to estimate the vocal tract information. Then vocal tract information is removed from speech signal and the resulting signal we get is known as LP residual. First we have chosen 50 speakers from TIMIT database. We have implemented LP using different orders to find out proper order in which it provides better performance. LP analysis is performed on the speech signal using 6th order prediction as it provides 78% performance. We can get same performance on some higher order but the problem is as we go for higher order it increases the complexity hence 6th order is being used in this system. In the below figure we have provided the performance of 50 speakers at different orders.

4.3. Speaker Recognition Performance Using LP Residual at Different Codebook Size

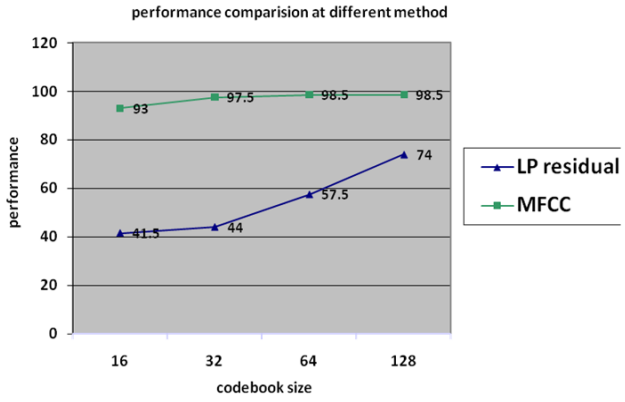

Figure 4.2. Performance comparison of MFCC with LP residual

Speaker recognition performance for a set of 100 speakers using linear prediction residual is given below. Thus we concluded that at codebook size of 128 & at 6th order speaker recognition performance is 74%, which is better performance among other performances.

Table 2. speaker recognition performances for set of 100 speakers using lp residual

Codebook size	16	32	64	128
Performance	41.5	44	57.5	74

Table 3. speaker recognition performance comparisons at different algorithm

model	codebook size			
	16	32	64	128
LP residual	41.5	44	57.5	74
MFCC	93	97.5	98.5	98.5

5. Conclusion

In this paper the MFCC algorithm is successfully employed to extract features from the speech. When MFCC is employed our performance of system improves further and further with increment of code book size. At the same time, we have employed linear prediction residual algorithm & found respective performance at different codebook size. By comparing these two algorithms it is found that performance of MFCC is superior & thus can be suggested for appreciable speaker reorganization system.

Acknowledgement

The authors sincerely thank to the H.O.D & all the staff of Dept. of ETC, SIET, DHENKANAL, ODISHA for constant encouragement and support directly or indirectly.

References

[1] L.R. Rabiner and R.W. Schafer, *"Digital Processing of Speech Signals"*, New Delhi: Prentice Hall of India. 2006.

[2] J. P. Campbell, JR., *"Speaker Recognition: A Tutorial"*, Proceedings of the IEEE, vol. 85, no. 9, pp. 1437-1462, Sep 1997. J.M.Naik, *"speaker verification: A Tutorial"*, IEEE Communication Magazine, pp.42-48, January 1990.

[3] J.M.Naik,*"speaker verification: A Tutorial"* ,IEEE Communication Magazine, pp.42-48, January 1990.

[4] Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani Md. Saifur Rahman, *"Speaker identification using mel frequency cepstral coefficients"* **3**rd International Conference on Electrical & Computer Engineering ICECE 2004, 28-30 December 2004, Dhaka, Bangladesh.

[5] John G. Proakis and Dimitris G. Manolakis, *"Digital Signal Processing"*, New Delhi:Prentice Hall of India. 2002.

[6] John G. Proakis and Dimitris G. Manolakis, *"Digital Signal Processing"*, New Delhi:Prentice Hall of India. 2002.

[7] Jr. J.D. Hansen, J. & Proakis, J. *"Discrete time processing of speech signal"*, 2nd edition, IEEE press, Newwork, 2000.

[8] Y. Linde, A. Buzo & R. Gray, *"An algorithm for vector quantizer design"*, IEEE Transactions on Communications, Vol. 28, issue 1, Jan 1980 pp. 84-95.

[9] Seddik, H.; Rahmouni, A.; Sayadi, M.; *"Text independent speaker recognition using the Mel frequency cepstral coefficients and a neural network classifier"* First International Symposium on Control, Communications and Signal Processing, Proceedings of IEEE 2004 Page(s): 631-634.