International Journal on Emerging Technologies **1**(1): 19-22(2010) **ISSN : 0975-8364**

MFCC and its applications in speaker recognition

Vibha Tiwari

Deptt. of Electronics Engg., Gyan Ganga Institute of Technology and Management, Bhopal, (MP) INDIA

(Received 5 Nov., 2009, Accepted 10 Feb., 2010)

ABSTRACT : Speech processing is emerged as one of the important application area of digital signal processing. Various fields for research in speech processing are speech recognition, speaker recognition, speech synthesis, speech coding etc. The objective of automatic speaker recognition is to extract, characterize and recognize the information about speaker identity. Feature extraction is the first step for speaker recognition. Many algorithms are suggested/developed by the researchers for feature extraction. In this work, the Mel Frequency Cepstrum Coefficient (MFCC) feature has been used for designing a text dependent speaker identification system. Some modifications to the existing technique of MFCC for feature extraction are also suggested to improve the speaker recognition efficiency.

Keywords : Feature extraction, Mel frequency cepstral coefficients (MFCC), Speaker recognition

I. INTRODUCTION

The human speech contains numerous discriminative features that can be used to identify speakers. Speech contains significant energy from zero frequency up to around 5 kHz. The objective of automatic speaker recognition is to extract, characterize and recognize the information about speaker identity. The property of speech signal changes markedly as a function of time. To study the spectral properties of speech signal the concept of time varying Fourier representation is used. However, the temporal properties of speech signal such, as energy, zero crossing, correlation etc are assumed constant over a short period. That is its characteristics are short-time stationary .Therefore, using hamming window, Speech signal is divided into a number of blocks of short duration so that normal Fourier transform can be used.

In this work, the Mel frequency Cepstrum Coefficient (MFCC) feature has been used for designing a text dependent speaker identification system. The extracted speech features (MFCC's) of a speaker are quantized to a number of centroids using vector quantization algorithm. These centroids constitute the codebook of that speaker. MFCC's are calculated in training phase and again in testing phase. Speakers uttered same words once in a training session and once in a testing session later. The Euclidean distance between the MFCC's of each speaker in training phase to the centroids of individual speaker in testing phase is measured and the speaker is identified according to the minimum Euclidean distance. The code is developed in the MATLAB environment and performs the identification satisfactorily.

II. SPEAKER RECOGNITION

Anatomical structure of the vocal tract is unique for every person and hence the voice information available in the speech signal can be used to identify the speaker. Recognizing a person by her/his voice is known as speaker recognition. Since differences in the anatomical structure are an intrinsic property of the speaker, voice comes under the category of biometric identity. Using voice for identity has several advantages. One of the major advantages is remote person authentication.

Like any other pattern recognition systems, speaker recognition systems also involve two phases namely, *training and testing*. Training is the process of familiarizing the system with the voice characteristics of the speakers registering. Testing is the actual recognition task. The block diagram of training phase is shown in Fig.1. Feature vectors representing the voice characteristics of the speaker are extracted from the training utterances and are used for building the reference models. During testing, similar feature vectors are extracted from the test utterance, and the degree of their match with the reference is obtained using some matching technique. The level of match is used to arrive at the decision. The block diagram of the testing phase is given in Fig.1.

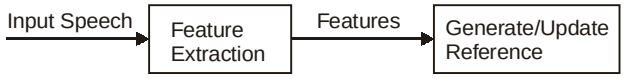

Fig.1. The block diagram of training phase.

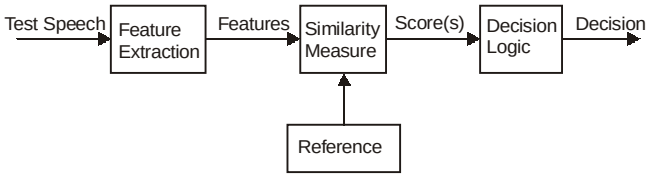

Fig.2. The block diagram of the testing phase.

A. Feature selection and measures

To apply mathematical tools without loss of generality, the speech signal can be represented by a sequence of feature vectors. The selection of appropriate features along with methods to estimate (extract or measure) them is known as feature selection and feature extraction.

Pattern-recognition models are divided into three components: feature extraction and selection, pattern matching, and classification. In speaker verification, the goal is to design a system that minimizes the probability of verification errors. Thus, the objective is to discriminate between the given speaker and all others.

B. Speaker recognition techniques

Speaker recognition concentrates on the identification task. The aim in speaker identification (SI) is to recognize the unknown speaker from a set of known speakers (closed-set SI).

A speaker recognition system is composed of the following modules:

1. Front-end processing - the "signal processing" part, which converts the sampled speech signal into set of feature vectors, which characterize the properties of speech that can separate different speakers. Front-end processing is performed both in training and testing phases.
2. Speaker modeling - this part performs a reduction of feature data by modeling the distributions of the feature vectors.
3. Speaker database - the speaker models are stored here.
4. Decision logic - makes the final decision about the identity of the speaker by comparing unknown feature vectors to all models in the database and selecting the best matching model.

III. TECHNIQUES OF FEATURE EXTRACTION

The general methodology of audio classification involves extracting discriminatory features from the audio data and feeding them to a pattern classifier. Different approaches and various kinds of audio features were proposed with varying success rates. The features can be extracted either directly from the time domain signal or from a transformation domain depending upon the choice of the signal analysis approach. Some of the audio features that have been successfully used for audio classification include Mel-frequency cepstral coefficients (MFCC), Linear predictive coding (LPC), Local discriminant bases (LDB). Few techniques generate a pattern from the features and use it for classification by the degree of correlation. Few other techniques use the numerical values of the features coupled to statistical classification method.

A. LPC

LPC (Linear Predictive coding) analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal is called the residue. In LPC system, each sample of the signal is expressed as a linear combination of the previous samples. This equation is called a linear predictor and hence it is called as linear predictive coding .The coefficients of the difference equation (the prediction coefficients) characterize the formants.

B. MFCC

MFCC is based on the human peripheral auditory system. The human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency *t* measured in Hz, a subjective pitch is measured on a scale called the 'Mel Scale' .The mel frequency scale is a linear frequency spacing below 1000 Hz and logarithmic spacing above 1kHz.As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 Mels.

C. LDB

LDB is an audio feature extraction and a multi group classification scheme that focuses on identifying discriminatory time-frequency subspaces. Two dissimilarity measures are used in the process of selecting the LDB nodes and extracting features from them. The extracted features are then fed to a linear discriminant analysis based classifier for a multi-level hierarchical classification of audio signals.

IV. MEL FREQUENCY CEPSTRAL COEFFICIENTS

The extraction and selection of the best parametric representation of acoustic signals is an important task in the design of any speech recognition system; it significantly affects the recognition performance. A compact representation would be provided by a set of mel-frequency cepstrum coefficients (MFCC), which are the results of a cosine transform of the real logarithm of the short-term energy spectrum expressed on a mel-frequency scale. The MFCCs are proved more efficient. The calculation of the MFCC includes the following steps.

A. Mel-frequency wrapping

Human perception of frequency contents of sounds for speech signal does not follow a linear scale. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale. The mel-frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000Hz .As a reference point ,the pitch of a 1 KHz tone ,40dB above the perceptual hearing threshold, is defined as 1000 mels. Therefore we can use the following approximate formula to compute the mels for a given frequency *f* in Hz.

$$\text{Mel}(f) = 2595*\log10(1 + f/700)$$

Ours approach to simulate the subjective spectrum is to use a filter bank, one filter for each desired mel-frequency component. That filter bank has a triangular band pass frequency response and the spacing as well as the bandwidth is determined by a constant mel-frequency interval. The mel scale filter bank is a series of l triangular band pass filters that have been designed to simulate the

band pass filtering believed to occur in the auditory system. This corresponds to series of band pass filters with constant bandwidth and spacing on a mel frequency scale.

B. Cepstrum

In this final step, we convert the log mel spectrum back to time. The result is called the Mel Frequency Cepstrum Coefficients (MFCC).The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the discrete cosine transform (DCT). In this final step log mel spectrum is converted back to time. The result is called the Mel Frequency Cepstrum Coefficients (MFCC).The discrete cosine transform is done for transforming the mel coefficients back to time domain.

$$C_n = \sum_{k=1}^{k} (\log S_k) \cos \left\{ n\left(k - \frac{1}{2}\right) * \frac{\pi}{k} \right\},$$

$$n = 1, 2, \ldots k$$

Whereas S_k, $K = 1, 2, \ldots K$ are the outputs of last step. Complete process for the calculation of MFCC is shown in

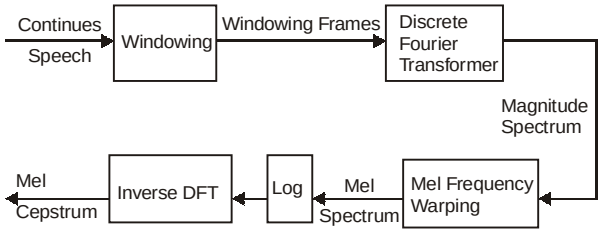

Fig.3. Complete pipeline for MFCC.

V. VECTOR QUANTIZATION

Vector Quantization is the classical quantization technique from signal processing which allows the modeling of probability density functions by the distribution of prototype vectors. It works by dividing a large set of points into groups having approximately the same number of points closest to them. Each group is represented by its centroid point. The density matching property of vector quantization is powerful, especially for identifying the density of large and high-dimensioned data. Since data points are represented by the index of their closest centroid, commonly occurring data have low error, and rare data high error. Hence, Vector Quantization is also suitable for lossy data compression.

A vector quantizer maps k-dimensional vectors in the vector space R^k into a finite set of vectors $Y = \{y_i : i = 1, 2, ..., N\}$. Each vector y_i is called a code vector or a codeword and the set of all the code words is called a codebook. Associated with each codeword, y_i, is a nearest neighbor region called Voronoi region, and it is defined by : $V_i = \{x \in R^k : \| x - y_i \| < \| x - y_j \|, \text{ for all } j \neq 1\}$. Given an input vector, the codeword that is chosen to represent it is the one in the same Voronoi region.

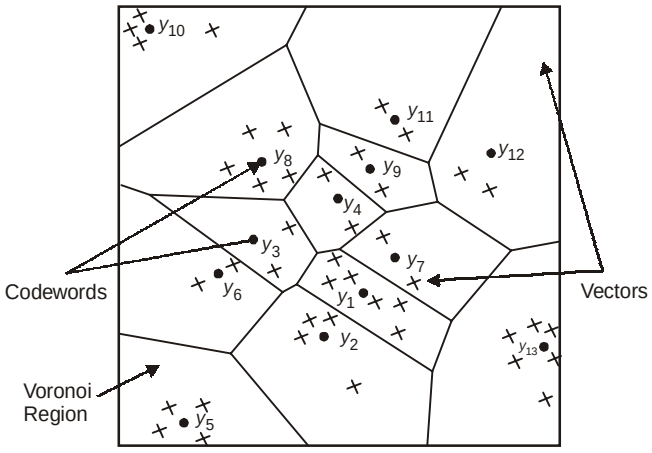

Fig.4. Codewords in 2-dimensional space. Input vectors are marked with an x, codewords are marked with circles, and the Voronoi regions are separated with boundary lines.

The representative codeword is determined to be the closest in Euclidean distance from the input vector. The Euclidean distance is defined by:

$$d(x, y_i) = \sqrt{\sum_{j=1}^{k} (x_j - y_{ij})^2}$$

where x_j is the jth component of the input vector, and y_{ij} is the j^{th} is component of the codeword y_i.

VI. COMPARISON OF DIFFERENT IMPLEMENTATIONS OF MFCC

The performance of the Mel-Frequency Cepstrum Coefficients (MFCC) may be affected by (1) the number of filters,(2) type of window.In this paper, several comparison experiments are done to find a best implementation.

A. Effect of number of filters

Results of the speaker recognition performance by varying the number of filters of MFCC to 12, 22, 32, and 42 are given. *The recognizer reaches the maximal performance at the filter nu*mber $K = 32$. Too few or two many filters do not result in better accuracy. Hereafter, if not specifically stated, the number of filters is chosen to be $K = 32$.

MFCC with 12 filters

Speaker	No. of Attempt	False Acceptance	False Rejection
*S*1	4	0	0
*S*2	4	0	1
*S*3	4	0	2
*S*4	4	0	0
*S*5	4	0	2
Total	**20**	**0**	**5**

Threshold value of distance = 130, Efficiency = 75%

MFCC with 22 filters

Speaker	No. of Attempt	False Acceptance	False Rejection
*S*1	4	0	0
*S*2	4	0	2
*S*3	4	0	2
*S*4	4	0	0
*S*5	4	0	3
Total	**20**	**0**	**7**

Threshold value of distance = 150, Efficiency = 65%

MFCC with 32 filters

Speaker	No. of Attempt	False Acceptance	False Rejection
*S*1	4	0	0
*S*2	4	0	0
*S*3	4	0	1
*S*4	4	0	0
*S*5	4	0	2
Total	**20**	**0**	**3**

Threshold value of distance = 150, Efficiency = 85%

MFCC with 42 filters

Speaker	No. of Attempt	False Acceptance	False Rejection
*S*1	4	0	0
*S*2	4	0	0
*S*3	4	0	2
*S*4	4	0	1
*S*5	4	0	1
Total	**20**	**0**	**4**

Threshold value of distance = 85, Efficiency = 80%

B. Effect of variation in type of window using 32 filters

Considering 32 filters as a standard number of filters we have changed the window type. In this experiment we have used two windows *viz*. Hanning Window and Rectangular window. Results show that efficiency is maximum while using hanning window.

(*a*) Hanning window

Speaker	No. of Attempt	False Acceptance	False Rejection
*S*1	4	0	0
*S*2	4	0	0
*S*3	4	0	2
*S*4	4	0	0
*S*5	4	0	3
Total	**20**	**0**	**5**

Threshold value of distance = 150, Efficiency = 75%

(*b*) Rectangular window

Speaker	No. of Attempt	False Acceptance	False Rejection
*S*1	4	0	0
*S*2	4	0	4
*S*3	4	0	2
*S*4	4	0	0
*S*5	4	0	3
Total	**20**	**0**	**9**

Threshold value of distance = 150, Efficiency = 55%

VII. CONCLUSION AND FUTURE WORK

In this paper several feature extraction techniques for speaker recognition were discussed. MFCC is well known techniques used in speaker recognition to describe the signal characteristics, relative to the speaker discriminative vocal tract properties. The goal of this project was to create a speaker recognition system, and apply it to a speech of an unknown speaker. By investigating the extracted features of the unknown speech and then compare them to the stored extracted features for each different speaker in order to identify the unknown speaker. In our results we find that

Number of filters	12	22	32	42
Efficiency	65%	75%	85%	80%

Types of window using 32 filters	Efficiency
Hanning	75%
Rectangular	55%

In future we will try to improve this system to be a text independent speaker identification system.

REFERENCES

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

[2] T. Fukuda, M. Takigawa and T. Nitta, "Peripheral features for HMMbased speech recognition," in *Proc. ICASSP*, **1:** 129-132(2001).

[3] M. Pandit and J. Kittler, "Feature selection for a dtw-based speaker verification system, in *Proceedings of IEEE Int. Conf. Acoust. Speech and Signal Processing*, **2:** 769-772 (1998).

[4] M. Dash and H. Liu, "Feature Selection for Classification," Intelligent Data Analysis, **1:** 131-156(1997).

[5] S. Furui, "An overview of speaker recognition technology, in Automatic Speech and Speaker Recognition (C.H. Lee, F.K. Soong, and K.K. Paliwal,eds), ch.2 pp.31-56Boston : Kluwer Academic, (1996).

[6] L.R. Rabiner and R.W. Schafer, *Digital Processing* of *Speech Signals.* Englewood Cliffs, NJ: Prentice-Hall, (1978).

[7] Atal, B.S. and S.L. Hanauer, 'Speech analysis and synthesis by linear prediction of the speech wave', *Journal of the acoustical society of America*, **50:** 637-655(1971).

[8] Automatic speaker recognition by S.Khan, Mohd Rafibul Islam, M. Faizul, D. Doll. *3rd international conference on electrical and computer engineering* (ICECE), 28-30th Dec. (2004), Dhaka, Bangladesh.

[9] Speaker recognition using MFCC by S. Khan, Mohd Rafibul Islam, M. Faizul, D. Doll, presented in *IJCSES (International Journal of Computer Science and Engineering System)* **2**(1)**:** 2008.

[10] Speaker identification using MFCC coefficients -Mohd Rasheedur Hassan, Mustafa Zamil, Mohd Bolam Khabsani, Mohd Saifur Rehman. *3rd international conference on electrical and computer engineering* (ICECE), (2004).

[11] Goutam Saha and Malyaban Das,On Use of Singular Value Ratio Spectrum as Feature Extraction Tool in Speaker Recognition Application, CIT-2003, pp. 345-350, Bhubaneswar, Orissa, India, (2003).

[12] Premakanthan and W.B. Mikhael, Speaker verification/ recognition and the importance of selective feature extraction: Review, *Proceedings of the 44th IEEE 2001, Midwest Symposium*, **1:** 14-17(2001).

[13] Molau, S, Pitz, M, Schluter, R, and Ney, H., Computing Mel-frequency coefficients on Power Spectrum, *Proceedings of IEEE ICASSP-2001*, **1:** 73-76(2001).

[14] C.D. Bei and R.M. Gray. An improvement of the minimum distortion encoding algorithm for vector quantization. *IEEE Transactions on Communications,* October (1998).

[15] Lawrence Rabiner and Biing-Hwang Juang, *Fundamental of Speech Recognition*", Prentice-Hall, Englewood Cliffs, N.J., (1993).