Volume 6, Issue 5, May 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Application of MFCC in Text Independent Speaker Recognition

Shipra Gupta

Vedant college of Engineering and Technology, Kota,
Rajasthan, India

Abstract*— *Recently speech processing is one of the important application area of digital signal processing. There are several parts of speech processing as speech recognition, speaker recognition, speech synthesis, speech coding etc. The objective of the presented work is to extract, characterize and recognize the speaker identity. Feature extraction is the key process for speaker recognition. In this work, the Mel Frequency Cepstrum Coefficient (MFCC) feature has been utilized for designing a speaker identification system which is independent of speech rather than previously reported text dependent techniques.

Index Terms*— *Feature extraction, Mel frequency cepstral coefficients (MFCC), Speaker recognition

I. INTRODUCTION

Speech processing is emerged as one of the significant application area of digital signal processing. So many fields for research in speech processing are recently emerging like speech recognition, speaker recognition, speech synthesis, speech coding etc.. Speaker recognition is the process of recognizing automatically who is speaking on the basis of individual information included in speech waves. This technique uses the speaker's voice to verify their identity and provides control access to services such as voice dialing, database access services, information services, voice mail, security control for confidential information areas, remote access to computers and several other fields where security is the main area of concern. In this work, the Mel frequency Cepstrum Coefficient (MFCC) feature has been used for designing a text independent speaker identification system. The extracted speech features (MFCC's) of a speaker are quantized to a number of centroids using vector quantization algorithm Vector Quantization is done by using Linde-Buzo-Gray 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 code is developed in the MATLAB environment and performs the identification satisfactorily.

II. SPEAKER RECOGNITION

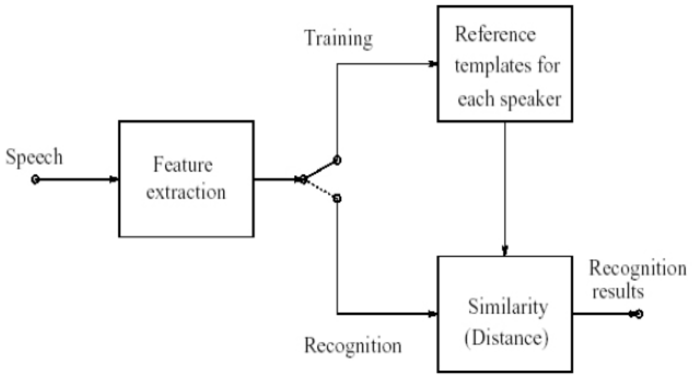
The problem of speaker recognition can be divided into two major sub problems: *speaker identification* and s*peaker verification*. Speaker identification can be thought of, as the task of determining who is talking from a set of known voices of speakers. It is the process of determining who has provided a given utterance based on the information contained in speech

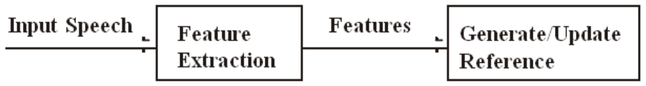
waves. The unknown voice comes from a fixed set of known speakers, thus the task is referred to as closed set identification. Anatomical structure of the vocal tract is unique for every person and hence the voice information available in the speech signal can be used to identify the speaker. Recognizing a person by her/his voice is known as speaker recognition.

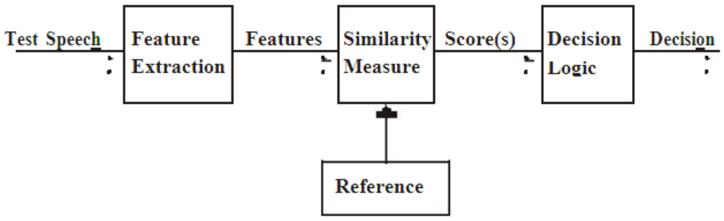
Speaker recognition methods can also be divide into *text dependent* and *text independent* methods. In case of text dependent methods a speaker is required to utter a predetermined set of words or sentences (e.g. a password). Features of voice are extracted from the same utterance. In case of text independent methods, there is no predetermined set of words or sentences and the speaker.s may not even be aware that they are being tested.

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.







A. Feature selection and measures

The speech signal can be represented by a sequence of feature vectors in order to application of mathematical tools without the loss of generality. Most of these features are also used for speaker dependent speech recognition systems. In practical real life systems, several of these features are used in combinations. In general the feature should preserve or highlight information and variation in the speech that is relevant to the basis being used for the speech recognition and at the same time minimize or eliminate any variation irrelevant to that task. Feature space should be relatively compact in order to enable easier learning of models from finite amounts of data. A feature representation that can be used without much consideration in most circumstances should be used. The process of feature calculation should be computationally inexpensive. Processing delay (i.e. how much of the ’future’ of the signal you have to know before you can emit the features) is a significant factor in some settings, such as real-time recognition.

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. FEATURE EXTRACTION TECHNIQUES

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 methods are the most widely used in speech coding, speech synthesis, speech recognition, speaker recognition and verification and for speech storage – LPC methods provide extremely accurate estimates of speech parameters, and does it extremely efficiently – basic idea of Linear Prediction: current speech sample can be closely approximated as a linear combination of past samples, i.e.,

B. MFCC

MFCC is based on human hearing perceptions which cannot perceive frequencies over 1Khz. In other words, in MFCC is based on known variation of the human ear's critical bandwidth with frequency [8-10]. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech.

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.

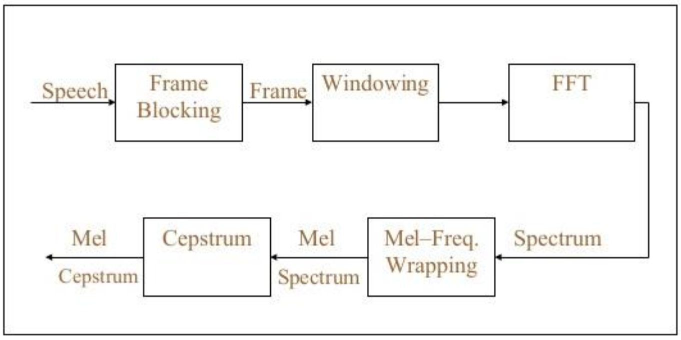
IV. MEL FREQUENCY CEPSTRAL COEFFICIENTS

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.

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



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.

$$k \qquad\qquad\qquad 1 \quad \pi$$

$$k \qquad\qquad\qquad -$$

$$C_n = \sum \quad \square 1 \, (\log S_k) \cos \quad nk - 2 \quad * \quad {}_k \, ,$$

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

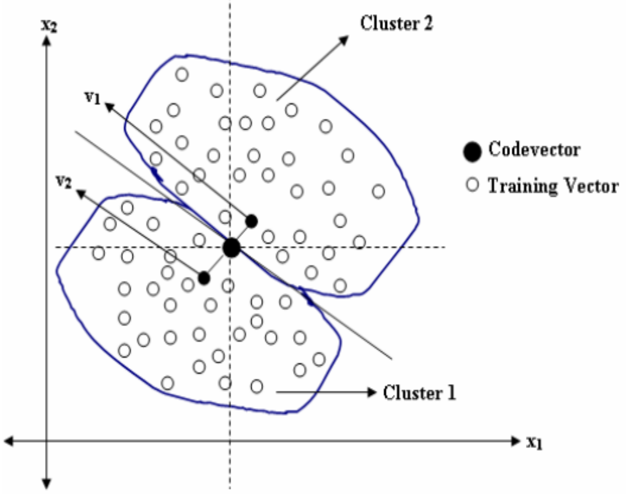
Whereas S_k, $K = 1, 2, \ldots K$ are the outputs of last step. Complete process for the calculation of MFCC is shown in above diagram.

V. VECTOR QUANTIZATION USING LINDE BUZO AND GRAY (LBG) ALGORITHM

Vector quantization (VQ) is one of the lossy data compression techniques and has been used in number of applications, like pattern recognition, speech recognition and face detection, image segmentation, speech data compression, Content Based Image Retrieval (CBIR) , Face recognition, iris recognition, tumor detection in mammography images etc.

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.

In this (LBG) algorithm centroid is computed by taking the average as the first codevector for the training set. In Figure 1 two vectors v1 & v2 are generated by using constant error addition to the codevector. Euclidean distances of all the training vectors are computed with vectors v1 & v2 and two clusters are formed based on closest of v1 or v2. This modus operandi is repeated for every cluster. The shortcoming of this algorithm is that the cluster elongation is +135o to horizontal axis in two dimensional cases resulting in inefficient clustering.



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 attempts	False Acceptance	False Rejection
U1	5	0	0
U2	5	0	0
U3	5	2	0
U4	5	1	0
U5	5	0	0
Total	25	3	0

Threshold value for distance=5.6 Efficiency=88%

MFCC with 22 filters

Speaker	No. of attempts	False Acceptance	False Rejection
U1	5	0	1
U2	5	0	0
U3	5	0	3
U4	5	0	1
U5	5	0	2
Total	25	0	6

Threshold value for distance=6 Efficiency=76%

MFCC with 32 filters

Speaker	No. of attempts	False Acceptance	False Rejection
U1	**5**	**0**	**1**
U2	**5**	**0**	**0**
U3	**5**	**0**	**0**
U4	**5**	**0**	**1**
U5	**5**	**0**	**1**
Total	**25**	**0**	**3**

Threshold value for distance=6.3 Efficiency=88%

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.

Hamming Window

Speaker	No. of attempts	False Acceptance	False Rejection
U1	**5**	**0**	**1**
U2	**5**	**0**	**0**
U3	**5**	**0**	**1**
U4	**5**	**0**	**0**
U5	**5**	**0**	**2**
Total	**25**	**0**	**4**

Threshold value for distance=6 Efficiency=84%

Rectangular Window

Speaker	No. of attempts	False Acceptance	False Rejection
U1	**5**	**0**	**2**
U2	**5**	**0**	**0**
U3	**5**	**0**	**1**
U4	**5**	**0**	**3**
U5	**5**	**0**	**3**
Total	**25**	**0**	**9**

Threshold value for distance=6 Efficiency=64%

VII. CONCLUSION

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
Efficiency	88%	76%	88%

windo **Types of w using 32 filters**	**Efficiency**
Hanning Rectangular	84% 64%

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 lslam, 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 lslam, 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).