Journal of the Nigerian Association of Mathematical Physics
Volume26, (*March, 2014*), pp 461 – 466

Feature Extraction and Identification of Speech Using Mel-Frequency Cepstral Coefficient Approach

Emagbetere J. O.

Department of Electrical/Electronic Engineering,
University of Benin, Benin City

Abstract

This paper evaluates the performance of the feature extraction module using Mel-Frequency Cepstral Coefficient (MFCC) technique. An MFCC algorithm is used to simulate the feature extraction module. The cepstral coefficients of the extracted features of the speech samples were obtained using MFCC algorithm. Extracted features of 4 test speech sample were compared with the codebooks from the database of 4 trained speech samples. The distortion distance measured in terms of the Euclidean distance between the features of the test samples and the database were recorded at different threshold values (6 MFCC and 12 MFCC) of the Mel-frequency cepstral coefficients. This work was simulated and analyzed in the Matlab®7.5 environment. The speaker recognition system achieved a 100% authentication rate with the 4 numbers of speech sample tested. Thus in this work, the feature extraction module performed satisfactorily with the MFCC technique used..

Keywords: Feature extraction, Mel-Frequency Cepstral Coefficient, Vector quantization, Euclidean distance, Voice recognition, speaker recognition

1.0 Introduction

Speech processing is an aspect of digital signal processing that is important and used for speech recognition. Speech is an important characteristic of the human being that is peculiar to an individual [1]. The human speech contains several features that can be used to identify speakers [2]. The speech signals can be represented by sequence of feature vectors.

The process of automatically recognizing who is speaking by distinguishing qualities in a speaker's voice is called speaker recognition [3]. The speaker recognition system basically performs two functions; speaker identifications and speaker verifications. These two functions can be realized in three main components; feature extraction and selection, pattern matching, and classification [2,3]. The basic modules for a speaker recognition system are; the front-end processing module, the speaker modeling module, the speaker database module, and the decision logic module. The processing module converts the sampled speech signal into sets of feature vectors which characterize the properties of the speech that can separate speakers [2,4,5]. The goal of the speaker recognition system is to design a system that can minimize the probability of verification errors, and the objective is to discriminate between the given speaker and all others.

The use of voice as the only way of providing speakers identity has certain associated problems. Problems such as the characteristics of the speaker voice being corrupted by the characteristics of the communication channel or by background noise [6]. Issues with speech being a time-dependent process; the same word said differently with the same duration can differ due to the difference in the part of the word spoken at different rate [7]. These problems have significant effect on the recognition system capability. As a result of these issues, some recognition systems have been designed to accept wide range of variations of a user's voice. However, this provision has its own shortcomings in that it may allow other speakers with similar voice characteristics to be accepted by the system. This has made the extraction of the features/characteristics of the voice signal a challenging task in speech signal processing. Although, features such as the pre-processing, pre-emphasis, and windowing schemes are employed to remove noise and silence occupying the voice spectrum, as well as minimize discontinuities of the signal, these attempts has not totally eliminate the problems. Thus, the effectiveness and performance of the extraction module is crucial to speaker recognition system [8, 9, 10].

Since the efficiency of the speaker recognition system depends on the performance of the speech signal feature extraction module, this paper presents a feature extraction module with Mel-frequency Cepstral (MFCC) approach, with a

Corresponding author: E-mail: joyokumo@yahoo.com, Tel.: +2348067147186

view to evaluate the performance of the module under varying Mel-frequency Cepstral (MFC) coefficient values. Different values of Mel-frequency Cepstral (MFC) coefficients were used for analysis in this paper. This was to verify the effect of the choice of MFCC value used by designers in simulating the extraction module [2]

2.0 Background of study

Speaker recognition system is a commonly used biometric today [1]. This recognition system is classified into speaker identification and speaker verification. The task of speaker verification system is to verify the claim identity of a person from his voice, while the speaker identification system decides who the person speaking is from a database [11]. System identification can be classified into text-dependent or text-independent system depending on if the identification is based on known utterance or any given utterance. Speaker identification consists of two stages; speaker enrollment or speaker modeling and speaker recognition [1, 6, 11]. In speaker enrolment, features are extracted from all speakers and models are built for them, while in speaker recognition features are extracted from the speaker under test and a model is built for and compared with models of all speakers in the database to find the closest speaker (matching).

In both speaker identification stages, features are extracted. This makes the feature extraction module and important module in speaker recognition system and the efficiency of the system depends on the extraction techniques employed. One of the most successful techniques used in speaker recognition is the Mel-frequency cepstrum coefficients for feature extraction and the vector quantization model for matching [1, 2, 4, 5, 11].

The human perception of the frequency contents of the sound for speech signals is non-linear [4, 11, 12,]. The human perception system perceives speech signals in the Mel-scale. Based on the non-linearity of the human perception system, Mel-frequency emerged to mimic it. Mel is a unit of measure of perceived pitch or frequency of a tone. This scale does not correspond linearly to the physical frequency of the tone. The Mel-frequency scale is a linear frequency spacing below 1kHz and logarithmic spacing above 1kHz [1, 2, 13, 14, 15].

The Mel-frequency cepstrum coefficients provide a compact representation of the speech signal for a given frame analysis. The coefficient is a result of the cosine transform of the logarithm of the short-term energy spectrum expressed on a Mel-frequency scale. Because the Mel-frequency coefficients are real numbers, the DCT helps to convert them to the time-domain, and the result is called the Mel-frequency cepstrum coefficients [2, 12]. The coefficients are obtained through Mel-scale filters.

The MFCC are used for designing a text-dependent speaker identification system such as the one considered in this paper. The extracted speech features (MFCC's) of a speaker are quantized to a number of centroids (code word) which constitute the codebook of that speaker MFCC's. The idea behind vector quantization (VQ) is that each feature vector is mapped to a finite number of regions in that space, and each region called a cluster is represented by its centre called a centroid [4, 7, 10]. Collection of all code words is called a codebook.

The VQ is used for matching samples. And one speaker speech sample can be discriminated from another based on the location of centroid. The distance from a speaker feature vector to the closest codeword of a codebook is called VQ-distortion. The total VQ-distortion is compared, and the speaker corresponding to the VQ codebook with the smallest total distance id identified. Thus, in speaker identification, the distortion distance of two vector sets can be measured based on the minimum Euclidean distance.

3.0 Research Method

The Mel-frequency cepstrum coefficient (MFCC) feature was used for the text-dependent speaker identification algorithm. This MFCC method was employed to extract features from the speech signal and compare test speech sample with the speaker features in the database. The block diagram of the MFCC algorithm used to simulate the feature extraction module is presented in Fig. 1. The feature extraction module converts the raw speech waveform (see Fig 2 – 5) of a speaker to a feature vectors. The feature vectors form representative codebooks of the speaker's voice pattern to create the database. The extracted speech features (MFCCs) of a speaker were quantized to a number of centroids (64 centroids) using vector quantization algorithm.

The input voice samples used for this study were obtained from students (two males and two females) within the average ages of 22yrs. Voice samples were recorded at a rate of 8 kHz to 16kHz, with a built-in microphone of a Nokia 6120 classic smart phone. Each speaker was made to utter the word 'close' four times, and from the pool of the voice samples, one sample was randomly picked per speaker to train, and store in the database. The recording environment was not noise-free. Using Fig.1, the speech signals were extracted in terms of Mel-frequency cepstrum coefficients. Because, the way data are represented is crucial in speech pattern recognition, the input voice speech signal data of the speakers used for this investigation were presented in time-amplitude waveform as shown in Figures 2 – 5. The Mel-frequency cepstrum coefficients were obtained for the training and testing phase. The codes were developed in the Matlab environment to compares the test sample with the pool of extracted speaker feature vectors in the database, in term of their VQ-distortion distance (Euclidean distance). The Euclidean distance between the MFCC of each speaker were obtained, and the speaker

with the minimum Euclidean distance identified.

The results obtained from the simulation process in terms of the Euclidean distance between the trained sample and the tested samples were presented for various threshold values of the Mel-frequency cepstrum coefficients. This was to determine the effectiveness of the system in terms of how well the system (Fig.1) can match a speaker with the database sample irrespective of the MFC coefficient used to develop the Matlab code. The Mel-frequency coefficient threshold values set for this work were 6, and 12. The results obtained are presented in Table 1, with the minimum Euclidean distance for each test speaker highlighted.

4.0 Data presentation and Analysis

Waveforms for the 4 input samples (female 1, female 2, male 1, and male 2) used for this study represented in time-amplitude speech pattern are shown in Figures 2 - 5.

5.0 Discussion

The speech samples used are presented in time-amplitude waveform pattern are shown in Figures 2 – 5. It is observed from Figures 2 – 5 that for female 1 and 2, the audible voice signal lies between 100ms to 150ms, while that for male 1 and 2 lies between 80ms to 110ms. The speech features were extracted using Fig.1. For each test speech sample compared with the extracted features of stored in the database, the Euclidean distance obtained were tabulated as shown in Table 1. In Table 1, the cells having the minimum Euclidean distance are highlighted as the most likely authenticated speaker.

It was also observed that the Euclidean distance obtained varied with different MFC coefficient values. With the use of 12 MFCC and 6 MFCC, the speaker recognition system's ability to discriminate accurately was high. All valid speakers were correctly identified, and their corresponding cells highlighted as shown in Table 1. For 12 MFCC value, the Euclidean distances obtained for unknown speakers were large compared to the identified speaker. Thus with 12 MFCC, the rate of false acceptance will be very minimal.

6.0 Conclusion

The feature extraction module presented in this paper using the Mel-frequency cepstral coefficients technique performed satisfactorily. The text-dependent system was trained with 4 input speech signals from 2 males and 2 females, and modeled using the VQ techniques. Each input speech signal tested against the database samples were accurately identified to their database samples with their minimum Euclidean distance. From the results obtained, the speaker recognition system achieved a 100% authentication rate with the 4 numbers of speech sample tested (see Table 1). Thus in this work, the feature extraction module performed satisfactorily with the MFCC technique used.

The study also revealed that 12 MFCC value is preferable to 6 MFCC when simulating the feature extraction module. This is to ensure that false acceptance rate is reduced to the minimum.

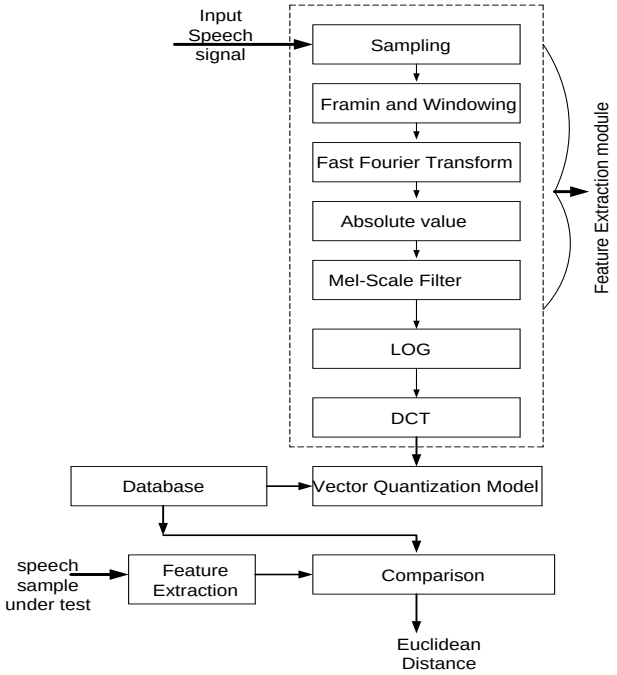

Fig 1: Feature extraction and identification module.

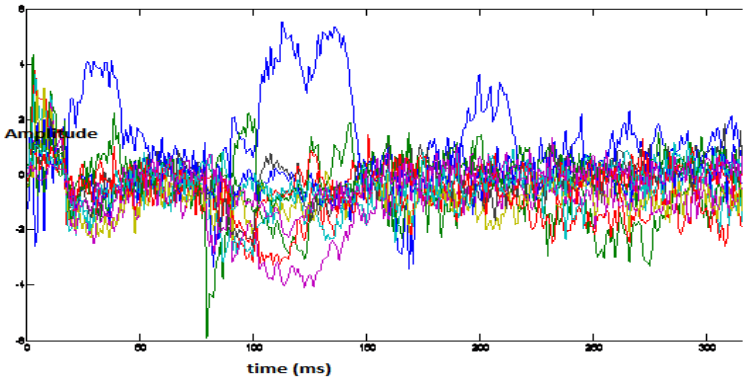

Fig.2: Female 1 voice sample of the word 'close'.

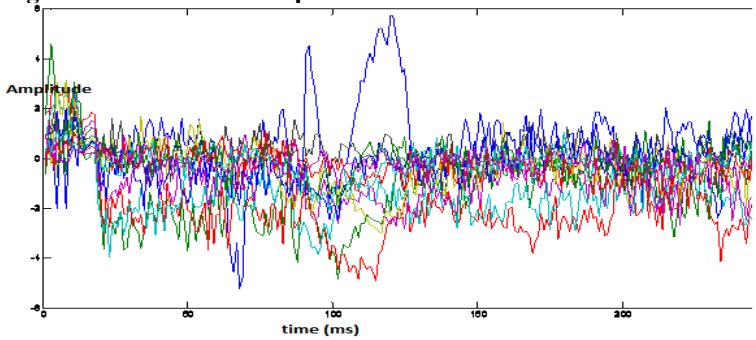

Fig.3: Female 2 voice sample of the word 'close'.

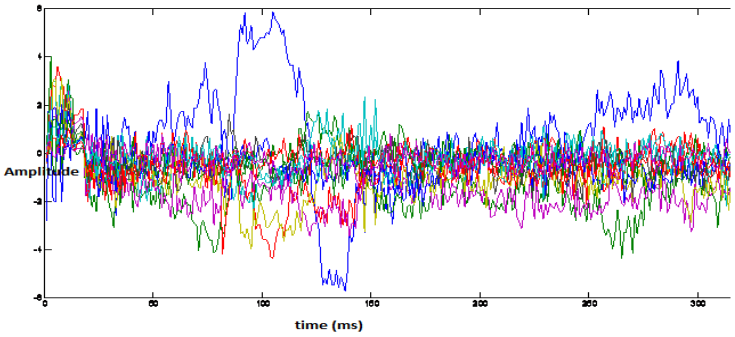

Fig.4: Male 1 voice sample of the word 'close'.

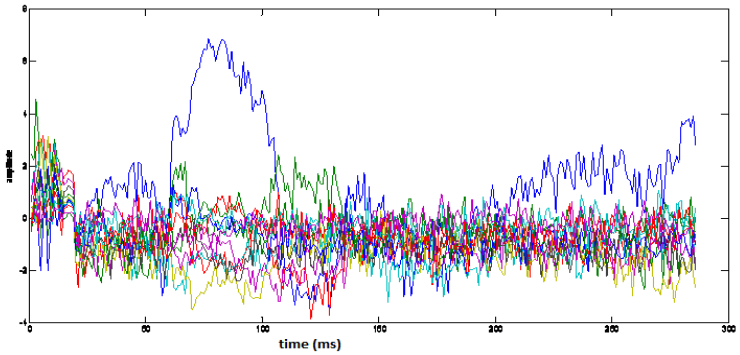

Fig.5: Male 2 voice sample of the word 'close'.

Table 1: Speaker identification with different MFCC threshold values.

			output		
		female1	female2	male1	male2
	female1	0.6849	2.5498	4.1484	4.0139
input	female2	2.2384	1.0211	4.8677	3.5653
	malecc1	4.9138	3.1641	0.8737	5.1282
	malecc2	5.0471	3.9051	6.5429	0.8737
mel-cepstral coefficients = 6					
			output		
		female1	female2	male1	male2
	female1	1.4866	4.0849	6.8283	7.3345
input	female2	4.5719	1.8155	6.9895	5.9398
	malecc1	6.9404	5.3882	1.8173	8.1146
	malecc2	7.9944	7.3892	9.6787	1.8871
mel-cepstral coefficiens = 12					

References

[1] Patel K, Prasad R.K, Speech Recognition and Verification Using MFCC and VQ, International Journal of Emerging Science and Engineering (IJESE), Vol.1, Issue 7, May 2013, pp 33 – 37.

[2] Vibha T, MFCC and its application in speaker recognition, Intenational Journal on Emerging Technologies, 1(1), 2010, pp 19 – 22.

[3] Dutta T, Text-dependent Speaker Identification based on Spectrograms, Proceedings of Image and Vision Computing, New Zealand 2007, pp 238 – 248.

[4] Srinivasan A, Speaker Identification and Verification using Vector Quantization and Mel Frequency Cepstral Coefficients, Research Journal of Applied Sciences, Engineering and Technology, 4(1), 2012, pp 33 – 40.

[5] Singh S, Rajan E.G, Vector Quantization Approach for Speaker Recognition using MFCC and Inverted MFCC, International of Computer Applications, Vol. 17, No.1, March 2011, pp 1 – 7.

[6] Singh S.K, Features and Techniques for Speaker Recognition, M.Tech. Credit Seminar Report, Electronic Systems Group, EE Dept, IIT Bombay, November 2003, pp 1 – 16.

[7] Renals S, Speech Recognition, COM326/COM646/COM678 lecture note, February 1998.

[8] Muda L, Begam M, Elamvazuthi I, Voice Recognition Algorithms Using Mel Frequency Cepstral Codfficient (MFCC) and Dynamic Time Warping (DTW) Techniques, Journal of Computing, Vol.2, Issue 3, March 2010, pp 138 – 143.

[9] Bala A, Kumar A, Birla N, Voice Command Recognition System Based on MFCC and DTW, International Journal of Engineering Science and Technology, Vol.2, No. 12, 2010, pp 7335 – 7342.

[10] Thakur A.S, Sahayam N, Speech Recognition Using Euclidean Distance, International Journal of Emerging Technology and Advanced Engineering, Vol.3, Issue 3, March 2013, pp 587 – 589.

[11] Abdulnasir H, Said A, A text-independent speaker identification system based on the Zak Transform, Signal Processing an International Journal (SPIJ), Vol. 4, Issue 2, pp 68 – 74.

[12] Ouzounov A, Cepstral Features and Text-dependent Speaker Identification – A comparative Study, Cybernetics and Information Technologies, Vol.10, No.1, 2010, pp 3 – 11.

[13] Iqbai S, Mahboob T, Khiyal M.S.H, Voice Recognition Using HMM with MFCC for Secure ATM, International Journal of Computer Science Issues, Vol.8, Issue 6, No.3, November 2011, pp 297 – 303

[14] Quatieri T.F, Discrete-time Speech Signal Processing: Principles and Practice, Prentice Hall, 2002.

[15] Deller J.R, Hansen J.H.L, Proakis, 'Discrete-Time Processing of Speech signal, IEEE Press, New York, NY, 2000