International Journal of Computer Trends and Technology
ISSN: 2231-2803 / https://doi.org/10.14445/22312803/IJCTT-V67I4P118
Volume 67 Issue 4, 79-84, April 2019

Original Article

Voice Recognition Using MKMFCC and VQ

Aswathi Menon[1], Anjali Krishnan[2], Arul V H[3]

[1] *PG Scholar, Department of Electronics and Communication Engineering A P J Abdul Kalam Technological University ,Thejus Engineering College, Kerala, India*

[2,3]*Assistant professor, Department of Electronics and Communication Engineering A P J Abdul Kalam Technological UniversityThejus Engineering College, Kerala, India*

Abstract *- Voice processing is emerging as an incredible means of human communication. Increased human dependency on machines led to the development of voice recognition, mainly for human-machine communication. In recent years, it has become an inevitable part of medical, military, and other security applications. Specific parameters that differentiate one speaker from another are the features; their extraction and classification describe every voice recognition system. Several voice recognition algorithms are evolved under various environmental conditions. MKMFCC (Multiple Kernel Weighted Mel Frequency Cepstral Coefficients), as a feature extraction algorithm, provides better feature extraction even in a noisy or degraded environment compared to other existing algorithms. Besides, VQ (Vector Quantization) is a good classifier because of its low computational burden. This paper proposes a voice recognition system combining the good qualities of MKMFCC and VQ, with MKMFCC as the feature extraction algorithm and VQ as the classifier. The proposed system provides an accuracy of about 92.6% for the voice recognition process.*

Keywords *- MKMFCC, VQ, Feature extraction, Classification*

I. INTRODUCTION

The expression of human thoughts that contains feelings, emotions, etc., can be called speech. Speech has been at the forefront of human communication for many decades. Analysis of human speech and its study is an area that has invited the attention of scientists till today. Our world is on a path of rapid development, i.e., rapid automation and rapid mechanization. So in such a situation, machines have replaced humans in many areas. Artificial intelligence and Robotics have created tremendous changes in various industries. In short, automation is ruling the world. Since machines have become an inevitable part of development and technologies, communication between humans and machines was necessary. Such a thought led researchers and scientists to focus on automatic speech recognition. With the help of some computer programs and algorithms, speech recognition makes the machine or computer understand what we are speaking. Thus a lot of time can be saved while doing tedious tasks, and as compared to other forms of communication such as writing and typing, speech is somewhat better and more comfortable. The methodologies and technologies developed by the interdisciplinary subfield of computational linguistics for recognizing and translating spoken language into text by computers is speech recognition.

Speech recognition systems are usually categorized into two: - Speaker dependent and Speaker independent systems. The former requires training before its use, whereas the latter recognizes most user voices with no training. A voice recognition system's performance highly depends on the features extracted from a particular speaker. Features are some parameters or specific quantity to determine a signal, or it distinguishes or determines the particular nature of a signal. These features will be different for different speakers. So, the extraction of such a unique feature is important. Various algorithms extract features from an incoming speech signal, which is analog. Extraction of features from the speaker's voice signal and classification is the basic foundation of a voice recognition system.

The incoming analog speech signal is initially digitized for ease of analysis. Those digital speech frames are then windowed to avoid sudden discontinuities in each frame. After noise removal, the speaker-specific features are extracted using available algorithms such as MFCC, LPC, LPCC, etc. Training and testing are the two phases of feature extraction. A trained database is created using the extracted features, which can be used later for matching. And finally, this trained database is compared against the test signals to obtain a match.

Many feature extraction algorithms, such as MFCC, LPC, etc., have been used for many years ago. All these algorithms have their pros and cons. The common drawback among them was the poor performance in a noisy background. Out of these, MFCC was most widely used as it accurately simulates the behavior of the human auditory system. So, by focussing on the advantages of MFCC and discarding its disadvantages, by weighting MFCC

coefficients with tangential and exponential components, a new algorithm was derived, MKMFCC (Multiple Kernel Weighted Mel Frequency Cepstral Coefficients). It can perform well in noisy as well as in any degraded environment. This noise-resistant peculiarity of MKMFCC coefficients made them well suited for voice recognition systems. Similar to feature extraction, the classification of features is also vital. It is the classifier's output that will provide us with the result. Because of its ease of implementation and low computational burden, VQ (Vector Quantization) is used here as the classifier.

This paper proposes a voice recognition system using MKMFCC as the feature extraction algorithm and VQ as the classifier.

Following the introduction, the paper is organized as follows: - A small review of existing research works related to voice recognition is included in Section 2. Section 3 deals with the proposed methodology that describes the voice recognition system using MKMFCC and VQ. Results and discussions of the proposed system are given in Section 4, and finally, Section 5 concludes the paper.

II. RELATED WORKS

Accordingly, about 30 research works related to voice recognition were reviewed and analyzed to develop the proposed system.

Voice recognition systems were used mainly in security and to preserve languages. Isolated word recognition systems for various languages such as Hindi, English [1], and Tamil [2] were reviewed. MFCC (Mel Frequency Cepstral Coefficients) was used as the common feature extraction algorithm. Because of the resemblance of the Mel scale to the human auditory system, MFCC was widely used for feature extraction. Researchers imply that MFCC can provide a better sound representation than other existing algorithms. A voice command recognition system based on MFCC and DTW [3] explain how an individual's voice can be used for security authorizations. The main drawback that everyone faces while developing voice recognition systems by using MFCC is its poor performance in the presence of noise. Other algorithms such as LPC and LPCC have found less use than MFCC. In a speech recognition system that uses MFCC along with VQ as the classifier for the Hindi language [4], a 90% success rate was obtained. Its authors have claimed that using VQ as the classifier model promotes great accuracy. While reviewing the research papers, classifier models found in use were KNN, SVM, and mostly VQ. VQ was used along with MFCC in many systems, in which either K-means clustering or LBG algorithm was used along with VQ for clustering the feature vectors of the speaker. The key point obtained through the literature survey was that MFCC had found wide use, but its use has been reduced considerably in recent years due to its poor performance in noisy backgrounds. Likewise, VQ was the feature classifier that was mostly used and provided low computational burden and high accuracy. A robust noise MKMFCC- SVM automatic speaker identification system [5] was developed in 2018, in which authors say that MKMFCC has a higher identification rate even in a noisy environment. Because of the simplicity of VQ, it was used along with MFCC as the classifier for identifying unknown speakers from a given set of registered speakers [6]. Many authors have performed comparison studies between various feature extraction techniques and classifiers [7], [8], [9].

III. PROPOSED METHODOLOGY

A detailed description of the proposed voice recognition system using MKMFCC as the feature extraction and VQ as the classifier model is provided in this section. The overall block diagram of the proposed system is shown in Fig1. Like other voice recognition systems, the common steps are pre-processing, feature extraction, modeling, and matching. Initially, the pre-processing stage accepts the audio signal from multiple speakers as input to the voice recognition system. This incoming analog audio signal is treated in the initial stages to perform well in the forthcoming stages. To improve the SNR of the audio signal and to attenuate interferences, speech enhancements are usually done in the pre-processing stage. This stage converts analog audio signals into digital signal frames for easy analysis, as it is difficult to analyze the analog speech signal, which is assumed to have a quasi-stationary nature. The final stage in pre-processing is the acoustic beamforming, which enhances the signal of our interest and suppresses all other unwanted signals. Following the pre-processing stage, extraction of speaker-specific characters, i.e., feature extraction, occurs. Here, the MFCC (Mel Frequency Cepstral Coefficients) cepstral coefficients are extracted, in which tangential and exponential components are weighted to obtain the so-called MKMFCC coefficients. The classifier provides the recognition result, as the extracted MKMFCC cepstral features are fed as training input to the VQ classifier. VQ uses LBG (Linde-Buzo-Grey) algorithm for clustering the speaker-specific features.

A. Feature Extraction

MKMFCC is the feature extraction method proposed here for the voice recognition system. In any voice recognition system, feature extraction is the core part. All speaker-specific parameters (Features) must be extracted exactly to recognize the unknown voice signal properly. The acoustic feature, i.e., MKMFCC, is considered in the proposed

system since it enhances and preserves the information from the spectral formant; also, it has better performance in noisy and degraded environments than other existing algorithms. This MKMFCC is born from the ever-used coefficients, MFCC (Mel Frequency Cepstral Coefficients), which differ from other coefficients as it approximately mimics the human auditory system.

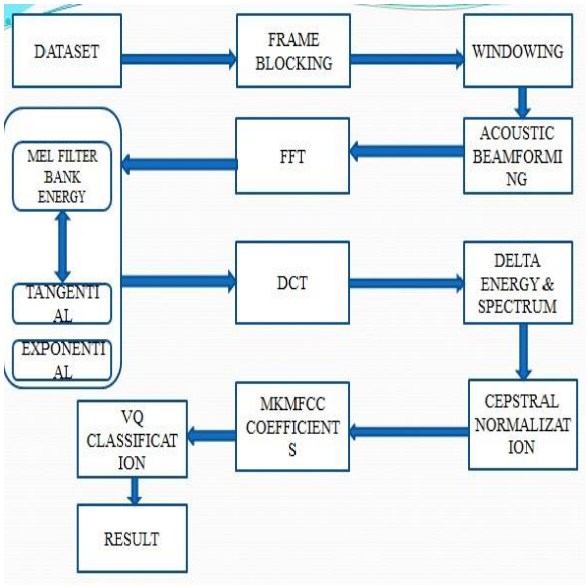

Fig. 1 Proposed voice recognition system using MKMFCC feature extraction and VQ classifier

a) MKMFCC

In MKMFCC, weightage is given to MFCC coefficients using two kernel functions, such as divergent and exponential functions. A natural way to merge and integrate various kinds of data features is provided by kernel weightage. Multiple kernels allow considering both high and low energy frames of an audio signal. Thus no portion of the signal is left unconsidered. The basic block diagram of MKMFCC is shown in Fig 2. The steps are being explained below:-

i. ***Pre-Emphasis****:* The incoming audio signal is passed through a filter that enhances the signal's higher frequency components, thereby providing higher energy to the acoustic signal in the higher frequencies.

ii. ***Framing****:* For easy analysis of the analog audio signal, which is quasi-stationary, they are sampled and combined to form frames. 20-40ms is the commonly used frame length. i.e., the signal is divided into *l* frames of *M* samples.

iii. ***Windowing (Hamming Window)****:* Windowing is done to avoid sudden discontinuities in the frame, as it diminishes the signal to zero at the beginning and end of each frame. Usually, Hamming window is used for this purpose, reducing the number of side lobes.

Hamming window function is given by W (m) = $0.56 - 0.46(\frac{2\pi m}{M-1})$; 0≤m≤M-1, m= 1 to M, represents the number of samples in the audio signal.

iv. ***Fast Fourier Transform (FFT):*** Time-domain to frequency domain conversion of the audio frames occurs in this step. Calculation of the periodogram spectral estimate also occurs.

v. ***Mel Filter bank processing:*** The FFT spectrum holds a wide range of frequency, but such a wide range cannot obey a linear scale. i.e., the spectrum still has unwanted information even though it was removed by acoustic beamforming. Thus, the bank of filters removes the unwanted information from the FFT spectrum through Mel scale processing. A triangular filter is used to filter the signal frequencies in the Mel scale.

Mel scale and linear scale are related as follows:

$$\text{Mel (f)} = 1125 * \ln(1 + \frac{f}{700})$$

The Mel filter bank equation is given by,

$$M_f(k) = \begin{cases} 0 & k < G(f-1) \\ \frac{k - G(f-1)}{G(f) - G(f-1)} & G(f-1) \le k \le G(f) \\ \frac{G(f+1) - k}{G(f+1) - G(f)} & G(f) \le k \le G(f+1) \end{cases}$$

Here, f=1 to F is the number of Mel filters, and G () is F+2 Mel spaced frequencies. About 20 Mel filter banks are created for this recognition system.

vi. ***Filter bank energy:*** The energy is obtained by multiplying the filter bank with the power spectrum and adding some coefficients. These coefficients used here are the multiple kernel weightage functions, tangential and exponential. On obtaining the energy, its log value is calculated, as it allows using cepstral mean subtraction, which is a channel normalization technique. Thus, the output of the filter bank is a log Mel spectrum.

The Tangential and Exponential functions are given by,

Tangential weightage function,

$$WT_{m1} = \tanh\left(\frac{-F}{2} + F.\left[\frac{f-1}{F-1}\right]\right)$$

Exponential weightage function,

$$WT_{m2} = \exp\left(\frac{-F}{2} + F.\left[\frac{f-1}{F-1}\right]\right)$$

vii. ***Discrete cosine transform (DCT):*** The log Mel spectrum obtained from the filter bank is converted to the time domain using DFT. The result of DCT is itself the MKMFCC coefficients. Using the obtained energy, the coefficient is calculated as follows;

$$WC_S(m) = \frac{1}{M'} \sum_{k=0}^{M'-1} E'(K) e^{jk(\frac{2\pi}{M'})m}$$

WC_S (m) is the proposed MKMFCC cepstral coefficient.

Where, $E'(K) = \begin{cases} E(l), & K = Kl \\ 0, & Otherwise \end{cases}$

viii. ***Delta Energy and Spectrum:*** Including the trajectories of coefficients and the power spectral envelop will increase the efficiency of the recognition system by some value.

ix. ***Cepstral Normalization:*** This is performed to obtain a zero mean for the corresponding coefficients of all the feature vectors.

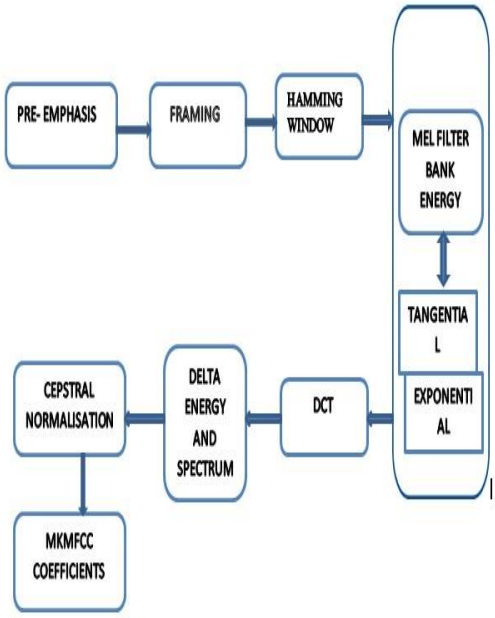

Fig. 2 MKMFCC feature extraction

B. Feature Classification

Just like feature extraction, its classification or clustering is also very crucial. The classification model proposed here to use is Vector Quantization (VQ). VQ is selected here as it reduces the time for training the feature vectors compared to other classifiers. VQ uses the Linde- Buzo- Grey (LBG) clustering algorithm to group features.

b) VECTOR QUANTIZATION

A vector quantizer maps K-dimensional vectors in the vector space R_k into a finite set of vectors given by; Y= {y_i: i=1, 2, 3…N}. Here, K is the number of feature coefficients, y_i is the code word, and Y is the codebook. In the proposed system, VQ works by creating codebooks of required sizes. The generation of a good codebook determines the performance of a classifier. The basic idea behind VQ is that it maps or clusters vectors from a large vector space into a finite number of regions in that space.

Here, MKMFCC cepstral features extracted from the speakers are fed to the VQ classifier as training input. Using the LBG algorithm, VQ cluster the features into many and provides the final result. LBG is an iterative algorithm in which the features extracted from the speakers are fed as the input. It generates a codebook using the splitting method. Codebooks created here are in the form of a matrix of size (20x16). Much of the codebook is generated depending on the number of training and testing data.

In the proposed system, a speaker-specific codebook is generated by clustering training vectors during the training phase. While during the testing phase, that trained codebook is used to vector quantizes the input utterances, and VQ distortion is computed. And finally, the speaker corresponding to the smallest VQ distortion is identified.

IV. RESULT AND DISCUSSIONS

(a) Software used

- The proposed system was implemented in the MATLAB R2014a version.
- GUI (Graphical User Interface) makes the code easy to use.

(b) Database Description

The database consists of about 500 utterances of different people consisting of both men and women around the age group of 20-30 who speak the word "Green Tree," which was taken from the open-source platform for machine learning, i.e., "Tensor Flow."

The utterances of each speaker were trained and tested manually.

(c) Performance Evaluation

The performance or accuracy of the proposed voice recognition system using MKMFCC and VQ was evaluated based on True positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) rates.

- True positive: - Positive cases that were correctly identified.
- True negative: - Negative cases that were incorrectly classified as positive.
- False-positive: - Negative cases that were correctly classified.
- False-negative: - Positive cases that were incorrectly classified as negative.

The equation gives accuracy,

$$\text{Accuracy} = \left(\frac{TP+TN}{FP+TP+FN}\right) * 700$$

A real-time dataset of about 20 different utterances from 5 different people was also provided to the proposed voice recognition system, and the dataset

was taken from the "Tensor flow ."The same performance was obtained for the real-time data also.

The proposed voice recognition system using MKMFCC and VQ has provided an accuracy of about 92.6%.

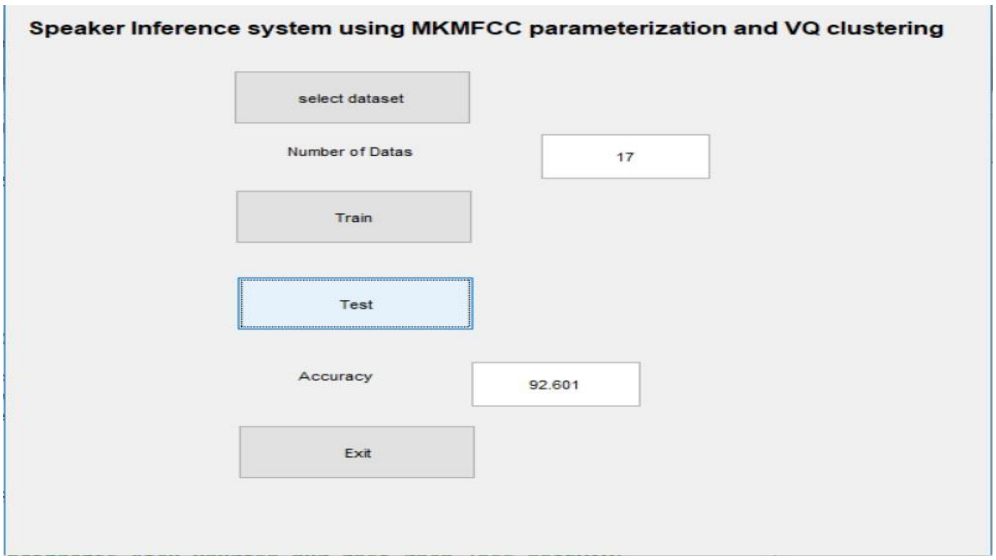

Fig. 3 GUI window that shows the overall steps

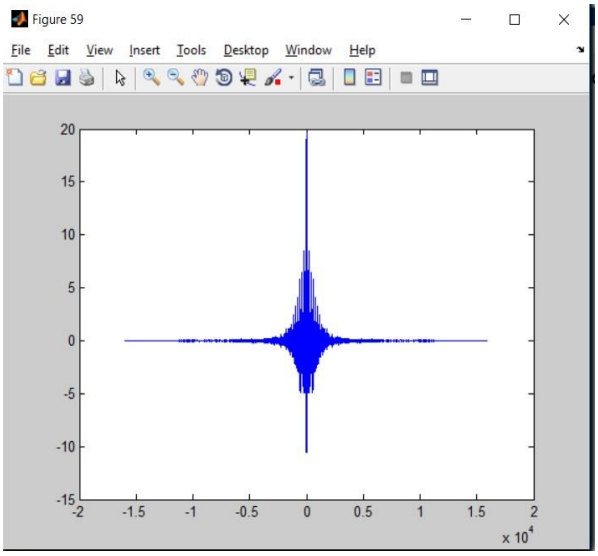

Fig. 4 Input speaker signal

Fig 3 shows the GUI window that appears while running the proposed system's code. About 17 datasets are taken for experimenting once, the fig shows the window of that time, and testing and training of the data are also shown in GUI.

Fig 4 shows one of the input speaker signals of the dataset.

V. CONCLUSION

An innovatory approach is proposed in the paper for a voice recognition system using MKMFCC feature extraction and the VQ technique of feature classification. The feature parameter chosen for the recognition system was multiple kernels weighted Mel frequency cepstral coefficients, which considers both high and low energy frames of the sound signal so that none of the signals is left unconsidered. Tangential and Exponential weighted functions are the so considered multiple kernel functions. Classification of the speaker features was done using the Vector Quantization method, which reduces the computational burden. VQ uses the LBG algorithm for clustering the parameters. Dataset taken from the open-source machine learning platform "Tensor flow" containing about 500 utterances from different people were used for the performance analysis of the proposed system based on True positive, True negative, False positive, and False-negative. The proposed voice recognition system using MKMFCC and VQ has provided an overall accuracy of about 92.6% for both the real-time dataset and the one taken from the open-source platform.

REFERENCES

[1] H B Kekre, A A Athawale, *Speech Recognition using Vector Quantization*, International Conference and Workshop on Emerging Trends in Technology (ICWET 2011) – TCET, Mumbai, India.

[2] Madhavaraj, A., & Ramakrishnan, A. G. (2017). *Design and develop a large vocabulary and continuous speech recognition system for Tamil.* 2017 14th IEEE India Council International Conference (INDICON).

[3] Kumar, A. N. A., & Muthukumaraswamy, S. A. (2017). *Text-dependent voice recognition system using MFCC and VQ for security applications.* 2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA.

[4] Suman K. Saksamudre, R. R. Deshmukh, *Isolated Word Recognition System for Hindi Language*, International Journal of Computer Sciences and Engineering

[5] Faragallah, O. S. (2018). *Robust noise MKMFCC–SVM automatic speaker identification.* International Journal of Speech Technology.

[6] Soong, F., Rosenberg, A., Rabiner, L., & Juang, B. (n.d.). *A vector quantization approach to speaker recognition*. ICASSP '85. IEEE International Conference on Acoustics, Speech, and Signal Processing.

[7] Smita B Magre, Ratnadeep R. Deshmukh, Pukhraj P. Shrishrimal, *A comparative study on feature extraction techniques in speech recognition*, International Conference on Recent Advances in Statistics and Their Applications.

[8] K. Sarmah, *Comparison Studies of Speaker Modeling Techniques in Speaker Verification System*, International Journal of Scientific Research in Computer Science and Engineering Vol.5, Issue.5

[9] Sarosi, G., Mozsary, M., Mihajlik, P., & Fegyo, T. (2011). *Comparison of feature extraction methods for speech recognition in noise-free and traffic noise environments.* 2011 6th Conference on Speech Technology and Human-Computer Dialogue (SpeD).