ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Single and Multispeaker Voice Recognition in Android

Tanmay Naskar[1], Anket Sah[1], Akshay Mense[1], Neha Patekar[1], Ms. Nikhatjahan Kankurti[1],
Dr. Neeta Deshpande[2], Suvarna Kadam[2]

B. E Scholar, Department of Computer Engineering, D Y Patil College of Engineering, Akurdi, Pune, MH, India[1]

Head of Department, Computer Engineering, D Y Patil College of Engineering, Akurdi, Pune, MH, India[2]

ABSTRACT: Voice is the most natural form of communication and the primary goal with this system is to make interacting with machines as easy as possible. Voice recognition and speech recognition are considered same however, they have distinct meanings. Speech recognition detects only the words, while the voice recognition disregards the language or anything linguistic for that matter and detects the identity of the person speaking. In this paper, we discuss two parts. In the first part of the paper, we discuss the means of processing the speech signal using the Mel Frequency Cepstral Coefficient and the basics of voice characteristics. The problem of voice recognition was already implemented at the hardware level, but here we propose a solution to the problem on the software end with a higher level of accuracy. Moreover, in the second part, we discuss the application developed on Android which can perform both single speaker and multi-speaker recognition. The multispeaker feature enables the system to identify more than one person speaking by splitting the input into smaller blocks and treating each block as a single speaker identification problem. In the single speaker recognition case, the algorithm is able to correctly identify the individual 90.3% of the time. In the multispeaker recognition case, up to 3 individuals are correctly identified roughly 40% of the time. The average identification time is 220ms across a database of 415 samples.

KEYWORDS: Voice biometric; Voice Recognition; voice identification; Multi Speakers; MFCC; LPCC; VAD; Smartphone Application.

I. INTRODUCTION

Voice recognition is a multi-disciplinary technology which uses vocal attributes of a speaker's voice and help us in identifying or confirming the identity of an individual based on his voice. Speaker Recognition is a branch of biometrics that is used for identifying, verifying and classifying individual or multiple speakers. Enormous capacity of information can be carried by a voice signal. Voice recognition system which involves voice identification as well as speaker verification, is based on the fact that a person's speech reflects its several unique attributes. Voice signal can be transmitted over long distances via telephone media and even seen as a non-evasive biometric that can be collected with or without the knowledge of person. A speaker's voice cannot be lost, stolen or at most forgotten unlike different forms of identification, such as a password or a key. A secure method of authenticating and identifying speakers is allowed by speaker recognition application. While there are several different biometric recognition systems available like fingerprint recognition system, retinal scans, face recognition etc. these are more dependable methods used across security and access control system for identification[3]. In near future, it is expected that voice recognition will be used widely to make it feasible, the use of speaker's voice for verifying and validating their identity and for controlling access to multiple services such as voice mail, information services, voice dial, telephone shopping, banking by telephone, security control for confidential information areas or providing remote access of computers. The work of

DOI: 10.15680/IJIRCCE.2017. 0505115

RnD (research and development) on voice recognition method and related techniques has been done for more than six decades until now and still it continues to be an active area. For representing the voice signal different feature matching and extraction methods are developed during this span of six decades. The two most popular algorithms of voice feature extraction presented are: Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Cepstral Coefficients (LPCC). After a test was done on 415 different voices with users from both genders, it was proven that the MFCC is more efficient than the LPCC and can reach up to an accuracy rate of 90% whereas the LPCC would hit 75%[7]. Therefore, using MFCC for feature extraction proved to be the best solution. Voice Activity Detection or VAD is also important for the system. It needs to know whether a person is currently speaking, or if it's just noise. On the basis of speech attributes, differentiation of voiced signal into silence and speech is done. Slicing of signal is done into contiguous frames. With each frame, a non-negative and real-valued parameter is associated. This parameter is average energy content along with number of Zero Crossings of the frame for the time-domain algorithms[4].This parameter is spectrum as well as variance of the spectrum of a frame for the frequency domain algorithms. The signal frame is classified as ACTIVE if this parameter exceeds a certain threshold, otherwise, it is classified as INACTIVE.

II. LITERATURE REVIEW

Sr. No	Paper Name	Author Name	Key Features	Refers To
1	Intelligent Voice Assistant Using Android Platform	Sutar Shekhar	Focused on Android development over the voice control recognition, generate and analyzecorresponding commands, intelligent responses automatically[1].	• HMM (Hidden Markov Model) • TTS (Text to Speech) • Intent • Android OS
2	Speech Recognition in Marathi Language on Android O.S.	SarojJadhav, Jayshree Ghorphade, Rishikesh Yeolekar	Studied different methods for feature extraction and comparison techniques for ASR systems and select the most appropriate technique for improving the accuracy of the system[2].	• Acoustic Model • Dynamic Time Warping • Language Model • Mel Frequency Cepstral Coefficient
3	Speech to Text Conversion using Android Platform	B. Raghavendhar Reddy, E. Mahender	Developed a system which acquires speech at run time through a microphone and processes the sampled speech torecognize the uttered text and the recognized text can be stored in a file[3].	• HMM • Android OS • Speech Recognition
4	Smart Voice Search Engine	Shahenda Sarhan	Built a domain independent search engine using a smart search by voice engine which searches user speech automatically without the users request and provide him with evidence on his speech, this engine was called SVSE[4].	• Speech recognition • Search engine • Search by voice

DOI: 10.15680/IJIRCCE.2017. 0505115

5	Voice recognition based secure android model for inputting smear test results	Teenu Therese Paul, Shiju George	Identifies users voice by using the user's voice sample a secure authentication system is developed where the unique features of the user's voice are extracted and stored at the time of registration[5].	• Voice Recognition • Speaker Identification • Smear Test
6	Voice-to-Text Transcription of Lecture Recordings	Dr.Stuart Dinmore, Dr. Jing Gao	Discusses the benefits of same-language transcription of media content and goes on to outline the details of a technical feasibility study[6].	• Transcription • Universal Design Learning • Same Language Subtitles

GOALS AND OBJECTIVE: The goal is to create an Android based Biometric Application which will help to identify the speaker based on his/her voice. Since, voice of a speaker has some unique characteristics which does not change even if the voice of the person changes due to any external environment. We can store this unique characteristic of voice in a database to identify the speaker. Objective of this project is to provide higher level of accuracy in identifying the speaker, the multi-speaker feature enables the system to identify more than one speaker at a time, Techniques that can help to improve its efficiency, use it as a biometric security.

III. COMPARISON BETWEEN EXISTING SYSTEM AND PROPOSED SYSTEM

Item	Existing System	Proposed System
Technique	Linear Predictive Cepstral Coefficients (LPCC).	• Mel Frequency Cepstral Coefficients • Voice Activity Detection (VAD)
Accuracy	Low	High
Complexity	Low	High
Explanation	In Existing system, Traditional voice based communication is used where Public Switched Telephone Networks (PSTN)[8]. Such systems are expensive when the distance between the calling and called subscriber is large because of dedicated connection. The current trend is to provide this service on data networks. Data networks work on the best effort delivery and resource sharing through statistical multiplexing. Therefore, the cost of services compared to circuit-switched networks is considerably less. However, these networks do not guarantee faithful voice transmission. Voice over packet or Voice over IP (VoIP) systems have to ensure that voice quality does not significantly deteriorate due to network conditions such as packet-loss and delays.	In our Proposed system, Mel Frequency Cepstral Coefficients algorithm is extremely effective for voice feature extraction. It provides security to the user. The Mel Frequency Cepstral Coefficients (MFCC). After a test was done on 415 different voices with users from both genders, it was proven that the MFCC is more efficient than the LPCC and can reach up to an accuracy rate of 90% whereas the LPCC would hit 75%. Therefore, using MFCC for feature extraction proved to be the best solution. The MFCC are always real and contain information about the physical aspects of the signal. The distance between the recorded voice print and the saved voice print is computed and compared with the Universal

	Therefore, providing Toll Grade Voice Quality through VoIP systems remains a challenge. **Disadvantage:** 1. The problem of voice recognition occurs at only hardware level.	Voice print in order to obtain a likelihood ratio. 1. Mel Frequency Cepstral Coefficients 2. Voice Activity Detection (VAD)

IV. SOFTWARE REQUIREMENT SPECIFICATION

The SRSfully describes what the software will do and how it will be expected to perform on different environments. The purpose of the proposed system is to provide higher level of accuracy in identifying the speaker. And the multi-speaker feature enables the system to identify more than one speakerat a time.To implement the techniques that can help in improving the overall efficiency of the system.The primary focus is touse it as a biometric security in multiple smart devices.

Major Constraints: Major constraints are to detect the interaction between two work modules i.e. the testing and training modules. There are no specific constraints on models or manufacturers when using regular PC microphones, headsets or the built-in microphones in laptops, smartphones and tablets while recording the users voice as an input. However, these factors should be noted:

- The same microphone model is recommended (if possible) for use during both enrollment and recognition, as different models may produce different sound quality.
- Settings for clear sound must be ensured; some audio software, hardware or drivers may have sound modification enabled by default. For example, the Microsoft Windows OS usually has, by default, sound boost enabled.
- A quiet environment for enrollment and recognition.
- User Behavior and Voice Changes.

Product Functions:

Functional requirements and characteristics: The functional requirements and its characteristics is to design the products that satisfy their target users, a deeper understanding is needed of their user characteristics and product properties in development. Voice Recognition system should be able to recognize the voice of multiple speaker speaking simultaneously. It should be able to identify the speaker from different voice samples store in the database. While the nonfunctional requirements

Nonfunctional requirements:Voice Recognition system in Android will always have some tolerance as the entire processing of the voice signal requires better hardware and software capabilities. To improve the performance of the system we need to increase the number of frames to be processed at a time in testing phase so that the speaker will be identified in real time. Reliability and Accuracy of the system can be improved by giving quality training of data samples in training phase in noise free environment.

Specific Software and Hardware Requirements:

Sr. No.	Constraints	Type	Minimum Requirements
1.	System	Android Smartphone	3.0 and above
2.	Hardware	RAM	1 GB
3.		Storage	1 GB
4.	Software	Android Studio	
5.	Programming Language	JAVA	

V. IMPLEMENTATION DETAILS

In this paper, we propose an innovative technique for finding and Detecting Voice. There is no limit for development, improving and evaluating. Our future vision is to implement and extend this system to support multiple languages rather than Arabic. Also hope to add some services not included in this version of the system such as Prayer Time service which can be used to send notifications to user on their mobiles or email or run some voice when Prayer Time is coming. the advantages of the proposed model include:

Advantages:

1. Provide higher level of accuracy.
2. The multi-speaker feature enables the system.
3. Techniques that help to improve its efficiency.

a) SYSTEM ARCHITECTURE

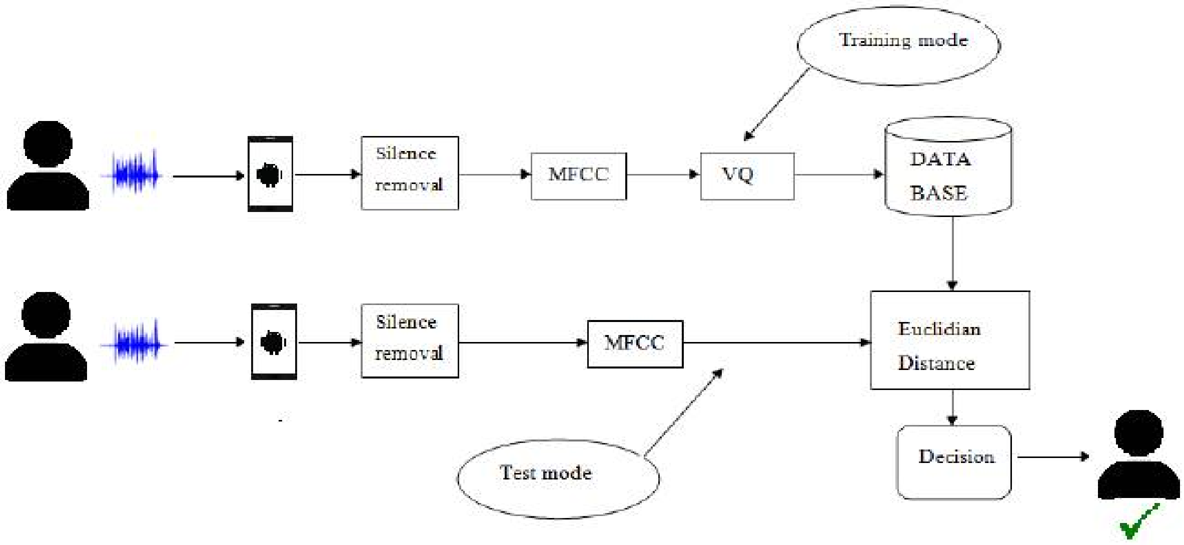

Figure 1: System Architecture

The System Architecture for Speaker Recognition Android Application is as shown in the figure below consist of two phases: 1. Training phase 2. Testing phase

1.Training phase: During this phase human voice is recorded with the microphone input from the system. The raw voice data is divided into frames of equal size and given to silence removal module. This module detects frames wherein there is no voice data and these frames are discarded. The remaining frames with voice data is given to MFCC module as input. The MFCC module makes use of mfcc algorithm and calculates 21 coefficients per frame which is then given as input to vector quantization system. The VQ system then categorizes these coefficients, links the username to coefficient datafile and creates an entry into database for storing the file into internal database of android.

2.Testing phase: During this phase human voice is recorded with the microphone input from the system. The raw voice data is divided into frames of equal size and given to silence removal module. This module detects frames wherein there is no voice data and these frames are discarded. The remaining frames with voice data is given to MFCC module as input[1]. The MFCC module makes use of mfcc algorithm and calculates 21 coefficients per frame which is then given as input to Euclidian distance module. This module calculates the likelihood of the recorded voice to the voice stored in database. If there is a match then respective username is displayed or else appropriate decision is taken for detecting the current person speaking using the decision module.

b) MFCC FEATURE EXTRACTION ARCHITECTURE:

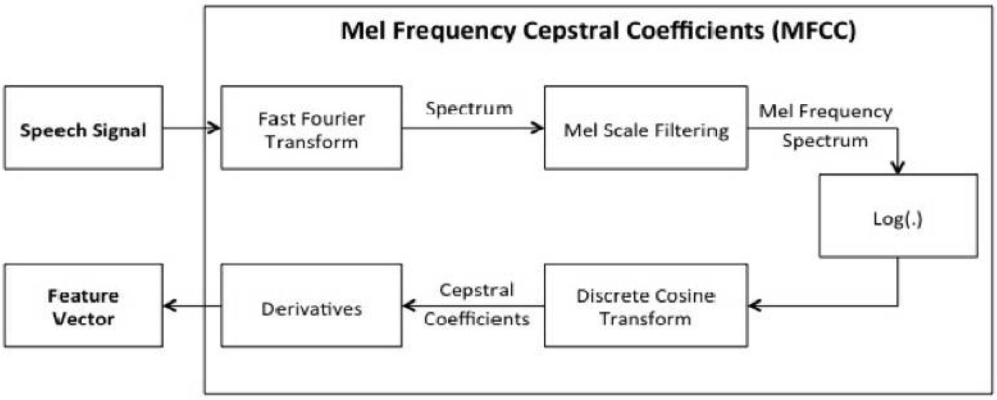

Figure 2: MFCC Feature Extraction

c) **PROPOSED SYSTEM FLOW:**The voice prints shown in figure 2 are created as such:

1. MFCC computed with 21 coefficients per frame.
2. Clustering is performed using the LBG variation of the K-means with K=16.
3. The clustered 16 x 21 vectors is encoded and stored in the database in the form of JSON.
4. The name of the individual serves as the primary key to the database record.
5. The entire records found in the database are then used and clustered to update the Universal Voice print which serves as an average model that will be used for comparison and likelihood estimation.

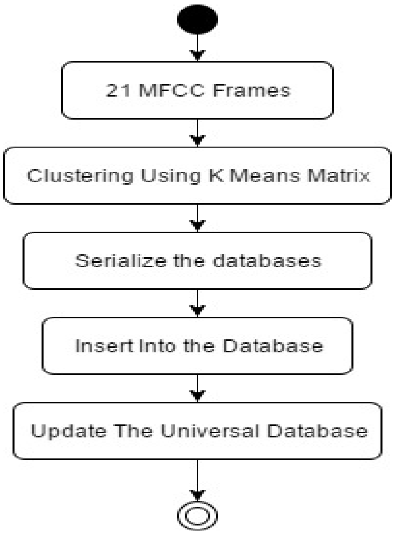

Figure 3:Voice Print Creation

d) ALGORITHM FOR RELEVANT FEATURE DISCOVERY USING ANDROID

Efficient Algorithms play important role in the relevant feature discovery from Voice using Android. The following steps explain the relevance feature Voice Extract.

MFCC ALGORITHM FEATURE EXTRACTION STEPS:

1. Frame the signal into 20-40 ms frames. 25ms is standard.
2. This means the frame length for a 16kHz signal is 0.025*16000 = 400 samples.
3. Frame step is usually something like 10ms (160 samples), which allows some overlap to the frames.
4. The first 400 sample frame starts at sample 0, the next 400 sample frame starts at sample 160 etc. until the end of the speech file is reached. If the speech file does not divide into an even number of frames, pad it with zeros so that it does.
5. We take the absolute value of the complex Fourier transform, and square the result. We would generally perform a 512 point FFT and keep only the first 257 coefficients.
6. Compute the Mel-spaced filter bank.
7. This is a set of 20-40 (26 is standard) triangular filters that we apply to the periodogram power spectral estimate.
8. Our filter bank comes in the form of 26 vectors of length 257.

9. Each vector is mostly zeros, but is non-zero for a certain section of the spectrum.
10. To calculate filter bankenergies, we multiply each filter bank with the power spectrum, then add up the coefficients. Once this is performed we are left with 26 numbers that give us an indication of how much energy was in each filter bank.
11. Take the log of each of the 26 energies from step 10. This leaves us with 26 log filter bank energies.
12. Take the Discrete Cosine Transform (DCT) of the 26 log filter bank energies to give 26 cepstral coefficients. For ASR, only the lower 12-13 of the 26 coefficients are kept.
13. The resulting features (12 numbers for each frame) are called Mel Frequency Cepstral Coefficients.

Steps:

1. Window size: 25ms
2. Window shift: 10ms
3. FFT size 512, 1024 or 2048
4. Pre-emphasis coefficient: 0.97
5. P=24 to P=30 filters in the Mel bank
6. 12 MFCC (Mel frequency cepstral coefficients)
7. 1 energy feature
8. 12 delta MFCC features
9. 12 double-delta MFCC features
10. 1 delta energy feature
11. 1 double-delta energy feature
12. Total 39-dimensional features

VI. EXPERIMENTAL SETUP AND RESULTS

1. RESULTS:

For MFCC algorithm with Euclidian distance the speaker recognition system is able to identify 18 out of 20 males speaking for single speaker recognition i.e. gives an accuracy of 90.3%. Similarly, for multispeaker recognition system the accuracy is around 34%. On the other hand, MFCC algorithm with dynamic time wrapping the speaker recognition system is able to identify 14 out of 20 males speaking for single speaker recognition i.e. gives an accuracy of 72.1%. Similarly, for multispeaker recognition system the accuracy is around 14%. On comparison of both system evaluation results, use of MFCC with Euclidian distance is proposed in the given system.

Comparison Results of MFCC with Euclidean Distance and With Dynamic Time Wrapping(DTW).

	MFCC WITH EUCLIDEAN DISTANCE	MFCC WITH DYNAMIC TIME WRAPPING
20 Males	90.3%	72.1%

2. RESULT EVALUATION:

Comparison among multiple speakers was also performed with a subset of the original database taking a sample of 20 individuals. Each test case was repeated 10 times with thesubjects changing between each test.

Test Cases	Results
2 Males Speaking at same time randomly	80%
2 Females Speaking at same time randomly	85%
3 People Speaking at same time randomly	87%

The results look promising with high accuracy as shown in the above table.

VII.CONCLUSION

This paper shows that the speaker recognition application can be accurately adapted to work on software, specifically on an Android mobile phone. The results obtained already look promising and with proper care, a voice print can be used as an additional security measure on our mobile phones, in addition to pin locking and finger-print locking. The system is intelligent enough to extract the voice characteristics from the voice sample and later in future the system can be enhanced by integrating it in the Android Operating system and using it as a biometric security, along with that, this system can also be integrated into robots for identifying its master. The system can be made more robust and efficient by providing more training datasets and using the machine learning concepts.

REFERENCES

[1] Sutar Shekhar "Intelligent Voice Assistant Using Android Platform", International Journal of Advance Research in 1Computer Science and Management Studies, Volume 3, Issue 3, March 2015.
[2] Saroj B. Jadhav, Jayshree Ghorphade and Rishikesh Yeolekar "Speech Recognition in Marathi Language on Android O.S", International Journal of Research in Computer and Communication Technology, Vol 3, Issue 8, August – 2014
[3] B. Raghavendhar Reddy, "Speech to Text Conversion using Android Platform", International Journal of Engineering Research and Applications (IJERA), Vol. 3, Issue 1, January -February 2013, pp.253-258
[4] Shahenda Sarhan, "Smart Voice Search Engine", International Journal of Computer Applications (0975 – 8887) Volume 90 – No.3, March 2014
[5] Teenu Therese Paul, "Voice Recognition Based Secure Android Model For Inputting Smear Test Results", International Journal of Engineering Sciences & Emerging Technologies, Dec. 2013.
[6] Dr. Stuart Dinmore, "Voice To Text Transcription Of Word Recording", show me the learning, Nov-27-30.
[7]George Frewat ,CharbelBaroud and andRoy Sammour,"Android Voice Recognition Application with Multi Speaker Feature", Department of Computer and Communication Engineering ,Faculty of Engineering ,April -20-16.
[8] RadosławWeychanAgnieszkaStankiewiczTomaszMarciniakAdamDabrowski "Improving of Speaker Identification from Mobile Telephone Calls", Poznan University of Technology, Chair of Control and Systems Engineering, Division of Signal Processing and Electronic Systems, ul. Piotrowo 3a, 60-965 Pozna´n, Poland,2014.