Voice Recognition Technique: A Review

Gunjan Chugh[1] Rakshita Saini[2] Nancy Gupta[3] Raghav Manchanda[4] Gunjan[5]
[1]Assistant Professor [2,3,4,5]Student
[1,2,3,4,5]Dr. Akhilesh Das Gupta Institute of Technology and Management, India

Abstract— Voice Recognition is a biometric technology. It is used to recognize the voice of a particular individual. When the voice is generated, the speech waves from the different voices serve as the basis of identification of speaker. Using voice recognition for security purpose where a person can enter his/her voice for authentication is one of the most important applications. Each voice has its unique characteristics called feature & extracting these features from an individual's voice is called feature extraction. The extracted features are further compared with the saved voices in the database for matching. This paper provides review of various voice recognition systems.

Keywords: Automatic Speech Recognition (ASR), Articulatory Movement, Motor Control, Speech Generation, Vocal Tract, Voice Recognition, MFCC (Mel Frequency Cepstral Coefficient), DCT (Discrete Cosine Transform), FFT (Fast Fourier Transform), IFT (Inverse Fourier Transform), LPC (Linear Prediction Coefficient), LPCCs (Linear Prediction Cepstral Coefficients)

I. Introduction

Voice recognition or speaker recognition refers to the identification or confirmation of the identity of an individual based on his voice. It's an automated method. Speaker Recognition are the best-known commercialized forms of voice biometrics. Basically, identification or authentication using speaker recognition consists of four steps which includes recording, feature extraction, pattern matching and decision [1][10].

The difference between voice biometrics and other biometrics is that voice biometrics are the only commercial biometrics that process acoustic information. Most other are image-based. Speaker recognition is a task of validating a user's claimed identity by using the characteristics which are extracted from their voices [1].

The unique characteristics, such as speed, tone, pitch, dialect etc. associated with an individual's voice are captured by voice recognition system and a non-replicable voiceprint is created which is also known as a speaker model. This derived voiceprint through mathematical modelling of multiple voice features is nearly impossible to replicate. A voiceprint is therefore a secure method for authenticating an individual's identity as it cannot be stolen, duplicated or forgotten like passwords or tokens [1].

II. Voice production mechanism

There are various organs which are involved in the production of speech and sound by the human beings. These organs are flexible and their shape and size changes on the command of control signals from motor control unit (driven by human brain), as per the type of speech and sound which is to be produced. Lungs are responsible for providing the necessary air force for the generation of sound in the form of acoustic wave. The air passes through the vocal tract, vocal cords, glottis, epiglottis, and other organs in the mouth and in the end comes out through mouth and nasal cavities in the form of acoustic wave. Various organs involved through which the air passes during the process of generation of speech and sound are shown in the fig(i) [3].

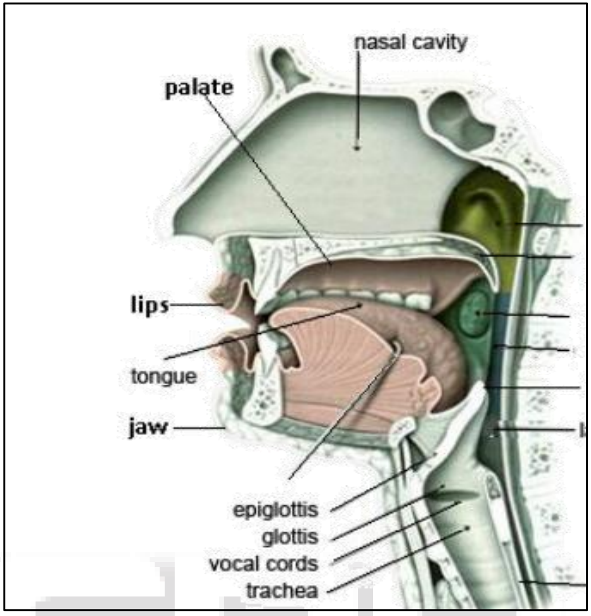


Fig. 1: Human Vocal Apparatus [3]

III. How the technology works

Every person's voice is unique and the factors that distinguishes it uniquely are pitch, tone and volume while there are other factors too that contributes to this such as the mouth, nose, teeth and throat which are referred as articulators. On the other hand, size, shape and tension of the vocal cords also makes the voice distinct. Different positions of the articulators create different sounds [1].

Spectrogram is the visual representation of the voice. It helps in the analysis as it displays the time, frequency of vibration of the vocal cords (pitch), and amplitude (volume) [1].

IV. Applications of voice recognition

Forensics department: There are certain forensics situations where only audio evidence is available to investigators, and that's when voice biometrics can be deployed to great effect [11]. Audio forensics is the field of forensic science that is related to the acquisition, analysis, and evaluation of sound recordings which in turn can be presented as an admissible evidence [12].

Transaction Authentication: One of the most popular applications of biometrics is in mobile payments and voice recognition has also made its way into this competitive arena.[12]. It is widely used in user verification for e-commerce and m- commerce [2].

Access control: Voice Recognition system provide access control to various services like automation in cars, homes and mobile phones by voice command [2].

V. Methodology

Enrolment and verification are the two phases of speaker recognition system. In the enrolment phase the speaker's voice is recorded and a number of features are extracted and voice print is created. In the verification phase, a previously created voice print is compared against a speech sample or "utterance" [1].

Speaker identification system falls into two categories, text-dependent and text-independent. If the text spoken by the speaker is identical to the text stored during training phase, that is, if there is dependency on the text for recognition purpose then it is known as text dependent voice recognition system. On the other hand, if any random text spoken by the speaker is used for voice identification, it is called as text independent voice recognition system [2].

VI. General Speaker Recognition System Architecture

The voice recognition system can be categorized into two ways

– Speaker Identification: The process of identifying a voice of a given speech from the group of speakers is called speaker identification [2].

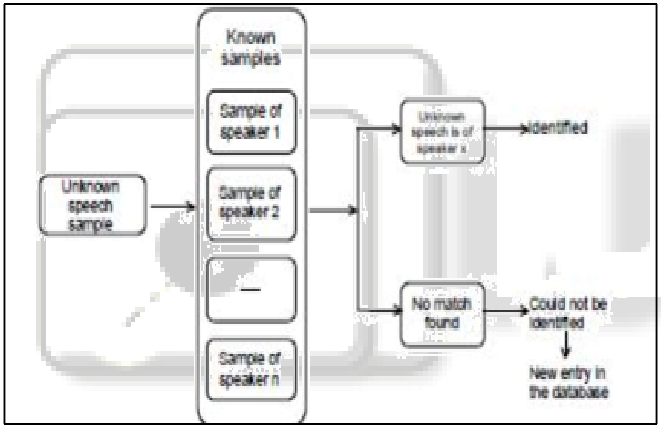


Fig. 2: Speaker Identification [2]

– Speaker Verification: In this process the claim of the speaker is accepted or rejected and thus in turns it verifies the speaker [2].

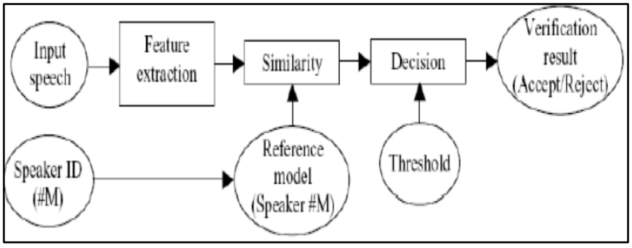


Fig. 3: Speaker Verification [13]

VII. Types of Voice Recognition Systems

– Speaker dependent system: In this voice recognition needs training before using, which requires us to read a series of words and phrases [8].

– Speaker independent system: Here the voice recognition software recognises most of the users without any training [8].

– Discrete speech recognition: The user is required to pause between words so that the speech recognition can identify each separate word [8].

– Continuous speech recognition: It allows the users to speak naturally while the content is determined by the computer [7][8].

– Natural language: The speech recognition not only understands the voice, but answers to the questions or other queries are also returned [8].

– Automatic Speech Recognition: Automatic speech recognition (ASR) is used to identify and process human voice. It identifies the words a person has spoken or authenticates the identity of the person speaking into the system. It is also known as automatic voice recognition (AVR), voice-to-text or simply speech recognition [9].

VIII. Feature Extraction Technique

Feature extraction is the process of recognition of the voice input of the speaker and its pattern matching process. There are various algorithms that are used to extract features.

A. Mel Frequency Cepstral Coefficient (MFCC)

Mel scale is used as an input to derive the cepstral representation of an audio clip. Cepstrum and the mel-frequency cepstrum differs as in MFC, the frequency bands are positioned logarithmically (on the Mel scale) which approximates to the response of human auditory system more closely than the frequency bands which are linearly-spaced obtained directly from FFT (fast Fourier transform) or DCT (discrete cosine transform) [4][6].

1) Cepstrum

On reversing the first four alphabets of the word spectrum we get the word cepstrum. There are different types of cepstrum such as complex, real, power and phase cepstrum. Power cepstrum has applications in human speech analysis [4].

2) Mel Frequency Cepstral Coefficients

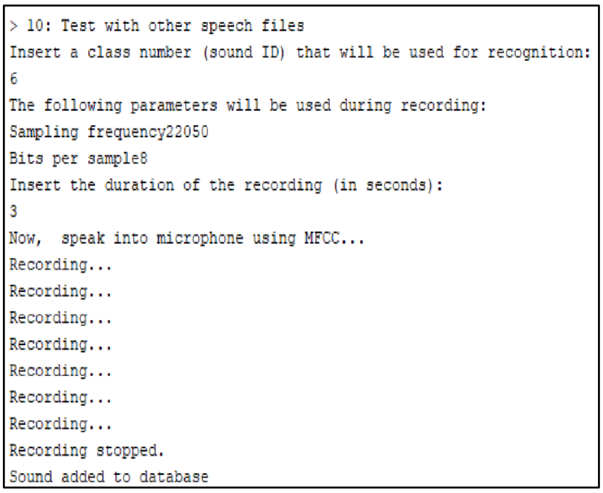
Before MFCCs were introduced, Linear Prediction Coefficients (LPCs) and Linear Prediction Cepstral Coefficients (LPCCs) were used as main feature type for ASR. Mel spectrum coefficients may be converted to the time domain using the Discrete Cosine Transform (DCT) since Mel spectrum coefficients and their logarithms too are real numbers. The equation below is used to calculate MFCC [4][7].

$$C_n = \sum_{k=1}^{K}(log S_k)\left[n\left(k-\frac{1}{2}\right)\frac{\pi}{2}\right]$$

Fig. 4: Equation [4]

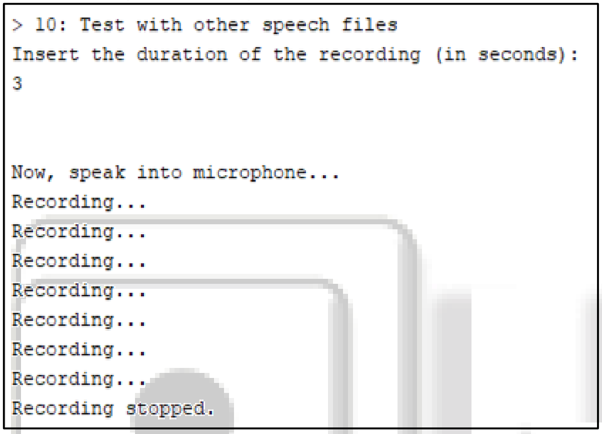
In the equation n varies from 1 to k. The first component C_0 is excluded from the DCT because it depicts the mean value of the input signal which carries less information specific to the speaker. This set of coefficients is called an acoustic vector [4].

IX. Proposed Work



```
> 10: Test with other speech files
Insert a class number (sound ID) that will be used for recognition:
6
The following parameters will be used during recording:
Sampling frequency22050
Bits per sample8
Insert the duration of the recording (in seconds):
3
Now,  speak into microphone using MFCC...
Recording...
Recording...
Recording...
Recording...
Recording...
Recording...
Recording...
Recording stopped.
Sound added to database
```

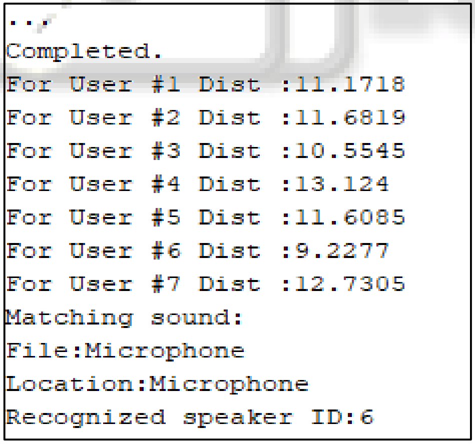
Fig. 5: Adding Sound to the Database



```
> 10: Test with other speech files
Insert the duration of the recording (in seconds):
3


Now, speak into microphone...
Recording...
Recording...
Recording...
Recording...
Recording...
Recording...
Recording...
Recording stopped.
```

Fig. 6: Recording sound for recognition purpose



```
...
Completed.
For User #1 Dist :11.1718
For User #2 Dist :11.6819
For User #3 Dist :10.5545
For User #4 Dist :13.124
For User #5 Dist :11.6085
For User #6 Dist :9.2277
For User #7 Dist :12.7305
Matching sound:
File:Microphone
Location:Microphone
Recognized speaker ID:6
```

Fig. 7: Speaker Recognized

X. Conclusion

The voice recognition system works in view to authenticate the input voice of the speaker. It compares the input voice with the available voices in the database by correlating the intricacies of the voices. These features are extracted by using certain algorithms such as MFCC in MATLAB. However real-time applications are liable to issues such as noise disturbance. Our main attempt is to gain at least some understanding in the domain- how the voice mechanism works, its applications and how to recognize the voice.

References

[1] Saquib, Zia & Salam, Nirmala & Nair, Rekha & Pandey, Nipun & Joshi, Akanksha. (2010). A Survey on Automatic Speaker Recognition Systems. Communications in Computer and Information Science. 123. 134-145.

[2] Nisha, “Voice Recognition Technique: A Review”, International Journal for Research in Applied Science & Engineering Technology(IJRASET) , Vol. 5 , India, May 2017.

[3] Harish Chander Mahendru , “Quick Review of Human Speech Production Mechanism” , International Journal of Engineering Research and Development, Vol. 9,pp. 48-54, India, January 2014

[4] Supriya Tripathi and Smirti Bhatnagar, “Speaker Recognition” , Third International Conference on Computer and Communication Technology, India , 2012.

[5] Kurzekar, Pratik & Deshmukh, Ratnadeep & Waghmare, Dr. Vishal & P Shrishrimal, Pukhraj. (2014).Contiguous speech recognition system: a review. Asian Journal Computer Science & Information Technology. 4. 62-66.

[6] Bala, Anjali & Kumar, Abhijeet. (2010).Voice command recognition system based on MFCC and DTW International Journal of Engineering Science and Technology. 2. 7335-7342.

[7] Karthik Selvan. Aju Joseph and Anish Babu K. K, “Speaker Recognition System For Security Applications”, IEEE Recent Advances in Intelligent Computational Systems(RAICS) ,2013.

[8] Voice Recognition. (2019, May 1). Retrieved from Computer Hope: https://www.computerhope.com/jargon/v/voicreco.htm

[9] What is automatic speech recognition? (n.d.) Retrieved from techopedia: https://www.techopedia.com/definition/6044/automatic-speech-recognition-asr

[10] Speaker Recognition. (n.d.). Retrieved from Biometric Solution: http://www.biometric-solutions.com/speak er-recognition.html

[11] Applications of voice recognition. (2016, May 18). Retrieved from Find biometrics: https://findbiometrics.com/4-applications-voice-recognition-305180/

[12] Audio forensics. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Audio_forensics

[13] an automatic speaker recognition system. (n.d.). Retrieved from http://minhdo.ece.illinois.edu/teaching/speaker_recognition/speaker_recognition.html