International Journal of Research in Engineering, Science and Management
Volume-2, Issue-6, June-2019
www.ijresm.com | ISSN (Online): 2581-5792

Text Independent Speaker Recognition System

Monika B. Duratkar[1], Anup A. Ladge[2], Rahul Wantmore[3]
[1,2]*Student, Department of MCA, NCRD's Sterling Institute of Management Studies, Navi Mumbai, India*
[3]*Assistant Professor, Dept. of MCA, NCRD's Sterling Institute of Management Studies, Navi Mumbai, India*

***Abstract*: This paper tends to provide an overview of automatic speaker recognition technology, with a special importance on text-independent recognition. The idea of the Speaker Recognition is to implement a recognizer which can identify a person by processing his/her voice. The basic goal is to recognize and classify the speeches of different persons. An automatic text-independent speaker recognition system is suitable for verification and identification purposes. Voice recognition systems helps the consumers to interact with the technology simply by speaking to it, enabling hands-free requests, reminders and other simple tasks. With AI, machine learning and consumer acceptance have matured, the uses for voice recognition have grown quickly. The text-independent speaker recognition system is based on spotting the vowels of the test utterance, extracting parameter vectors and classifying them into a speaker-dependent reference database.**

***Keywords*: Speaker recognition, Speech recognition, Voice recognition, Speech analysis, Speech processing.**

1. Introduction

Voice recognition or Speaker recognition is referred to the automated method of identifying or confirming the identity of an individual based on his/her voice. Speech recognition is defined as the process of capturing spoken words using a microphone or telephone and then converting them into a digitally stored set of words. The voice can be both a physiological and a behavioural biometric factor:

The physical shape of the subject's voice tract is the physiological component of speaker recognition. The physical movement of jaws and tongue is the behavioural component of speaker recognition.

Speech is a diverse field with many applications. When speech signal travels, the speech has to be recognized, language has to be recognized and speaker has to be recognized. Speaker identification helps in determining which registered speaker has spoken. Speaker verification accepts or rejects the claimed identity of a speaker.

2. Literature Review

Speech signal is the most ancient form of communication. Speech signal provides extensive information like the message which is conveyed from the speaker to the listener as well as gender, emotion, language, health condition and the identity of the speaker. Speech signal provides the linguistic information present in the signal whereas speaker recognition deals with preserving features related to the identity of the speaker. Speaker recognition systems are widely used in biometric systems using speech as an access control parameter and in security systems. Speaker recognition are of two types, viz. speaker identification and speaker verification. Speaker identification is divided into text independent and text dependent. Speaker utters the same word or phrase during training and testing of text dependent speaker identification while speaker has the liberty of using different words or phrases for text independent speaker identification.

3. Problem Definition

The current problems faced by the users using the voice biometric is that the accent through which the user speaks may not be recognizable to the system. Due to which users are not able to use their own device. There are higher chances that mimicry artist who has the voice resemblance of the authorized person can use the authorized person's device. Another problem faced by the users are that suppose when a user is having health issue related to his voice, the system is not able to recognize the voice of the authorized user. There also arises problem when an authorized person has the recording of the authorized person which he or she can use to access the authorized person's device. So this research paper intends to solve all these problems.

4. Objective

The objective of this research paper is to provide all possible solutions to the users who are facing the problems while accessing their any device using voice recognition. So the solution that have been provided through this research paper is the introduction of speaker recognition system. Speaker recognition system generally consists of a feature extraction system, speaker modelling system and a classification system. The speaker extraction system mainly used to extract the voice of speaker. Speaker recognition system intends to solve the problems that have been faced by the users using voice recognition system.

5. Research Methodology

The Block Diagram of Speaker Recognition system is as shown in Fig. 1. Speaker recognition system generally consists of a feature extraction system, speaker modelling system and a classification system. Pre-processing aids in removing

unvoiced part from the speech signal.

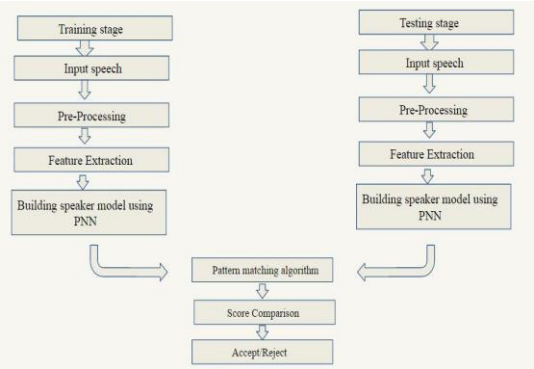

Fig. 1. Block diagram of speaker recognition system

Training stage is used to model the preprocessed speech and store in the form of a neural network which is employed during the testing stage for speaker recognition.

Feature extraction MFCC is known to be an effective method in modelling speaker information. In this paper we propose a novel method of framing signals into overlapping frames based on the energy overlap ratio rather than evaluating system for a particular frame overlapping ratio.

A. *Pre-processing*

Fig. 2. Pre-processing

Pre-processing involves converting the analog input speech signal into a digital signal and separating the voiced part from the unvoiced/silenced part. As most of the speaker dependent information lies in the voiced part of the speech signal, so the feature extraction module converts the resultant signal into data which is used for speaker identification.

The block diagram of pre-processing is shown in fig. 2.

1) Pre-emphasis, 2) Energy efficient framing, 3) Windowing

1) Pre-emphasis: Process of applying a high pass filter to boost high frequencies and compensate for the attenuation at high frequencies caused by glottal voice source y(n) = x(n) + α∗x(n−1) where α is varied in the range (0.9,1)

2) Energy efficient framing: Speech signal is a non-stationary signal. Speech signal is divided into overlapping frames to avoid loss of information. The frame overlap is dependent on the pre-decided energy overlap ratio which is given by

Energy Overlap ratio = Overlap energy Total Energy of a discrete signal is given by

Energy = lim N→∞

N Xn = −N

|x(n)|2

3) Windowing: The purpose of windowing is to create smooth and less distorted spectrum. Hamming window is the most popularly used window function.

B. *Feature extraction*

Fig. 3. MFCC feature extraction

Feature Extraction is the important part of speech recognition. It plays important role in separating one speech from other. The technique of computing MFCC is based on the short-term analysis, and thus from each frame a MFCC vector is computed.

FFT: Fast Fourier Transform converts time domain signal in frequency domain signal FFT is applied to every speech signal and its magnitude is calculated.

Mel-scale filter bank: It has been proved that human ears are more sensitive and have higher resolution to low frequency compared to high frequency. Hence, the filter bank is designed to emphasize the low frequency over the high frequency. Mel scale is linear below 1kHz and logarithmic above 1kHz.

$$f_{mel} = 2595 * \log_{10}\left(1 + \frac{f_{lin}}{700}\right)$$

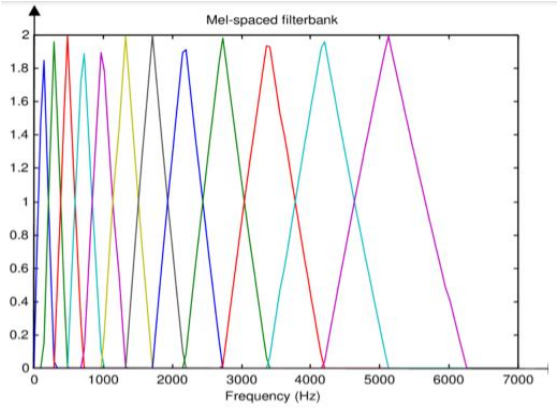

Fig. 4. MFCC

Discrete Cosine Transform (DCT): Output of Mel-scale filter bank then passed through logarithm block. This block is used for normalization purpose. After that normalized signal is then passed through DCT block that de-correlates the log energies of filters. Finally, the output of DCT block provides the MFCC coefficient values.

Speaker modelling: PNNs are much faster, more accurate and are relatively insensitive to outliers. PNN networks generate accurate predicted target probability score. PNN can be said to be a special case of Back Propagation Neural Network (BPNN) where the sigmoid function is replaced by an exponential

function, providing advantages of increased speed and real timely computation over other ANN architectures. Probabilistic Neural Network is a feed forward neural network, mainly used for classification and pattern recognition problems. In PNN algorithm, the parent probability distribution function (PDF) of each class is approximately identified by a Parzen window and a non-parametric function. Then, by using PDF of each class, the class probability of a new input data is estimated and Bayes' rule is then employed to allocate the class with highest posterior probability to new input data. Using this method, the probability of misclassification is minimized.

6. Analysis and Findings

We observe that using PNN model the accuracy decreases as we increase the number of speakers. The experiment was carried out for 80% of energy overlap ratio. Also for GMM based model, better accuracy is observed. Figure shows various results.

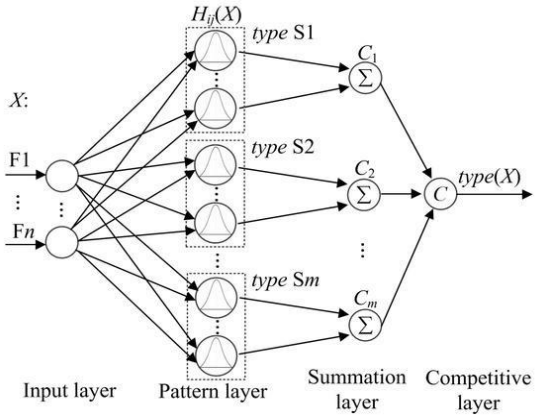

Fig. 5. Probabilistic neural network

Table 1
Results

Type of Methodology used	Type of Database	Inputs	Percentage Accuracy
PNN	Real time	2 speakers (M.F)	82.23
PNN	Real Time	3 speakers (2M.F)	73.17
PNN	Stored	2 Speakers (2M)	94
GMM	Real Time	2 Speakers (2M)	97.83
GMM	Real Time	3 Speakers (2M.F)	94.667

7. Limitations

Speech Recognition works best only when the microphone is close to the user. More distant microphones will tend to increase the number of errors.

If the system can only recognize English language, then user would not be able to speak other regional languages as a part of speech recognition.

The user must have the knowledge of using the speaker recognition system.

8. Future Scope

Entertainment: Voice recognition can be used for changing TV or radio channels, open and close screens, and play movies. It can also help personalize customer experience. For instance, services such as Netflix can be personalized by determining the age of the user through voice analysis, enabling them to access age-appropriate content.

Healthcare: The global healthcare biometric market is expected to reach USD 14.5 billion by 2025. In an industry where data security is paramount, physicians can use the voice biometrics to record patient's health conditions directly into the system and securely retrieve patient's personal history. This can provide benefit to patients who need to share medical records between various doctors. The system can provide help to reduce fraud for providers and payers by automating payment collection, and improve patient satisfaction by offering an additional payment option.

Banking: Customers can use the voice authentication for operating bank lockers. Banks can leverage the system to enable highly secure and advanced voice-based payments. With fraud on the rise, credit card companies and banks such as Citibank, use voice biometrics to proactively identify fraudsters and authenticate callers at their call centers.

Education: Educational institutions can use the voice recognition to provide flexibility to students with visual disability, helping them take online exams using voice authentication.

9. Conclusion

The use of MFCC and its Delta Derivatives calculated using Mel spaced Gaussian Filter Banks with Energy efficient framing. If overlap percentage is kept fixed and features are extracted, it is likely possible that some frames might have less energy than other frames. It is possible to have frames with fixed energy values by keeping a fixed threshold criteria (say 80%) for overlapping of frames. Length is thus kept variable. The accuracy of the system can be further tested by varying the various parameters of the speech signal like frame size, number of filter banks used in order to create a more robust speaker recognition system.

References

[1] S. G. Bagul and R. K. Shastri, "Text independent speaker recognition system using GMM," *2013 International Conference on Human Computer Interactions (ICHCI)*, Chennai, 2013, pp. 1-5.
[2] https://www.slideshare.net/deepeshlekhak/text-independent-speaker-recognition-system
[3] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," in *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72-83, Jan. 1995.
[4] A. Roland, M. Carey and Harvey Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification Systems," Digital Signal Processing, vol. 10, no. 1–3, pp. 42-54, January 2000.