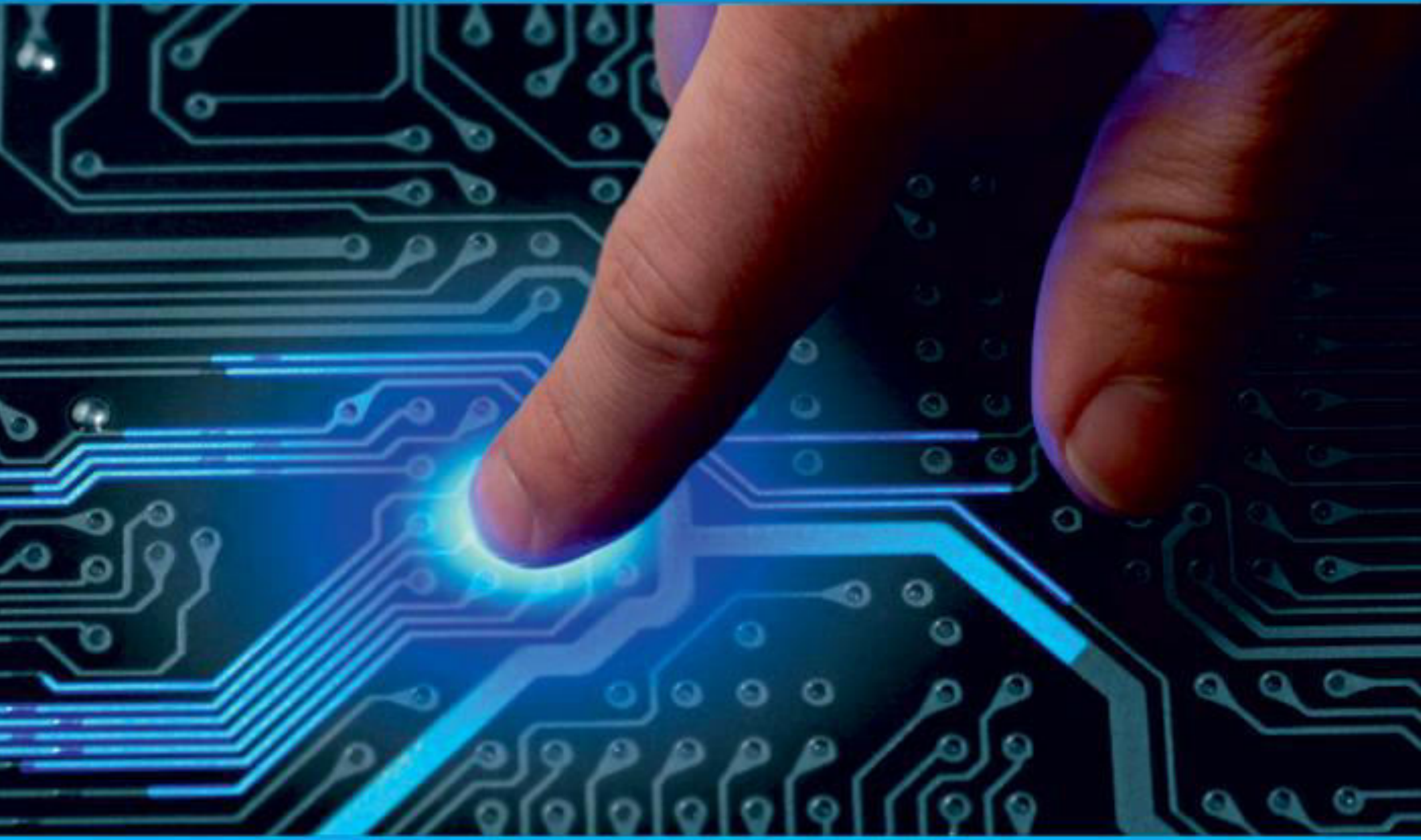
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 3, March 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

9940 572 462 6381 907 438 ijircce@gmail.com www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203200|

A Survey on Speaker Recognition: Exploring Feature Extraction and Classification Methods

M Sai Akash, S Praveena

I M. Tech, DECE, Department of ECE, MGIT, Hyderabad, India

Associate Professor, Department of ECE, MGIT, Hyderabad, India

ABSTRACT: Speech processing is becoming increasingly popular for enhancing security measures. Speech is widely utilized for authentication purposes, and speaker recognition plays a key role in verifying and identifying individuals based on their voice. It is important to distinguish between speech recognition systems and speaker recognition systems. Speaker recognition is extensively applied across various domains, including industries, hospitals, and laboratories. Its key advantages include enhanced security, ease of implementation, and user-friendliness.

In areas where high security is crucial, speaker recognition is one of the most commonly employed techniques. It is also recognized as a popular biometric method. This paper provides an overview of various techniques used in speaker recognition applications, such as LPC, LPCC, and MFCC, and explores different classifiers, including DTW, GMM, VQ, and SVM. The primary goal of this review is to summarize widely recognized techniques for speaker recognition systems.

KEYWORDS: Key Words: Speaker recognition, Mel frequency cepstral coefficients(MFCC), Linear predictive coding (LPC), Linear Predictive Cepstral Coefficients (LPCC), Gaussian Mixture Model(GMM), Vector Quantization(VQ), Support Vector Machine(SVM), Dynamic Time Warping(DTW)

I. INTRODUCTION

Speech signals convey multiple levels of information [14]. They can be utilized in various applications, such as speech recognition, speaker recognition, or voice command systems [3]. Among these, speaker recognition plays a significant role in numerous speech processing applications, particularly in the domains of security and authentication. In today's world, security is a critical requirement.

It is essential to differentiate between speech recognition and speaker recognition, as they are closely related but distinct systems [14]. Speech recognition focuses on identifying the content of what is spoken, whereas speaker recognition aims to determine the identity of the speaker. Speech recognition involves automatically identifying the spoken words based on the information embedded in the speech signal [3]. Speaker recognition, on the other hand, is divided into two categories: speaker identification and speaker verification. The primary objective of speaker recognition is to identify the speaker by extracting, characterizing, and analyzing the information contained in the speech signal [14]. Speech recognition systems are further categorized as either speaker-dependent or speaker-independent.

Machines process human speech through feature extraction and feature matching. The fundamental model of a speaker recognition system is depicted in Figure 1 [3].

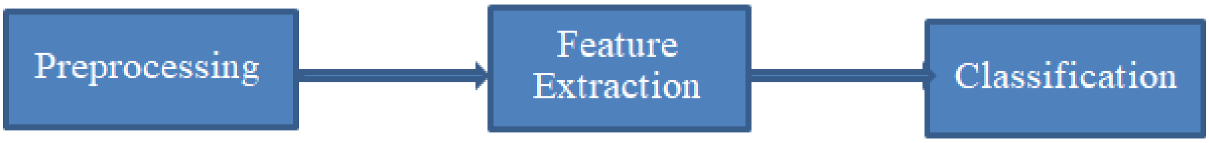


Figure 1: Basic Model of a Speaker Recognition System

The speaker recognition process involves three main steps. The first step is pre-processing, which removes silent segments from the speech signal [3]. During the feature extraction stage, various techniques such as Linear Predictive Coding (LPC), Linear Predictive Cepstral Coefficients (LPCC), and Mel Frequency Cepstral Coefficients (MFCC) are employed. In the classification stage, different classifiers are used, including Support Vector Machines (SVM), Vector Quantization (VQ), Gaussian Mixture Models (GMM), and Dynamic Time Warping (DWT).

II. RELATED WORK

Tiwari et al. [1] employed feature extraction techniques such as LPC, LDB, and MFCC, alongside the VQ classifier. Their findings revealed that MFCC, when used with a Hanning window and 32 filters, demonstrated higher efficiency. Additionally, the density-matching property of vector quantization was noted as particularly powerful in their research. K. Kaur et al. [2] utilized LPC, LPCC, and MFCC for feature extraction, paired with classifiers such as VQ, GMM, SVM, DWT, and HMM. They concluded that the MFCC technique aligns more consistently with human hearing compared to LPCC. Among the classification models, GMM was identified as the best due to its superior classification accuracy and minimal memory usage. K. Dhameliya et al. [3] focused on MFCC and LPC for feature extraction and employed GMM and ANN for classification. They suggested that combining one or more techniques could enhance speaker recognition performance.

The literature survey highlights that speaker recognition is one of the most commonly employed techniques in areas where high security is essential. It is also a widely recognized biometric method [8]. Although other biometric methods exist, speech-based techniques often yield superior results. Various feature extraction methods are available, including LPC, LPCC, and MFCC. Among these, MFCC is preferred for its effectiveness, particularly at lower filter orders.

III. TECHNIQUES IN SPEAKER RECOGNITION

Speaker recognition can be classified into several categories based on different criteria.

3.1Open Set vs Closed Set

This classification is based on the set of trained speakers available in the system. An open set system can accommodate any number of trained speakers, while a closed set system only includes a predefined group of registered users [8].

3.2 Speaker Identification vs Speaker Verification

Speaker identification involves determining which registered speaker is providing a given utterance from a set of known speakers. In contrast, speaker verification is the process of accepting or rejecting a speaker's identity claim [8]. Speaker verification is typically faster than speaker identification.

3.3Text Dependent vs Text Independent

In a text-dependent system, the same text is spoken during both the training and testing phases. On the other hand, a text-independent system places no restrictions on the text spoken [2]. Recognition accuracy tends to be higher in text-dependent systems compared to text-independent systems.

IV. FEATURE EXTRACTION TECHNIQUES

Various techniques are employed for feature extraction in speech processing, including Linear Prediction Coding (LPC), Linear Predictive Cepstral Coefficients (LPCC), and Mel-Frequency Cepstral Coefficients (MFCC).

4.1 Linear Predictive Coding (LPC)

LPC operates under the assumption that a speech signal is produced by a buzzer at the end of a tube. It analyzes the speech signal by estimating the formants, which describe the resonance frequencies of the vocal tract. LPC removes the formant effects from the signal and estimates the intensity and frequency of the remaining "buzz." The technique used to eliminate formants is called inverse filtering, and the remaining signal is termed the residue [8]. However, one drawback of LPC is its performance degradation in the presence of noise [2].

4.2 Linear Predictive Cepstral Coefficients (LPCC)

LPCC is a widely used technique for feature extraction from speech signals. LPC parameters effectively describe the energy and frequency spectrum of speech frames [8]. LPCC offers a smoother spectral envelope and a more stable representation compared to LPC [2]. However, one limitation of LPCC is that it uses linearly spaced frequency bands, which may not be optimal for certain speech processing tasks [2].

4.3 Mel-Frequency Cepstral Coefficients (MFCC)
Introduced by Davis and Mermelstein in the 1980s, MFCC is one of the most popular and widely used techniques for feature extraction in speech processing. In speech and speaker recognition systems, feature extraction is the first crucial step. It involves identifying the components of an audio signal that are useful for recognizing the content while discarding irrelevant information, such as background noise or emotional influences. The core principle of MFCC is based on the use of a filter bank to capture perceptually relevant features of the speech signal [2].

The process of extracting MFCC involves the following steps:

4.3.1 Pre-emphasis

Pre-emphasis is a simple signal processing method that boosts the amplitude of higher-frequency components while attenuating the lower frequencies [6]. Higher frequencies carry more important information for distinguishing speech signals.

4.3.2 Framing

Since audio signals continuously change over time, they are divided into smaller, manageable segments or frames for analysis. If the frames are too short, the signal may not contain enough samples for reliable spectrum estimation. Conversely, longer frames may result in inaccurate representation due to changes in the signal over time. Typically, frames are defined by a number of samples, with adjacent frames overlapping [6].

4.3.3 Windowing

To minimize signal discontinuities at the edges of the frames, windowing techniques are applied. A common technique is applying a Hamming window to each frame, which ensures the continuity of the signal at both ends of the frame.

4.3.4 Fast Fourier Transform (FFT)

FFT is used to transform each frame from the time domain into the frequency domain. By applying FFT to each frame, the magnitude of the frequency response of the signal can be obtained, assuming the signal is periodic and continuous when wrapped around.

4.3.5 Mel Filter Bank Processing

Since human perception of frequency is nonlinear, the output from the FFT is processed by a set of triangular bandpass filters, which simulate the way the human ear perceives sound. These filters help capture the log energy of each frequency band, providing a more perceptually relevant representation of the speech signal [6].

4.3.6 Discrete Cosine Transform (DCT)

DCT is applied to convert the log Mel spectrum into the time domain, producing the Mel Frequency Cepstral Coefficients (MFCCs) [3].

V. FEATURE CLASSIFICATION

In speaker recognition system another important part is classifications[12]. In classification stage the patterns are classified into different classes. There are many classifiers are used such as DWT, GMM, SVM, VQ, etc. From that selection of classifier is an important task. But there is no fix criteria for the selection of classifier. Many pattern classifiers are explored for developing speech systems like, emotion classification, speech recognition, speaker verification, speaker recognition.

5.1 Dynamic Time Warping (DTW)

The Dynamic Time Warping algorithm calculates the distance between two sequences that may vary in time or speed. Non-linear time warping is applied to align the timing differences between the test utterance and the reference template. Once aligned, the time-normalized distance is computed between the patterns. The authentic speaker is identified based on the minimum time-normalized distance. This technique is advantageous for handling variable-length input features and requires less storage space [2].

5.2 Vector Quantization (VQ)

Vector Quantization is a traditional quantization technique from signal processing that models probability density functions by distributing prototype vectors. Using VQ, the extracted speech features of a speaker are quantized into a set of centroids, which together form the speaker's codebook [1]. VQ is useful for data compression and requires minimal storage. It is computationally efficient and can be easily adapted for real-time applications [2].

5.3 Gaussian Mixture Model (GMM)

The Gaussian Mixture Model is a density estimation technique that belongs to unsupervised learning algorithms. GMM requires less training and test data compared to other methods. The Expectation-Maximization (EM) algorithm is used to estimate the GMM parameters from the training data. A sequence of features extracted from the input signal is processed, and the log-likelihood of the sequence is computed to assess the distance from the model. GMM is effective because it requires a smaller amount of data to train the classifier, resulting in lower memory requirements [2].

5.4 Support Vector Machine (SVM)

Support Vector Machines (SVM) are simple yet effective supervised learning algorithms commonly used for speech or speaker recognition. While SVM excels in binary classification tasks, it may not perform as well in speaker recognition due to its fixed-length vector constraints, which limit its adaptability to diverse speaker characteristics [2].

VI. CONCLUSION

This review paper provides an overview of speaker recognition techniques that are widely used in various speech processing applications, particularly for security and authentication purposes. It highlights the most commonly employed feature extraction techniques, with a focus on MFCC, which is the most widely used method. Additionally, different feature classification techniques for speaker recognition are discussed, shedding light on their strengths and weaknesses.

REFERENCES

[1] V. Tiwari, DzMFCC and its applications in speaker recognitiondz, IEEE International Jouranal on Emerging Technologies, Volume-1, Issue-7, May 2013, pp 33-37.

[2] K. Kau and N. Jain, DzFeature Extraction and Classification for Automatic Speaker Recognition System – A Reviewdz, International Journal of Advanced Research in Computrt Science and Software Engineering, Volume 5, Issue 1, January 2015, pp. 1-6.

[3] K. Dhameliya, DzFeature Extraction And Classification Techniques for Speaker Recognition: A Reviewdz, IEEE International Conference on Electrical, Electronics, Signal, Communication and Optimization, January 2015, pp. 1-4.

[4] S. Nakagawa, L. Wang and S. Ohtsuka, DzSpeaker Identification and Verification by Combining MFCC and Phase Information,dz IEEE Transaction on Audio, Speech and Language Processing, Vol. 20, No. 4, May 2012, pp. 1085-1095.

[5] M. AdaMmsk, B. VonSolms, Dz An Open Speaker Recognition Enabled Identification And Authentication Systemdz, IST-Africa 2014 Conference Proceedings 2014,pp 1-8.

[6] Lindasalwa M. , M. Begam and I. Elamvazuthi, DzVoice Recognition Algorithm using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniquesdz, Journal Of Computing, Volume 2, Issue 3, March 2010, pp 138-143.

[7] E. Chandra, K. Manikandan, M. Kalaivani, DzA Study on Speaker Recognition System and Pattern Classification Techniquesdz, International Journal Of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, Vol 2, Issue 2, February 2014, pp. 963-967.

[8] P. Chaudhary and M. Vagadia, DzA Review Article on Speaker Recognition with Feature Extractiondz, International Journal of Emerging Technology and Advanced Engineering, Volume 5, Issue 2, February 2015,pp. 94-97.

[9] R. Bharti and P. Bansal, DzReal Time Speaker Recognition System using MFCC and Vector Quantization Techniquedz, International Journal of Computer Applications, Volume 117 No. 1, May 2015, pp, 25-31.

[10] S. Suuny, D. Peter, K. Jacob, DzPerformance Of Different Classifiers In Speech Recognitiondz, International Journal of Research in Engineering and Technology, Volume: 02 Issue: 04 Apr-2013, pp. 590-597.

[11] S. Madikeri and H. Murthy, DzMel Filter Bank Energy Based Slope Feature and Its Application to Speaker Recognition,dz IEEE National Conference on communication (NCC), Bangalore, January 2011, pp. 1-4.

[12] S. K. Singh. Dz Features And Techniques For Speaker Recognitiondz, M. Tech. Credit Seminar Report, Electronic Systems Group, EE Dept, IIT Bombay submitted Nov 03.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com