ETHNIC RECOGNITION SYSTEM FOR MALAY LANGUAGE SPEAKERS USING GAMMATONE FREQUENCY CEPSTRAL COEFFICIENTS PITCH (GFCCP) AND PATTERN CLASSIFICATION

RAFIZAH BINTI MOHD HANIFA

A thesis submitted in
fulfillment of the requirement for the award of the
Doctor of Philosophy in Electrical Engineering

Faculty of Electrical and Electronic Engineering
Universiti Tun Hussein Onn Malaysia

MARCH 2022

To my beloved mother who taught me to trust in Allah and believe in hard work. To my husband and children who have always stood by me and understand my difficulties in completing this thesis.

ACKNOWLEDGEMENT

In the Name of Allah, the Most Merciful, the Most Compassionate, and all praise be to Allah, the Lord of the worlds; prayers and peace be upon Muhammad His servant and messenger. First and foremost, I must acknowledge my limitless thanks to Allah, the Ever-Magnificent and the Ever-Thankful, for His help and blessings. This work would never be accomplished without His guidance.

I would like to express my deep and sincere gratitude to my research supervisor, Ts. Dr Khalid bin Isa for allowing me to research under his supervision and for providing invaluable guidance. It was a great privilege and honour to be under his direction. I would also like to thank him for his friendship, empathy and great sense of humour.

I am extending my thanks to Universiti Tun Hussein Onn Malaysia for sponsoring my studies and providing moral support during my research work.

I am incredibly grateful to my mother, Hamidah, for her love, prayers, care, and sacrifices to educate and prepare me for my future. I am very much thankful to my husband, Shamsul, daughter, Amani, and sons, Irfan and Syahmie, for their love, understanding, prayers and continuous support that enabled me to complete this research work. I also thank my sisters, brothers, sisters-in-law and brothers-in-law for their support and valuable prayers.

Finally, my thanks go to those who have supported me to complete the research work directly or indirectly.

ABSTRACT

Malaysia is a multi-racial country consisting of many ethnic groups such as the Malay, Chinese, Indian, and Bumiputera, also known as a multilingual society. The Malay language is a non-tonal language, which does not need lexical stress. The study on recognizing the speaker's ethnicity is important as it has many potential and useful applications such as improving the interaction between robots and humans, audio forensic, telephone banking, and electronic commerce. Feature extraction, voice text-independent, and variability coverage are issues related to speaker recognition systems. The research focused on establishing a novel method, Gammatone Frequency Cepstral Coefficients and pitch (GFFCP) coupled with the K-Nearest Neighbours (KNN) and the voice text-independent system were used to identify the speaker's ethnicity. The speech corpus consisted of a collection of readings of Malay texts by both genders with ages ranging from 10 to 48 years old and classified into three ethnic groups: Malay, Chinese, and Indian. GFCC and Mel Frequency Cepstral Coefficients (MFCC) were used to represent the human auditory system. Pitch was added to MFCC and GFCC, as it contributes to the differences in the human voice and is difficult to imitate. The use of Naïve Bayes, Support Vector Machine (SVM), and KNN as classifiers was to quantify the pattern classification performance. The dataset used the hold-out validation methods (80% training, 20% testing) to split the data for training and testing. The system's performance was assessed based on the validation and prediction accuracy. The results revealed that the GFCCP obtained the highest validation and prediction accuracy from the KNN classifier. The validation accuracy was 100%, 99.6%, and 99.2% for 12, 24, and 34 speakers, respectively, while the prediction accuracy was 89.98%, 73.56%, and 72.36% for 12, 24, and 34 speakers, respectively. An important finding in the study is that the combination of the pitch with MFCC and GFCC provided better accuracy, with the latter performing better than the former, compared with those of MFCC and GFCC alone under noisy conditions.

ABSTRAK

Malaysia merupakan negara berbilang kaum yang terdiri daripada pelbagai etnik seperti Melayu, Cina, India, dan Bumiputera, dan dikenali sebagai masyarakat berbilang bahasa. Bahasa Melayu merupakan bahasa *non-tonal*, yang tidak memerlukan tekanan leksikal. Kajian pengecaman etnik penutur penting kerana berpotensi dan berguna dalam aplikasi untuk meningkatkan interaksi antara robot dan manusia, forensik audio, perbankan telefon, dan perdagangan elektronik. Pengekstrakan ciri, bebas teks suara dan liputan kebolehubahan antara isu yang berkaitan dengan sistem pengecaman penutur. Penyelidikan ini menumpukan kepada mewujudkan kaedah baru, di mana *Gammatone Frequency Cepstral Coefficients* dan nada (GFFCP) ditambah dengan *K-Nearest Neighbours* (KNN) menggunakan sistem bebas teks suara untuk mengenal pasti etnik penutur. Korpus pertuturan terdiri daripada koleksi bacaan teks Melayu oleh kedua-dua jantina dengan umur antara 10 hingga 48 tahun dan diklasifikasikan kepada tiga kumpulan etnik: Melayu, Cina, dan India. GFCC dan *Mel Frequency Cepstral Coefficients* (MFCC) digunakan kerana mewakili sistem pendengaran manusia. Nada ditambah kepada MFCC dan GFCC, kerana ia dapat membezakan suara manusia dan sukar ditiru. Penggunaan *Naïve Bayes*, Mesin Vektor Sokongan (SVM), dan KNN sebagai pengelas bertujuan mengukur prestasi pengelasan corak. Set data menggunakan kaedah *hold-out* (80% latihan, 20% ujian) untuk memisahkan data latihan dan ujian. Prestasi dinilai berdasarkan ketepatan pengesahan dan ramalan. Keputusan menunjukkan GFCCP memperoleh ketepatan pengesahan dan ramalan tertinggi daripada pengelas KNN. Ketepatan pengesahan adalah 100%, 99.6%, dan 99.2% untuk 12, 24, dan 34 penutur, masing-masing, manakala ketepatan ramalan ialah 89.98%, 73.56% dan 72.36% untuk 12, 24, dan 34 penutur, masing-masing. Penemuan penting kajian ialah gabungan GFCC dan MFCC dengan nada memberi ketepatan lebih baik, berbanding MFCC dan GFCC sahaja dalam situasi hingar.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABI	-	Accents of British Isles
ADAM	-	Advanced Development Autonomous Machine
AI	-	Artificial Intelligence
ANN	-	Artificial Neural Network
ASR	-	Automatic Speaker Recognition
AUC	-	Area Under the Curve
CER	-	Character Error Rate
CLSP	-	Centre for Language and Speech Processing
CNN	-	Convolutional Neural Network
DA-DNN7L	-	Data Augmentation Deep Neural Network 7 Layers
DBN	-	Deep Belief Network
DCT	-	Discrete Cosine Transform
DFT	-	Discrete Fourier Transform
DL	-	Deep Learning
DNA	-	Deoxyribonucleic Acid
DNA	-	Deep Neural Architecture
DNN	-	Deep Neural Network
DT-CWPT	-	Dual-Tree Complex Wavelet Packet Transform
DTW	-	Dynamic Time Warping
DWPT	-	Discrete Wavelet Packet Transform
ECG	-	Electrocardiogram
EEG	-	Electroencephalogram
EER	-	Equal Error Rate
ELM	-	Extreme Learning Machine
ELSDSR	-	English Language Speech Database for Speaker Recognition

EM	-	Expectation Maximization
ERB	-	Equivalent Rectangular Bandwidth
ERICA	-	ERATO Intelligent Conversational Android
FAR	-	False Acceptance Rate
FCM	-	Fuzzy C-Means
FFT	-	Fast Fourier Transform
FIR	-	Finite Impulse Transform
FN	-	False Negative
FNR	-	False Negative Rate
FP	-	False Positive
FPR	-	False Positive Rate
FRR	-	False Rejection Rate
FVQ	-	Fuzzy Vector Quantization
FVQ2	-	Fuzzy Vector Quantization2
GFCC	-	Gammatone Frequency Cepstral Coefficient
GFCCP	-	Gammatone Frequency Cepstral Coefficient Pitch
GMM	-	Gaussian Mixture Model
GMM-JFA	-	Gaussian Mixture Model-Joint Factor Analysis
GMM-UBM	-	Gaussian Mixture Model-Universal Background Model
HASR	-	Human Assisted Speaker Recognition
HMM	-	Hidden Markov Model
HRI	-	Human-Robot Interaction
IoT	-	Internet of Things
KNN	-	K-Nearest Neighbour
LID	-	Language Identification
LPC	-	Linear Prediction Coding
LPCC	-	Linear Prediction Cepstral Coefficient
MFCC	-	Mel-Frequency Cepstral Coefficient
MFCCP	-	Mel-Frequency Cepstral Coefficient Pitch
MGFCC	-	Modified GFCC
ML	-	Machine Learning
MLAN	-	Multi-level Adaptive Network
MLP	-	Multi-Layer Perception
MNN	-	Modular Neural Network

NB	-	Naïve Bayes
NICO	-	Neuro-Inspired Companion
NIST	-	National Institute of Standards and Technology
NIST 2003	-	National Institute of Standards and Technology 2003
PD	-	Partial Discharge
RCC	-	Real Cepstral Coefficient
RNN	-	Recurrent Neural Network
ROC	-	Receiver Operating Characteristics
SNR	-	Signal to Noise Ratio
SRE	-	Speaker Recognition Evaluation
STE	-	Short-Term Energy
STFT	-	Short-Time Fourier Transform
SVM	-	Support Vector Machine
TIMIT	-	Texas Instruments and Massachusetts Institute of Technology
TN	-	True Negative
TP	-	True Positive
TPR	-	True Positive Rate
VQ	-	Vector Quantization
WER	-	Word Error Rate
WPT	-	Wavelet Packet Transform
ZCR	-	Zero-Crossing Rate

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

Biometrics is widely used to identify and authenticate individuals trustworthily and promptly through unique biological characteristics. As shown in Figure 1.1, biometrics can be classified into physiological and behavioural categories (Porta et al., 2021; Rousan and Intrigila, 2020).

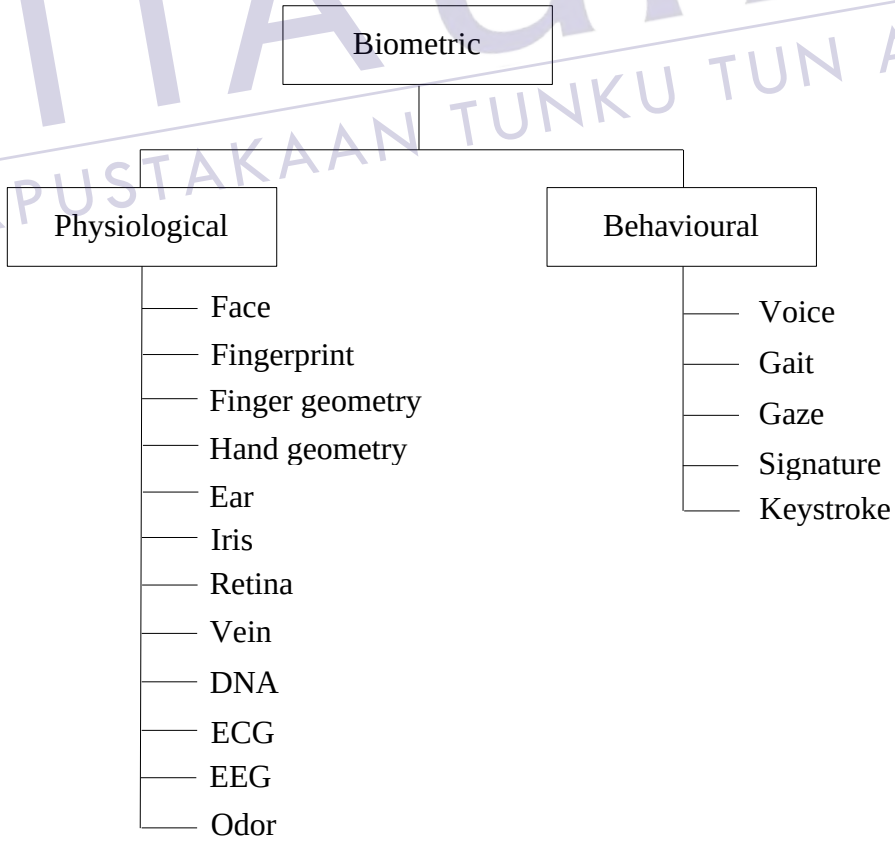

Figure 1.1: Types of biometrics: physiological and behavioural

The former refers to features identified through the five senses, i.e., sight, sound, smell, taste, and touch. For example, face, fingerprint, iris, retina, vein, ECG, odour, etc. The latter is usually based on how people conduct themselves, including voice, gait, gaze, signature, and keystroke (Rousan and Intrigila, 2020).

Biometric technology has various characteristics, by which we can distinguish their applications. Table 1.1 compares the most used biometric types based on the characteristics of biometric technology such as distinctiveness, complexity, universality, quantifiability, performance, comparison, collect capacity, acceptance, cost, and use.

Table 1.1: A comparison of biometric types based on the characteristics of biometric (Rousan and Intrigila, 2020)

Biometric Identifier	Distinctiveness	Complexity	Universality	Quantifiability	Performance	Comparison	Collect Capacity	Acceptance	Cost	Use
Fingerprint	M	L	H	H	M	H	H	H	M	H
Iris	H	L	H	H	H	H	H	H	H	M
Facial	M	L	H	H	M	M	H	H	M	M
Palm	M	H	H	H	M	M	L	L	H	M
Ear	M	H	H	H	L	L	L	L	H	L
Footprint	M	H	M	M	L	L	L	L	H	L
Finger vein	H	H	H	L	H	H	L	L	H	L
Voice	M	H	H	M	M	M	L	L	H	L
Signature	L	H	H	H	L	L	M	H	L	L
Keystroke dynamics	L	M	M	L	L	L	L	L	H	L

H = High; M = Medium; L = Low

Based on the information in the table, it can be deduced that voice is one of the useful technologies. Furthermore, a study by Sharma (2019) asserted that voice is a useful biometric because it provides comparable and much higher levels of security. In addition, the study by Zheng and Li (2017) stated that voice could be used to differentiate people because each person's voice has some unique characteristics. Before going any further, it is vital first to understand the essential characteristics of the voice.

In general, any sound produced by humans to communicate meanings, ideas, opinions, etc., is called the voice. In a more specific term, voice is any sound produced by vocal fold vibration, which occurs when air is under pressure from the lungs (Zhaoyan, 2016). Voice is the most natural communication tool used by humans. It conveys the speaker's traits, such as ethnicity, age, gender, and feelings. Lungs, larynx, pharynx, nose, and various parts of the mouth are all involved in producing voice

(Holmes and Holmes, 2002), as shown in Figure 1.2. A voice's features are dependent on its pace or speed, volume, pitch level, and quality, while articulation rate and speech pauses rely on the speaker's speaking style (Sujiya and Chandra, 2017).

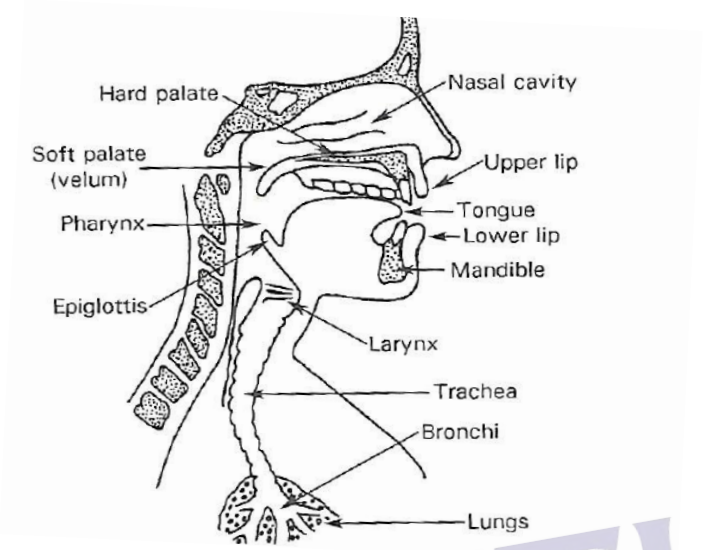

Figure 1.2: Diagrammatic cross-section of a human head showing vocal organs (Holmes and Holmes, 2002)

In speech processing, speaker and speech recognition are the two applications commonly used by researchers to analyse uttered speech (Sharma, 2019). Before delving further into the concept of speaker recognition, it is vital to understand the difference between speaker recognition and speech recognition. Although the terms 'speaker recognition' and 'speech recognition' have often been used interchangeably, they are different. Speech recognition is concerned with the spoken words, while speaker or voice recognition aims to recognise/identify the speaker rather than the words.

Speech recognition is helpful for people with various disabilities, such as those with physical disabilities who find typing the words difficult, painful, or impossible, and those who have difficulties recognising and spelling words, such as people with dyslexia. Since speech recognition deals with converting audio into text, its effectiveness depends heavily on the language and the text corpus (Sharma, 2019).

On the other hand, speaker recognition is to identify the person who is speaking. Speaker recognition scans the features of the speech uttered by an individual, which is distinctive due to their physiology and behavioural patterns. Pitch, speaking

style, and accent are some features that contribute to the differences. Speaker recognition technology has been used in various applications, such as biometrics, security, and even human-computer interaction. Table 1.2 summarises the differences between speaker recognition and speech recognition in terms of several features: recognition, purpose, focus, and application.

Table 1.2: Speaker recognition vs speech recognition

Features	Speaker Recognition	Speech Recognition
Recognition	Recognises who is speaking by measuring voice pattern, speaking style, and other verbal traits.	Recognises what is being said and converts them into text.
Purpose	To identify the speaker.	To identify and digitally record what the speaker is saying.
Focus	Biometric aspects of the speaker, such as pitch, intensity, etc., to recognise them.	Convert the vocabulary words of what is being said by the speakers into digital texts.
Application	Voice biometrics.	Speech to text.

Malaysia is a multi-racial country consisting of many ethnic groups such as the Malay, Chinese, Indian, and Bumiputera, which can further be classified as Iban, Kadazan, Melanau, Murut, Bidayuh, and Bajau (Nagaraj et al., 2009). Malaysia is also a multilingual society with hundreds of languages that more than a million native speakers speak (Lim, Huspi, and Ibrahim, 2021). The speech sound is concerned with phonetics, whereas phonology involves language functions. Malay is the national language, while English is the second language in Malaysia. The various ethnic groups speak both languages in Malaysia, but they might pronounce the same word slightly differently without affecting the meaning. Accents in a particular language are common in speech, especially when the language is spoken by non-native speakers (Juan, Besacier, and Tan, 2012).

Since Malay and English are the two important languages in Malaysia that began from British colonization, thus the comparison between these two languages is made in terms of vocals and diphthongs, place, and manner of articulations. There are six vocals, 27 consonants, and three diphthongs in the Malay sound system, whereas there are 12 vocals, 24 consonants, and eight diphthongs in the English sound system

(Alam, Zilany, and Davies-Venn, 2017). According to Kristin Denham and Anne Lobeck, there are seven important places of articulation in English, i.e., bilabial, labiodental, dental, alveolar, palatal, velar, and glottal. Whereas Malay phonology has labio-velar and no labiodentals and dental sounds (Azmi et al., 2016). As for the manner of articulation, Malay and English phonologies have six manners with voiced and unvoiced pronunciation. In Malay, they are plosive or affricate, fricative, nasal, trill, approximant, and lateral, while in English, they are stop, fricative, affricate, nasal, approximant, and glide.

1.2 Research Motivation

Humans have long dreamed of creating robots that can socially interact just like humans interact with each other. Applications based on social robots, which are a kind of humanoid robots, have recently emerged as a platform with huge potential in the field of human-robot interaction (HRI). Sophia, Jia Jia, ERICA, Nadine, Pepper, and NICO are some examples of humanoid robots that have been enhanced with human-like traits to improve the communication between robots and humans. If Nadine, a sitting robot designed as a companion for the elderly or children with special needs (Indramalar, 2016), Pepper is another personal humanoid robot that is used in Japan by pre-school children to help them study English at home and at retail stores to greet customers and provide information about products and services (Tanaka, Isshiki, and Takahashi, 2015). Unfortunately, those mentioned social humanoid robots can only converse in English despite being developed by researchers from China and Japan. Since each language reflects the culture of the particular social group, a humanoid robot must be sensitive to the pitch and intonation of each language for it to interpret correctly and give an appropriate response when communicating with users. ADAM, the Malaysian humanoid robot, currently converses only in English. It would be great if ADAM could interact with Malaysian people in the Malay language. It is the country's national language and a common language spoken by various ethnic groups. The Malay language is also commonly spoken in the region, such as in Indonesia, Singapore, Brunei, and South Thailand.

The pitch period refers to the interval of periodic motion caused by vocal cord vibration when an individual is uttering. Thus, it represents the vocal cords' speed

vibrating (Zhang and Yao, 2020). Vocal folds vibrate when the air is under pressure from the lungs. The tension in the vocal muscles controls pitch. The faster the vibration, the higher the pitch. The pitch period is the inverse of the vocal cord frequency *R* and thus, becomes an important parameter for speech signal analysis. The range of frequencies for the normal speaking human voice is 85 – 180 Hz for males and 165 – 255 Hz for females (Salleh et al., 2018). Thus, men have denser and longer vocal folds. Mustafa, Don, and Knowles (2013) highlighted that the Malay language is a non-tonal language that does not need lexical stress. Lexical stress, also known as word stress, is the stress placed on a given syllable in a word.

Language identification (LID) is an interesting field to be studied as it identifies a particular natural language from the given set of speech corpus that consists of a group of languages (Roy and Das, 2021). Besides, LID is an interactive process between humans and the system where users can directly interact with the system, identify the language spoken by the user, and respond in the recognized language. The application of LID can be applied in forensics. It can be done if a speech sample is recorded during the crime, and the suspect's voice can be compared for voice sample matching. The result can prove the criminal's identity and discharge the innocent during a court case.

1.3 Problem Statement

Speaker recognition emphasised raw audio and related data to identify the uniqueness in the way people speak to recognise the speaker's identity. The following drawbacks of the current speaker recognition system have been identified as having potential improvements.

The main challenge in the speaker recognition system is the extraction of discriminative features from speech signals that can improve performance from the classification algorithm (Jahangir et al., 2020). The terms 'feature extraction' and 'feature selection' have often been used interchangeably. The main idea behind the feature extraction is to compress the data to maintain most of the relevant information to improve the predictive performance of the models. In contrast, feature selection involves selecting a subset of the original features to simplify the model complexity (Olukoya and Musiliu, 2020). In other words, feature selection keeps a subset of the

original features while feature extraction creates brand new ones. Mel Frequency Cepstral Coefficient (MFCC) is the most used cepstrum feature in speaker recognition and is mainly designed using the knowledge of the human auditory system. One primary problem with MFCC is that it involves a plethora of information such as phone content, channels, noises, etc., thus, making it difficult to be used for speaker recognition (Tazi, 2016; Li et al., 2015). Furthermore, for the Mel scale uniform distribution, frequency division is based on the centre frequency and does not conform to the concept of critical bandwidth characteristics of hearing completely (Zhang and Ni, 2017). Due to the dynamic characteristics of MFCC, its characteristics are relatively stable, but it is easy to imitate (Zhang and Yao, 2020).

Although the early successful speaker recognition systems were all voice text-dependent, the voice text-independent has become a trend nowadays (Shaver and Acken, 2016). Speaker recognition systems can be categorised based on speech modality, i.e., voice text-dependent and voice text-independent. In a voice text-dependent system, the speaker utters the same text during the training and testing phases. As the spoken phrase is known beforehand, a voice text-dependent system is generally more robust and can achieve better performance. Unfortunately, this approach results in problems related to spoofing attacks. Once the imposter steals the information, the system will be easily broken up (Zheng and Li, 2017). One way to rectify the problems is to use a voice text-independent system. There are no constraints or restrictions in the spoken text during the training and testing phases in a voice text-independent system. The text could be user-selected phrases or conversational speech, thus leading to a more flexible system suited for identification. On top of that, in real life, a voice text-independent system is more commercially attractive and flexible (Sharma, 2019) than a voice text-dependent system because it is not convenient for the speaker to utter the same text a few times.

Another issue in speaker recognition is whether everyone in the sample corresponds to a different and identifiable feature's distribution or a wide overlap between speakers. The acoustic parameters extracted should exhibit a large between speaker variability and a low within the speaker. Several sources contribute to the variation, such as type and quality of recordings, the education level, gender, accents, etc. Furthermore, the individual feature distribution's estimation must be considered and often based on a small sample. Indeed, the ideal validation protocol would include

a database of vocal signals, represent the specific language, and have enough information (Zheng and Li, 2017).

1.4 Research Questions

The specific research questions identified to address the earlier problem statement are presented below.

RQ 1: What other feature extraction technique can be used to represent the human auditory system?

RQ 2: What parameter contains the information of the voice frequency that can prevent imitation?

RQ 3: How will speakers from different ethnicities' utterances be compiled?

RQ 4: What are the contents of the speech corpus?

RQ 5: How will the classification algorithms be used?

1.5 Research Objectives

This research work embarked on the following objectives:

(a) To develop a dataset based on the gammatone frequency cepstral coefficients pitch (GFCCP),

(b) To design a structured framework for a speaker ethnicity recognition system based on the Malay language, and

(c) To evaluate the performance of the speaker ethnicity recognition system based on the K-Nearest Neighbours (KNN) model and GFCCP features.

1.6 Scope of Study

This research focused on improving the speaker ethnicity recognition system, accurately classifying the speakers' ethnicity. The scope of the study is as follows:

(a) The respondents read texts (voice text-independent) from a local news website in standard Malay language. The speech corpus consisted of males and females aged 10 to 48 and classified into three ethnic groups: Malay, Chinese, and Indian. An open-set system was used in this research, as the system had no limits on the number of trained speakers, and the test speeches may comprise other than the trained speeches.

(b) This research adopted two different cepstral features, i.e., Mel Frequency Cepstral Coefficients (MFCC) and Gammatone Frequency Cepstral Coefficients (GFCC). Pitch is an additional feature concatenated to MFCC (MFCCP) and GFCC (GFCCP). It contains the information of the voice frequency structure that reflects the vocal cords' characteristics; thus, it is not easy to imitate. Besides, the pitch contributes to the differences in the human voice.

(c) The use of Naïve Bayes (NB), Support Vector Machine (SVM), and K-Nearest Neighbours (KNN) as classifiers was to quantify the pattern classification performance. The train-test split procedure is used to estimate the performance of machine learning (ML) algorithms to make predictions on data not used to train the model. The hold-out validation split the data, 80% for training and 20% for testing.

(d) The performance of the system was assessed based on validation and prediction accuracy.

A detailed explanation of how the research was conducted, including the activities involved, is discussed in Chapter 3. Figure 1.3 illustrates the scope covered in this research.

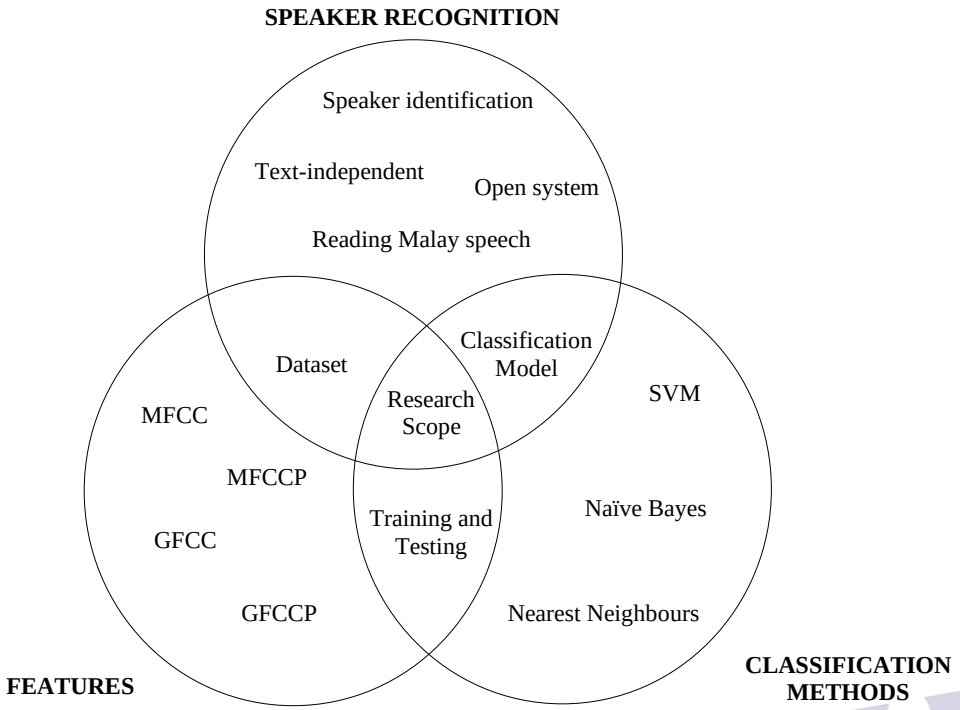

Figure 1.3: Illustration of the research scope

1.7 Significance of Study

The findings from this work, such as the convenient feature extraction parameters, methods used for recognising the ethnicity of the speakers, and the algorithm and system design, contribute to the knowledge in speech processing. Besides, it can be used to improve the humanoid robots' capabilities in interacting with humans and help in improving decision making on the innocent or guilty person for forensic cases based on the voice sampling matching.

1.8 Outline of the Thesis

In this **Introduction chapter**, the research background is presented. The motivation of the research is also briefly explained. The problem statement is discussed, and the specific research questions that reflect the problem statement are outlined. The primary objectives of the research are listed. The scope and significance of the research are also highlighted in this chapter.

The **Literature Review chapter** provides a basic of human speech. The historical overview of speaker recognition is described. The basic representation of a speaker recognition system consists of the speech signal, pre-processing, feature extraction, classification, assessment, and decision. Popular databases that are available and widely used for speaker recognition are highlighted. The commonly used features extracted from speech signals, such as MFCC, LPC, etc., are also described. Machine learning classifiers are also discussed in this chapter. The assessment of the trained classifiers is explained based on the confusion matrix, ROC and AUC, and EER. Previous research studies on multi-ethnic progress are summarised, highlighting the research gap.

The **Methodology chapter** explains the method adopted in this work. In organising this chapter, a framework was designed, showing the steps taken to accomplish the work. The research method is presented in algorithms to direct the flow of the research: speech corpus, pre-processing, feature extraction, pattern classification, assessment, and decision. The experimental setup described how the research method was implemented using the machine learning approach to identify the speaker's ethnicity based on the features extracted from the recorded speech. The prediction accuracy was based on the new speech that was not used during training. The model with the highest prediction accuracy was chosen as the best design in predicting the speaker's ethnicity.

The **Results and Analysis chapter** presents the results based on the proposed method. Twelve different datasets with different feature parameters and numbers of speakers are used to be trained using three different classifiers, i.e., Naïve Bayes, SVM, and KNN, for comparison purposes. The assessment of the models is made using the confusion matrix, ROC and AUC, and EER. The model with the highest validation accuracy is tested with a new set of data to determine its prediction accuracy. The analysis is based on validation and prediction accuracy before deciding the best model. The results obtained in this research are compared with another research.

The **Conclusion chapter** presents the overall work carried out in the research. Research contributions are highlighted, and the achievements of the research objectives identified in the research are revisited. In addition, some recommendations for future work are provided.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Speech signal processing technology has become a popular communication technology, as many applications today use speech to enhance everyday human life. Human speech reveals a lot of information, as the human voice forms a vital characteristic of an individual (Saste and Jagdale, 2017; Shaver and Acken, 2016; Jain and Sharma, 2013; Doddington, 1985). Accent, language, speech, emotion, gender, and the speaker's identity are some of the human voice's information (Zheng and Li, 2017; Muda, Begam, and Elamvazuthiet, 2010), as shown in Figure 2.1.

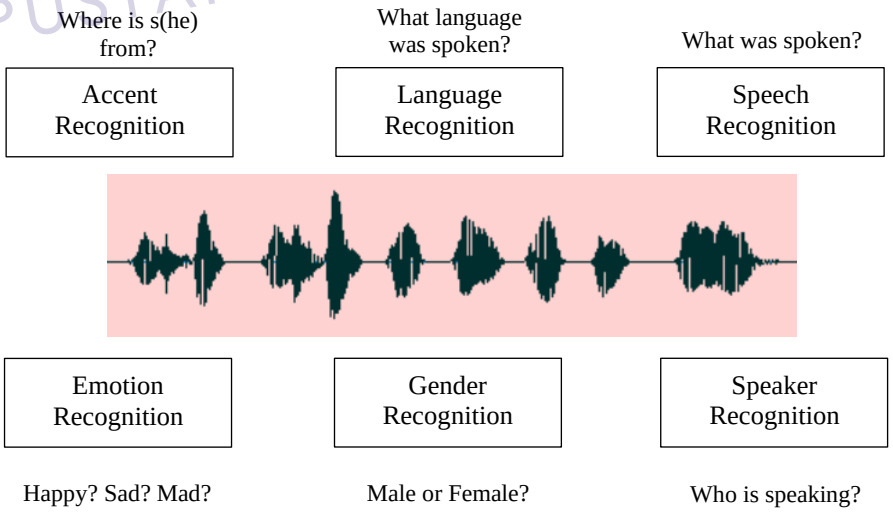

Figure 2.1: Some of the information contained in spoken language (Zheng and Li, 2017)

Automatic Speaker Recognition (ASR) is an essential tool for recognising people based on their voice (Zheng and Li, 2017; Singh, N., 2014; Sharma and Bansal, 2013). The field of speaker recognition has gained more attention lately. Although researchers have been working on speaker recognition in the last eight decades, advancements in technology, such as the Internet of Things (IoT), intelligent devices, voice assistants, and smart homes have made it popular (Sharma, 2019).

2.2 Historical Overview of Speaker Recognition

Advancements in various fields have increased the importance of speaker recognition systems, especially in identifying a person's identity. Research on speaker recognition first started in the 1930s. In March of 1932, the kidnapping and killing of Charles and Anne Lindbergh's baby boy led to the investigation into the speaker's speech signal. During the suspected kidnapper's trial, Charles Lindbergh claimed that the kidnapper's voice, Bruno Hauptmann, was the same as the voice he heard while waiting in a car nearby where the ransom was paid (Singh, Argawal, and Khan, 2018). Frances McGehee, who was inspired by the case, conducted the first academic research on the reliability of ear witnesses in 1937, which later became a topic of interest in forensic and psychology research (Singh et al., 2018). In 1946, scientists Potter, Kopp, and Green at Bell Laboratories developed a visual representation of speech called a spectrogram, displaying the frequency and intensity of a speech signal concerning time (Shaver and Acken, 2016).

In 1960, a Swedish professor named Gunner Fant developed a physiological model of human speech. Two years later, Lawrence Kersta, a physicist at Bell Laboratories, published an article entitled ‘Voiceprint Identification’, which later became one of the types of evidence used in US courts (Zheng and Li, 2017). Kersta's method was an aural-visual method, by which the spectrogram was inspected visually for pattern matching and scored by an interpreter. In 1963, Bogert, Healy, and Tukey published their study on a new method for detecting echo by taking the spectrogram of a log-magnitude spectrum in seismic signal (Shaver and Acken, 2016). In the same year, Sandra Pruzansky, who also worked at Bell Labs, was the first to research filter banks and look at the correlation between two digital spectrograms for a similarity measure. A year later, she worked with Max Mathews to improve this technique, which was subsequently developed by Li, Damman, and Chapmann using linear

discriminators (Furui, 2009). In 1964, Michael Noll, inspired by the echo-detecting cepstrum (Bogert, Healy, and Tukey), explored its use for the pitch detection of the human voice (Shaver and Acken, 2016). In 1965, Cooley and Tukey published their work on digital implementation for the Fourier Transform, which later became known as the Cooley-Tukey Fast Fourier Transform (FFT) (Singh et al., 2018). Alan Oppenheim and Ronald Schafer, who was inspired by a subsequent work of Michael Noll, introduced Real Cepstral Coefficients (RCC) in 1967 (Ganchev, 2011). James Luck initiated the cepstrum technology for speaker recognition two years later, which succeeded (Zheng and Li, 2017).

In the early 1970s, Leonard Baum and Lloyd Welch developed the Hidden Markov Model (HMM), which was later widely used in speaker recognition (SR) systems during the 1980s (Kouemou, 2011). George Doddington developed the first successful autonomous SR system in 1971. He used digital filter banks to conduct spectral analysis, a text-dependent system (Zheng and Li, 2017). Atal and Hanauer proposed the Linear Prediction Coefficients (LPC) in the same year. After three years, the former proposed the Linear Prediction Cepstral Coefficients (LPCC) to improve further the precision of cepstral coefficients (Ganchev, 2011). Research on the Support Vector Machine (SVM) started after Vladimir Vapnik developed the statistical learning theory in 1979 (Shaver and Acken, 2016). The original SVM algorithm was invented by Vladimir Vapnik and Alexey Chervonenkis in 1974. SVM is used to classify data. This classifier's advantage is that it uses an optimised non-linear decision boundary to minimise false reject and false accept error rates (Shaver and Acken, 2016).

In 1981, Sadaoki Furui proposed cepstral coefficients and their orthogonal polynomial coefficients as frame-based features to increase robustness against distortions by the telephone system (Furui, 2009). Score normalisation attempts in minimising error by removing speaker model score vectors away from the decision boundary began in 1988 when Li and Porter normalised the score distribution of the imposter model (Shaver and Acken, 2016). Since then, many variations of score normalisation have arisen, such as the T-norm and Z-norm. In 1992, Douglas Reynolds introduced the Gaussian Mixture Model (GMM). His work led to a new paradigm in speaker recognition due to its flexibility, high efficiency, and good robustness (Zheng and Li, 2017). In 2000, he developed the Gaussian Mixture Model-Universal Background Model (GMM-UBM), which had a high impact as the speaker recognition

technology moved from lab experiments to actual practical use. This model used a set of not authenticated people (Shaver and Acken, 2016). Realising the significant impact of the GMM thus, Dehak, Dumouchel, and Kenny proposed the Gaussian Mixture Model-Joint Factor Analysis (GMM-JFA) (Dehak, Dumouchel, and Kenny, 2007). They extracted the pitch and energy of continuous prosodic features to be modelled using GMM and compensate speaker and session variability effects using JFA.

The first Speaker Recognition Evaluation (SRE) was performed by the National Institute of Standards and Technology (NIST) in 1996 for text-independent systems (Shaver and Acken, 2016). Researchers from the Centre for Language and Speech Processing (CLSP) at Johns Hopkins University researched the SuperSID project in 2003, aimed at analysing, characterising, extracting, and applying high-level information to the speaker recognition task (Reynolds et al., 2003). In 2010, a Human Assisted Speaker Recognition test was included in NIST-SRE as an attempt to lower error rates by allowing humans to supplement the autonomous systems (Shaver and Acken, 2016). A year later, Ke Chen and Ahmad Salman proposed a deep neural architecture for learning speaker-specific characteristics from Mel Frequency Cepstral Coefficients (MFCC), an acoustic representation commonly used in both types of speaker recognition systems (Chen and Salman, 2011).

Table 2.1 summarises some of the major speaker recognition advances by the researchers mentioned above.

Table 2.1: Timeline of major speaker recognition advances

Year	Advancement in the field of the speaker recognition system
1937	**Frances McGhee**: First academic research into speaker recognition
1946	**Potter, Koop and Green**: Development of spectrogram at Bell Laboratories
1960	**Gunner Fant**: Development of Physiological model of speech
1962	**Lawrence Kersta**: Published an article on Voiceprint
1963	**Bogert, Healy and Tukey**: Published a study on echo detection in seismic signal
	Sandra Pruzansky: Initiate research on filter banks and correlation spectrograms
1964	**Michael Noll**: Cepstrum pitch determination
1965	**Cooley and Tukey**: Development of Fast Fourier Transform (FFT)
1967	**Oppenheim and Schafer**: Introduced Real Cepstral Coefficients (RCC)
1969	**James Luck**: Applied cepstrum technology to speaker recognition
1970	**Leonard Baum and Lloyd Welch**: Development of Hidden Markov Model

Table 2.1 (continued)

Year	Advancement in the field of the speaker recognition system
1971	**George Doddington**: Text-dependent system at Texas Instruments
	Atal and Hanauer: Proposed Linear Prediction Coefficients (LPC)
1974	**Binshu Atal**: Proposed Linear Prediction Cepstral Coefficients (LPCC)
1979	**Vapnik and Chervonenkis**: Invention of Support Vector Machines (SVM) algorithm
1981	**Sadaoki Furui**: Cepstrum based system
1988	**Li and Porter**: Proposed score distribution of the imposter model
1992	**Douglas Reynolds**: Gaussian Mixture Model (GMM) based system
1996	**National Institute of Standards and Technology (NIST):** Performed the first Speaker Recognition Evaluation
2000	**Douglas Reynolds**: Proposed Gaussian Mixture Model-Universal Background Model (GMM-UBM)
2003	**Centre for Language and Speech Processing (CLSP) Group**: Super SID project
2007	**Dehak, Dumouchel and Kenny**: Proposed Gaussian Mixture Model – Joint Factor Analysis (GMM-JFA)
2010	**National Institute of Standards and Technology (NIST):** Included a Human Assisted Speaker Recognition (HASR) test
2011	**Chen and Salman**: Proposed Deep Neural Architecture (DNA)

2.3 Speaker Recognition System

Humans interact with machines via devices that require physical movements, such as the mouse or keyboard. The emergence of speech technology has changed human and machine interaction through speech, making it more popular due to speed, ease of use, and more comfortable for some users. Speaker recognition is an Artificial Intelligence (AI) technology that lets the machine process, interpret, and respond to human language. Speaker recognition is not an easy task, as many factors create variances in the speech signals during the training and testing sessions, such as changes in people's voices due to time, health conditions, speaking rates, etc. (Suchita and Bindu, 2015).

The basic representation of the speaker recognition system consists of the speech signal, pre-processing, feature extraction, classification, assessment, and decision, as shown in Figure 2.2.

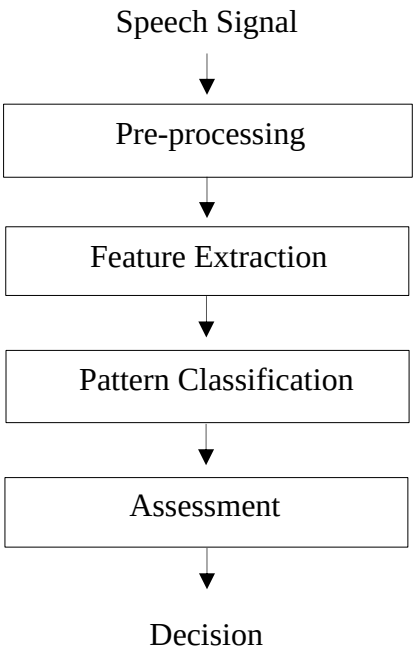

Figure 2.2: Process flow in speaker recognition

In the following subsections, the discussion on these processes is explained in more detail.

2.3.1 Speech Signal

Speech databases are needed to get adequate amounts of speech to train and test the speaker recognition system. The application of speaker recognition leads to a diversity of the structure and content of speaker recognition databases. Table 2.2 shows some popular available and widely used databases for speaker recognition.

Table 2.2: Popular databases used for speaker recognition
(Barai et al., 2017)

Database	Language	Recording Device (s)	Utterance Type				No. of Speakers	Creator	Distributor
			Sentence	Word	Digit	Spontaneous			
IITG-MV SR (Phase I, II III & IV)	ENG (IND) + 13 Regional Languages	Mobile (2), DVR, Tablet PC, Headset mic	YES	NO	NO	YES	100	Indian Institute of Technology, Guwathi	IITG
Russian Speech Database	RUS	mic	YES	NO	NO	NO	89	STC	ELRA

Table 2.2 (continued)

Database	Language	Recording Device (s)	Utterance Type				No. of Speakers	Creator	Distributor
			Sentence	Word	Digit	Spontaneous			
XM2VTS	ENG (GB)	mic, video	NO	NO	YES	NO	295	Univ. of Surrey	Univ. of Surrey
Brent	ENG (GB)	tel.	YES	YES	YES	NO	100	BT	BT
Millar	ENG (GB)	mic	NO	NO	YES	NO	63	BT	BT
SpeechDat (FDB+SDB)	ENG (GB)	tel.	YES	YES	YES	NO	5120	GPT Ltd.	ELRA
Polycost	ENG	tel.	YES	NO	YES	YES	134	COST250	ELRA
EUROM-1	DAN	mic	YES	NO	YES	NO	60	Tele Denmark, CPK	ELRA
TIMIT/ NTIMIT	ENG (US)	mic, tel.	YES	NO	NO	NO	630	MIT, TI, SRI	LDC
YOHO	ENG (US)	mic	NO	NO	YES	NO	138	ITT, Oklahoma State Univ.	LDC
Switchboard-1	ENG (US)	tel.	NO	NO	NO	YES	325	TI, NIST, LDC	LDC
Switchboard-2 (Phase I & II)	ENG (US)	tel.	NO	NO	NO	YES	657 + 679	LDC	LDC
KING-92	ENG (US)	mic, tel.	NO	NO	NO	YES	51	ITT	LDC
LLHDB	ENG (US)	mic	YES	NO	NO	YES	53	MIT-LL	LDC

The above table summarizes some available speech databases to support speaker recognition research and evaluation. As can be seen, most of the available databases are in the English language. Microphones and telephones are widely used in terms of recording devices compared to other devices such as tablets, videos, etc. Most of the utterances are recorded based on the sentence and digit. Apart from the abovementioned speech databases, other databases are also preferred by the researchers, such as the National Institute of Standards and Technology (NIST), NIST 2003, English Language Speech Database for Speaker Recognition (ELSDSR), VoxForge, and LibriSpeech. Although there is a plethora of available databases, some researchers developed their corpus based on their spoken languages such as Czech (Král, 2010), Marathi (Jawakar et al., 2013), Arabic (Tolba et al., 2015), Chinese (Li et al., 2015) and Malay (Abdullah et al., 2019; Salleh et al., 2018; Juan, Besacier, and Tan, 2012). As for the number of populations, most researchers, especially those using self-generated corpus, only used a small sample (Abdullah et al., 2019; Desai and Tahilramani, 2017; Soleymanpur and Marvi, 2017; Jawarkar et al., 2011) possibly to reduce training time.

In summary, the speech database plays an important role as, without it, there would be no research on speaker recognition. There is also a trend in developing own spoken language database as it reflects the country's identity. The careful selection or decision of a corpus drives the directions of research.

2.3.2 Pre-processing

Pre-processing is the first step in speech signal processing, and it involves converting an analog signal into a digital signal (Imam, Bansal, and Singh, 2017; Singh, Agrawal, and Khan, 2015). Interference due to noise often occurs during speech recording, causing the performance to degrade. The pre-processing stage's main objective is to modify the speech signal to be suitable for feature extraction analysis (Ibrahim, Odiketa and Ibiyemi, 2017; Suchitha and Bindu, 2015). Different methods can be adopted for noise-reduction algorithms, and the two most frequently used are spectral subtraction and adaptive noise cancellation (Ibrahim et al., 2017). However, Cutajar et al. (2013) highlighted that the function to be used during the pre-processing stage is dependent on the approach employed at the feature extraction stage. Some commonly used functions include noise removal, endpoint detection, pre-emphasis, framing and windowing, and normalisation (Singh et al., 2015; Suchitha and Bindu, 2015). The pitch was the only characteristic of a source in the region of voiced speech. Thus, to distinguish between silence and speech, the simplest method that can be applied is to analyse the short-term energy (STE) and zero-crossing rate (ZCR) for each frame, as explained in the following subsections.

2.3.2.1 Short-Term Energy (STE)

The amplitude of unvoiced segments is generally much lower than the amplitude of voiced segments. The short-time energy of the speech signal provides a convenient representation that reflects these amplitude variations. The STE can be calculated by using the following equation (Rabiner and Schafer, 2007; Schafer and Rabiner, 1975):

$$E_T = \sum_{m=-\infty}^{\infty} s^2\,(m) \tag{2.1}$$

where E_T is the total energy and $s(m)$ is the discrete time signal.

The STE of the voiced signal is always greater than that of unvoiced signals.

2.3.2.2 Zero-Crossing Rate (ZCR)

ZCR is defined as the number of times the zero axes is crossed per frame (Nandhini and Shenbagavalli, 2014). If zero crossings are more in a given signal, the signal contains high frequency and is termed unvoiced speech. In contrast, if zero crossing is less, the signal has low frequency and is termed voiced speech (Bachu et al., 2010; Schafer and Rabiner, 1975;). ZCR is defined as the weighted average of the number of times the speech signal changes sign within the time window, and it is given by the following equation (Rabiner and Schafer, 2007):

$$Z_n = \sum_{m=-\infty}^{\infty} \left| sgn\left(x[m] - sgn(x[m-1])\right) \right| w[n-m] \tag{2.2}$$

where

$$sgn(x[n]) = \begin{cases} 1, for\ x[n] \geq 0 \\ -1, otherwise \end{cases} \tag{2.3}$$

and

$$w[n] = \begin{cases} \frac{1}{2N}, for\ 0 \leq n \leq N-1 \\ 0, otherwise \end{cases} \tag{2.4}$$

In conclusion, pre-processing is a crucial and critical step, as improper pre-processing conducted on the recorded speech input will decrease the classification performance.

2.3.3 Feature Extraction

Feature extraction is a significant issue in voice text-independent speaker recognition systems (Sharma, 2019). It can be considered a process to extract the speaker's feature traits. The basic principle of feature extraction is to extract a sequence of features for each short-time frame of the input signal, assuming that such a small segment of

speech is sufficiently stationary to allow for better modelling (Reddy Gade and Sumathi, 2021; Saste and Jagdale, 2017). In other words, feature extraction is accomplished by changing the speech waveform to a parametric representation at a relatively minimized data rate for subsequent processing and analysis (Alim and Rashid, 2018). This phase is vital for the next step, as it affects the behaviour of the modelling process. The speaker signal is a dependent speech system. The speech signal is analysed to get less variability and identify more discriminative features by converting a speech signal to parametric values (Singh et al., 2018). The various techniques used for extracting speech features are in the form of coefficients, including Linear Prediction Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC), Mel Frequency Cepstral Coefficients (MFCC) and Gammatone Frequency Cepstral Coefficients (GFCC). The following subsections discuss each of these techniques in more detail.

2.3.3.1 Linear Prediction Coefficients (LPC)

LPC is based on a mathematical approximation of the vocal tract (throat, tongue, and lips) and tube diameter (Atal, 1974). This technique analyses the speech signal by estimating the formants where it removes the effects of formants from the speech signal and calculates the remaining buzz's intensity and frequency (Alim and Rashid, 2018). Removing the formants is known as inverse filtering, and the remaining signal is called the residue. Each speech signal sample conveyed as a linear combination of the previous samples is a linear predictor (Imam et al., 2017). Hence, the process is called linear prediction coefficients (LPC). LPC may decrease the bit rate significantly, and this reduction rate has a distinctive artificial sound, which causes a loss in the quality of the signal (Sharma and Bansal, 2013). The basic procedure to get an LPC coefficient is shown in Figure 2.3 (Alim and Rashid, 2018; Sanjaya, Anggraeni, and Santika, 2018; Rajasekhar and Hota, 2018; Amrutha et al., 2016).

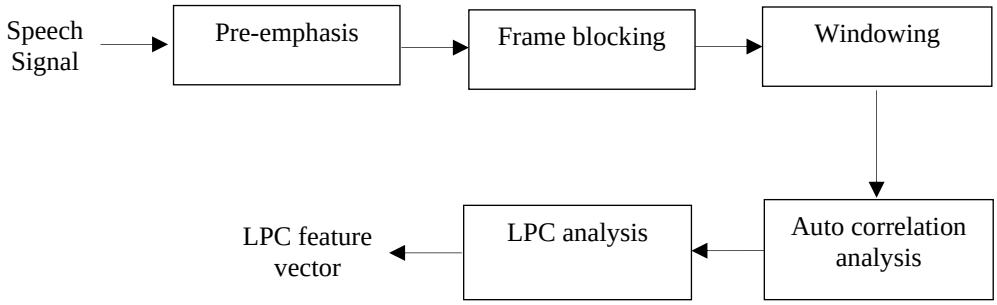

Figure 2.3: Block diagram of LPC extraction
(Amrutha et al., 2016)

Since this technique does not represent the vocal tract's characteristics from the glottal dynamics, it takes more time and computational cost to implement the speaker's model (Imam et al., 2017). Kaur and Jain (2015) pointed out that LPC's inconsistency with human hearing tends to provide detail to all the frequencies equally, which usually results in additional noise. Thus, LPC is only suitable for encoding speech at a low bit rate (Saste and Jagdale, 2017).

2.3.3.2 Linear Prediction Cepstral Coefficients (LPCC)

LPCC is often used in speaker recognition systems. LPCC is an improved LPC form, which considers the differences in the biological structure of the vocal tract in human beings (Atal, 1974). In estimating the common parameters of the speech signal, i.e., pitch period, speech frame energy and formant, LPCC has become one of the important features to consider. The aim is to display speech signals through finite numbers of signal measures. LPCC is derived through different translations into cepstral coefficients through LPC using autocorrelation (Gupta and Gupta, 2016). Since LPCC is an extension of LPC, the block diagram is the same as that for LPC but with an additional phase, which is the LPC parameter conversion (highlighted), as shown in Figure 2.4.

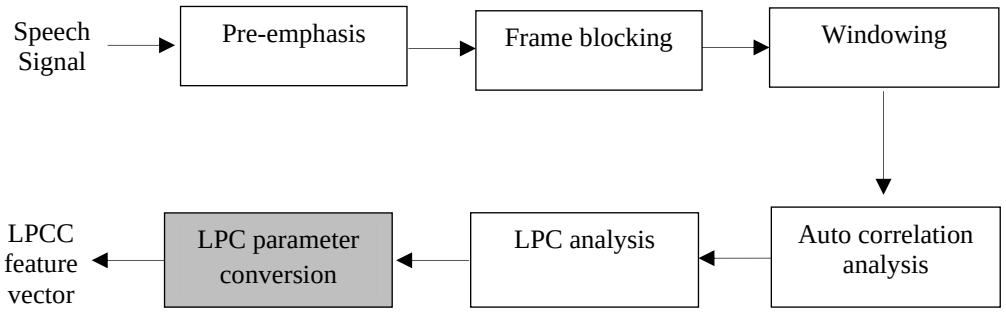

Figure 2.4: Block diagram of LPCC extraction
(Alim and Rashid, 2018)

LPCC parameters can effectively describe sound frames' energy and frequency spectrum. In other words, LPCC can include more information on the acoustic signal, but at the same time, it also increases the computational complexity.

2.3.3.3 Mel Frequency Cepstral Coefficients (MFCC)

MFCC is one of the most renowned voice feature extractions for speech signals (Li et al., 2020; Kaphungui and Kandali, 2019; Rajasekhar and Hota, 2018). MFCC is modelled to match the human auditory system (Kaphungui and Kandali, 2019; Imam et al., 2017). Figure 2.5 depicts the procedure for extracting the MFCC feature vector from speech (Saste and Jagdale, 2017).

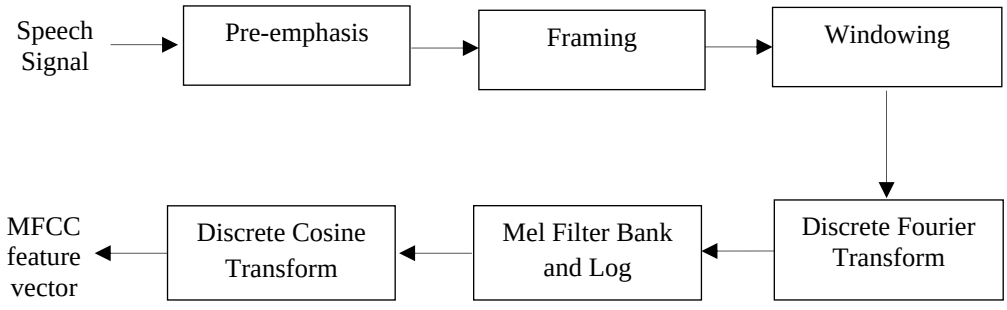

Figure 2.5: Block diagram of MFCC extraction
(Saste and Jagdale, 2017)

Pre-emphasis

The first step in MFCC feature extraction is to boost the energy to high frequencies by passing the signal through a finite impulse response (FIR) filter that aids in increasing the signal's energy at a higher frequency (Joshi and Cheeran, 2014; Gaurav et al., 2012). Equation (2.5) represents the FIR filter (Gaurav et al., 2012):

$$H(z) = 1 - \alpha z^{-1}, \ 0.9 \leq \alpha \leq 1.0 \tag{2.5}$$

Framing

Since speech is a non-stationary signal, the continuous speech signal is segmented into frames of N samples with the adjacent frames being separated by M $(M < N)$ during this step (Suchita and Bindu, 2015; Abdull Sukor, 2012; Muda et al., 2010). In other words, framing is the process of segmenting the speech signal into a small frame within the range of 20 to 40 ms (Jain and Sharma, 2013). Figure 2.6 illustrates the framing of speaker utterance.

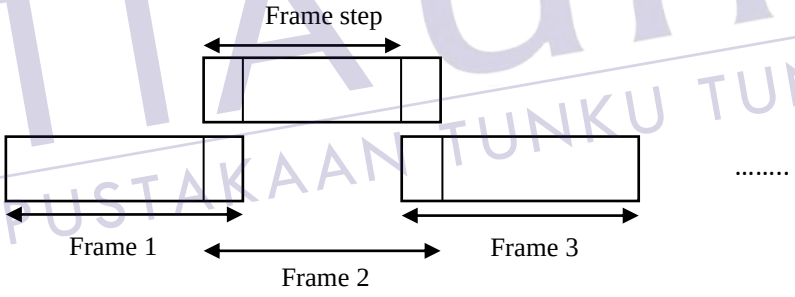

Figure 2.6: Framing of speaker utterance

Windowing

Windowing is the process of creating a window for each frame to minimise the discontinuity of the signal at the beginning and the end of each frame (Jain and Sharma, 2013; Ghadge et al., 2010; Rabiner and Juang, 1993). The most uncomplicated window is a rectangular window. However, the problem with such a window is that it can suddenly cut the signal at its boundaries. The equation for the rectangular window is as follows (Haggerty, 2008):

$$rectangular \quad w[n] = \begin{cases} 1 & 0 \leq n \leq L-1 \\ 0 & otherwise \end{cases} \tag{2.6}$$

REFERENCES

Abdull Sukor, A.S. (2012). "Speaker Identification System Using MFCC Procedure and Noise Reduction Method". (Master's Thesis). Retrieved from http://eprints.uthm.edu.my/id/eprint/2428/1/Abdul_Syafiq_Abdull_Sukor.pdf.

Abdullah, R., Muthusamy, H., Vijean, V., Abdullah, Z. & Che Kassim, F.N. (2019). "Real and Complex Wavelet Transform Approaches for Malaysia Speaker and Accent Recognition". *Pertanika Journal of Science and Technology*, 27(2), pp. 737-752.

Alam, M. S. M., Zilany, S. A. & Davies-Venn, E. (2017). "Effects of Speech-shaped Noise on Consonant Recognition in Malay". *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, 2017, pp. 22-25, doi: 10.1109/R10-HTC.2017.8288897.

Alim, S. A. & Rashid, N. K. A. (2018). "Some Commonly Used Speech Feature Extraction Algorithms, From Natural to Artificial Intelligence - Algorithms and Applications", Ricardo Lopez-Ruiz, IntechOpen, DOI: 10.5772/intechopen.80419. Available from: https://www.intechopen.com/books/from-natural-to-artificial-intelligence-algorithms-and-applications/some-commonly-used-speech-feature-extraction-algorithms.

Amrutha, R., Lalitha, K., Shivakumar, M. & Michahial, S. (2016). "Feature Extraction of Speech Signal using LPC". *International Journal of Advanced Research in Computer & Communication Engineering*, Vol. 5, Issue, 12, December 2016.

Anggraeni, D., Sanjaya, W.S.M., Nurasyidiek, M.Y.S. & Munawwaroh, M. (2017). "The Implementation of Speech Recognition using Mel-Frequency Cepstrum Coefficients (MFCC) and Support Vector Machine (SVM) Method Based on Phython to Control Robot Arm". *The 2nd Annual Applied Science and Engineering Conference (AASEC)*, pp.1-10.

Ashar, A., Bhatti, M. S. & Mushtaq, U. (2020). "Speaker Identification Using a Hybrid CNN-MFCC Approach". *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*, pp. 1-4, doi: 10.1109/ICETST49965.2020.9080730.

Atal, B.S. (1974). “Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification”. Journal of the Acoustical Society of America Volume 55, Issue 6, Pages 1304 – 1312, June 1974.

Auda, G. & Kamel, Mohamed S. & Raafat, Hazem. (1996). “Modular Neural Network Architectures for Classification”. 1279 - 1284 vol. 2. 10.1109/ICNN.1996.5490.

Azmi, M.N., Ching, L.T.P., Norbahyah, Haziq, M.N., Habibullah, M., Yasser, M.A. & Jayakumar, L. (2016). “The Comparison and Contrasts between English and Malay Languages”. English Review, 4(2), 209-218.

Babu, M. (2014). “Whether MFCC or GFCC is Better for Recognizing Emotion from Speech? A Study”. *International Journal of Research in Computer Applications and Robotics*, vol. 2, issue 6, pp. 14-17.

Babu, M., Arun Kumar, M.N. & Santosh, S.M. (2014). “Extracting MFCC and GFCC Features for Emotion Recognition from Audio Speech Signals.” *International Journal of Research Computer Applications and Robotics*, 2(8), pp 46-63.

Bachu, R.G., Kopparthi, S., Adapa, B. & Barkana, B.D. (2010). “Voiced/Unvoiced Decision for Speech Signals based on Zero Crossing Rate and Energy”. Advanced Techniques in Computing Sciences and Software Engineering, pp. 279-282.

Barai, B., Das, D., Das, N., Basu, S. & Nasipuri, M. (2017). “An ASR system using MFCC and VQ/GMM with Emphasis on Environmental Dependency”. *2017 IEEE Calcutta Conference (CALCON)*, 2017, pp. 362-366, doi: 10.1109/CALCON.2017.8280756.

Beigi, H. (2011). “Speaker Recognition”. 10.5772/17058.

Bjaili, H., Daqrouq, K. & Al-Hmouz, R. (2014). “Speaker Identification using Bayesian Algorithm”. *Trends in Applied Sciences Research, Academic Journals Inc*., pp. 472-479.

Bradley, A.P. (1997). “The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms”. *Pattern Recognition*, 30(7), pp. 1145-1159.

Chakroun, R. & Frikha, M. (2020). “Robust Text-independent Speaker recognition with Short Utterances using Gaussian Mixture Models”. *2020 International Wireless Communications and Mobile Computing (IWCMC)*, pp. 2204-2209, doi: 10.1109/IWCMC48107.2020.9148102.

Chen, K. & Salman, A. (2011). "Learning Speaker-Specific Characteristics with a Deep Neural Architecture. Neural Networks". IEEE Transactions on. 22. 1744 - 1756. 10.1109/TNN.2011.2167240.

Chethana, C. (2021). "Prediction of Heart Disease using Different KNN Classifier". *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 1186-1194, doi: 10.1109/ICICCS51141.2021.9432178.

Chowdhury, A. & Ross, A. (2020). "Fusing MFCC and LPC Features Using 1D Triplet CNN for Speaker Recognition in Severely Degraded Audio Signals". *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1616-1629, doi: 10.1109/TIFS.2019.2941773.

Cofiño, A.S. & Gutiérrez, J. M. (2001). Optimal Modular Feedfroward Neural Nets Based on Functional Network Architectures. Lecture Notes in Artificial Intelligence. 2083. 308-315. 10.1007/3-540-45720-8_35.

Cutajar, M., Gatt, E., Grech, I., Casha, O. & Micallef, J. (2013). "Comparative Study of Automatic Speech Recognition Techniques". *IET Signal Processing*, 7(1), pp. 25-46.

Dehak N., Humouchel, P. & Kenny, P. (2007). "Modeling Prosodic Features with Joint Factor Analysis for Speaker Verification". IEEE Transactions on Audio, Speech, And Language Processing. 2007 Sep; 15(7):2095–103 https:// doi.org/10.1109/TASL.2007.902758.

Desai, N. & Tahilramani, N. (2016). "Digital Speech Watermarking for Authenticity of Speaker in Speaker Recognition System". *International Conference on Micro-Electronics and Telecommunication Engineering*, pp.105-109.

Doddington, G.R. (1985). "Speaker Recognition – Identifying People by Their Voices". *Proceedings of the IEEE*. 73(11), pp. 1651-1665.

Du, K.-L & Swamy, M.N.S. (2014). "Recurrent Neural Networks". 10.1007/978-1-4471-5571-3_11.

Du, S. & Li, J. (2019). "Parallel Processing of Improved KNN Text Classification Algorithm Based on Hadoop". *2019 7th International Conference on Information, Communication and Networks (ICICN)*, pp. 167-170, doi: 10.1109/ICICN.2019.8834973.

Elmissaouri, R., Sakly, A. & M'Sahli, F. (2013). "Optimized FPGA Implementation of an Aritificial Neural Netwirk for Function Approximation". *International*

Journal of Emerging Trends in Engineering & Development, 3(1), pp. 474-490.

Falaka, B., Saputra, R., Setianingsih E. C. & Murti, M. A. (2021) "Sea Wave Detection System Using Web-Based Naive Bayes Algorithm". *2021 3rd International Conference on Electronics Representation and Algorithm (ICERA)*, pp. 57-60, doi: 10.1109/ICERA53111.2021.9538697.

Fawcett, T. (2006). "An Introduction to ROC Analysis". *Pattern Recognition, Lett*., vol 27, no. 8, pp. 861-874.

Furui, S. (2009). 40 Years of Progress in Automatic Speaker Recognition. In: Tistarelli M. & Nixon M.S. (Eds). "Advances in Biometrics". ICB 2009. Lecture Notes in Computer Science, vol 5558. Springer, Berlin, Heidelberg.

Gaikwad, S.K., Gawali, B.W. & Yannawar, P. (2010). "A Review on Speech Recognition Technique". *International Journal of Computer Applications*, 10(3), pp.16-24.

Ganchev, T. (2011). "Contemporary Methods for Speech Parameterization". Springer New York Dordrecht Heidelberg London. Retrieved from: books.google.com.my/books.

Gaurav, Deiv, D.S., Sharma, G.K. & Bhattacharya, M. (2012). "Development of Application Specific Continuous Speech Recognition System in Hindi". *Journal of Signal and Information Processing*, 3, pp. 394-401.

Ghadge, S.A., Janvale, G.B. & Deshmukh, R.R. (2010). "Speech Feature Extraction Using Mel-Frequency Cepstral Coefficient (MFCC)", *Proceedings of Emerging Trends in Computer Science, Communication and Information Technology*, pp. 503-506.

Gish, H. & Schmidt, M. (1994). "Text-Independent Speaker Identification". *IEEE Signal Processing Magazine*, pp. 18-31, October 1994.

Gupta, H. & Gupta, D. (2016). "LPC and LPCC Method of Feature Extraction in Speech Recognition System". *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, Noida, 2016, pp. 498-502, doi: 10.1109/CONFLUENCE.2016.7508171.

Haggerty, M. (2008). "Chapter 9: Automatic Speech Recognition". Unpublished work by Pearson Education Inc. (Melinda_Haggerty@prenhall.com).

Hariharan, M., Fook, C.Y., Sindhu, R, Adom, A.H. & Yaacob, S. (2013). "Objective Evaluation of Speech Dysfluencies using Wavelet Packet Transform with Sample Entropy". Digital Signal Processing. Elsevier Inc. pp. 952-959.

Holmes, J. N. & Holmes, W. (2002). "Speech Synthesis and Recognition". London: Taylor and Francis.

Ibrahim, Y.A., Odiketa, J.C. & Ibiyemi, T.S. (2017). "Preprocessing Technique in Automatic Speech Recognition for Human Computer Interaction: An Overview". Retrieved from https://anale-informatica.tibiscus.ro/download/lucrari/15-1-23-Ibrahim.pdf.

Imam, S.A., Bansal, P. & Singh, V. (2017). "Review: Speaker Recognition Using Automated Systems". *AGU International Journal of Engineering and Technology (AGUIJET),* Vol. 5, pp. 31-39.

Indramalar, S. (2016). "Meet Nadine, the Robot Who Can One Day Help Seniors". The Star Online. Retrieved from: https://www.thestar.com.my/lifestyle/living/2016/08/29/meet-nadine-the-robot-who-can-one-day-help-seniors.

Jahangir, R., Teh, Y.W., Memon, N.A., Mujtaba, G., Zareei, M., Ishtiaq, U., Akhtar, M.Z. & Ali, I. (2020). Text-Independent Speaker Identification through Feature Fusion and Deep Neural Network. *IEEE Access,* vol. 8, pp. 32187-32202, doi: 10.1109/ACCESS.2020.2973541.

Jain, A. & Sharma, O.P. (2013). "A Vector Quantization Approach for Voice Recognition Using Mel Frequency Cepstral Coefficient (MFCC): A Review". *International Journal of Electronics & Communication Technology*. 4(4), pp. 26-29.

Jamal, N., Shanta, S., Mahmud, F. & Sha'abani, MNAH. (2017). "Automatic Speech Recognition (ASR) based Approach for Speech Theraphy of Aphastic Patients: A Review". *AIP Conference Proceedings*, 1883(1).

Janse, P.V., Magre, S.B, Kurzekar, P.K. & Deshmukh, R.R. (2014). "A Comparative Study Between MFCC and DWT Feature Extraction Technique". *International Journal of Engineering Research & Technology (IJERT),* 3(1), pp. 3124-3127.

Jawarkar, N., Holambe, R. & Basu, T. (2011). "Use of Fuzzy Min-Max Neural Network for Speaker Identification", in Recent Trends in Information Technology (ICRTIT), 2011 International Conference on, pp. 178-182.

Jawarkar, N. P., Holambe, R. S. & Basu, T. K. (2013). "Speaker Identification Using Whispered Speech", in Communication Systems and Network Technologies (CSNT), 2013 International Conference on, 2013, pp. 778-781.

Jiang, K., Pan, D., Jiang T. & Yuan, Y. (2018). "Ocean Surface Stochastic Channel Modeling based nn Hidden Markov Model". *2018 IEEE Asia-Pacific Conference on Antennas and Propagation (APCAP)*, 2018, Pp. 440-441, Doi: 10.1109/Apcap.2018.8538148.

Joshi, S.C. & Cheeran, A.N. (2014). "MATLAB Based Feature Extraction Using Mel Frequency Cepstrum Coefficients for Automatic Speech Recognition". *International Journal of Science, Engineering and Technology Research (IJSETR)*, 3(6), pp. 1820-1823.

Juan, S. S., Besacier, L. & Tan, T. (2012). "Analysis of Malay Speech Recognition for Different Speaker Origin". *2012 International Conference on Asian Language Processing*, pp. 229-232, doi: 10.1109/IALP.2012.23.

Kamble, B.C. (2016). "Speech Recognition Using Artificial Neural Network – A Review". *International Journal of Computing, Communications, and Instrumentation Engineering (IJCCIE)*, 3(1), pp. 1-4.

Kaphungkui, N.K. & Kandali, A.B. (2019). "Text Dependent Speaker Recognition with Back Propagation Neural Network". I*nternational Journal of Engineering and Advanced Technology (IJEAT)*, 8(5), pp. 1431-1434.

Kaur, K. & Jain, N. (2015). "Feature Extraction and Classification for Automatic Speaker Recognition System – A Review". *International Journal of Advanced Research in Computer Science and Software Engineering*, 5(1), pp. 1-6.

Kouemou, G. L. (2011). "History and Theoretical Basics of Hidden Markov Models, Hidden Markov Models, Theory and Applications". Przemyslaw Dymarski, IntechOpen, DOI: 10.5772/15205. Available from: https://www.intechopen.com/books/hidden-markov-models-theory-and-applications/history-and-theoretical-basics-of-hidden-markov-models.

Král, P. (2010). "Discrete Wavelet Transform for Automatic Speaker Recognition", in Image and Signal Processing (CISP), 2010 3rd International Congress on, 2010, pp. 3514-3518.

Kumar, P. & Chandra, M. (2011) "Hybrid of Wavelet and MFCC Features for Speaker Verification". *IEEE World Congress on Information and Communication Technologies (WICT)*, Mumbai, pp. 1150-1154.

Lantz B. Machine learning with R. 2nd ed. Birmingham: Packt Publishing; 2015:1.

Li, L., Lin, Y., Zhang, Z., L. & Wang, D. (2015). "Improved Deep Speaker Feature Learning for Text-Dependent Speaker Recognition", in APSIPA Annual Summit and Conference, pp. 426-429.

Li, R., Sun, X., Liu, Y., Yang, D. & Dong, L. (2019). "Multi-resolution auditory cepstral coefficient and adaptive mask for speech enhancement with deep neural network". *EURASIP Journal Advances in Signal Processing*. 22 (2019). https:/doi.org/10.1186/s13634-019-0618-4.

Li, Q., Yang, Y., Lan, T., Zhu, H., Wei, Q., Qiao, F., Liu, X. & Yang, H. (2020). "MSP-MFCC: Energy-Efficient MFCC Feature Extraction Method with Mixed-Signal Processing Architecture for Wearable Speech Recognition Applications". *IEEE Access*, vol. 8, pp. 48720-48730, 2020, doi: 10.1109/ACCESS.2020.2979799.

Lian, Z., Xu, K., Wan, J. & Li, G. (2017). "Underwater Acoustic Target Classification Based on Modified GFCC Features". *IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 2017, pp. 258-262, doi: 10.1109/IAEAC.2017.8054017.

Lim, H. T., Huspi, S. H. & Ibrahim R. (2021). "A Conceptual Framework for Malay-English Mixed-language Question Answering System". *2021 International Congress of Advanced Technology and Engineering (ICOTEN)*, 2021, pp. 1-8, doi: 10.1109/ICOTEN52080.2021.9493503.

Malik, S. & Afsar, F. A. (2009). "Wavelet Transform based Automatic Speaker Recognition". *IEEE 13th International Multitopic Conference, INMIC, Islamabad*, pp. 1-4.

Mengxi, Z. & Zhiguo, T. (2020). "Research on Failure Identification of Partial Discharge Ultrasonic Signal Based on GFCC". *2020 IEEE Electrical Insulation Conference (EIC)*, 2020, pp. 412-416, doi: 10.1109/EIC47619.2020.9158683.

Mishra, M. & Srivastava, M. (2014). "A View of Artificial Neural Network". *IEEE International Conference on Advances in Engineering & Technology Research (ICAETR)*, India.

Mohammed, R.A., Ali, A.E. & Hassan, N.F. (2019). J*ournal of Al-Qadisiyah for Computer Science and Mathematics*, 11(3), pp. 21-30.

Moinuddin, M. & Kanthi, A.N. (2014). "Speaker Identification based on GFCC using GMM". *International Journal of Innovative Research in Advanced Engineering (IJIRAE).*, pp. 224-232.

Mouaz, B., Abderrahim, B-h. & Abdelmajid, E. (2019). "Speech Recognition of Moroccan Dialect Using Hidden Markov Models". IAES International Journal of Artificial Intelligence (IJ-AI), vol. 8(1), March 2019, pp. 7-13, ISSN: 2252-8938.

Muda, L., Begam, M. & Elamvazuthi, L. (2010). "Voice Recognition Algorithms Using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques". *Journal of Computing*. 2(3), pp. 138- 143.

Mustafa, M. B., Don, Z. M. & Knowles, G. (2013). "Context-dependent Labels for an HMM-based Speech Synthesis System for Malay HMM-based Speech Synthesis System for Malay". *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, 2013, pp. 1-4, doi: 10.1109/ICSDA.2013.6709884.

Nagaraj, S., Nai-Peng, T., Chiu-Wan, N., Kiong-Hock, L. & Pala, J. (2009). "Counting Ethnicity in Malaysia: The Complexity of Measuring Diversity". In: P. Simon, V. Piché, A. Gagnon (eds), *Social Statistics and Ethnic Diversity. IMISCOE Research Series*, Springer, Cham.

Najafian, M., Safavi, S., Hansen, J. H. L. and M. Russell, "Improving Speech Recognition using Limited Accent Diverse British English Training Data with Deep Neural Networks", *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2016, pp. 1-6, doi: 10.1109/MLSP.2016.7738854.

Nandhini, S. & Shenbagavalli, A. (2014). "Voiced/Unvoiced Detection using Short Term Process", *IJCA Proceedings on International Conference on Innovations in Information, Embedded and Communication Systems ICIIECS* (2): 39-43.

Nematollahi, M.A. & Al-Haddad, S.A.R. (2015). "Distant Speaker Recognition: An Overview". *International Journal of Humanoid Robotics*, vol. 12, pp. 1-45.

Nugroho, E. Noersasongko, K., Purwanto, Muljono, Setiadi, D.R.I.M. (2021). "Enhanced Indonesian Ethnic Speaker Recognition using Data Augmentation Deep Neural Network". *Journal of King Saud University – Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2021.04.002.

Olukoya & Musiliu, B. (2020). "Comparison of Feature Selection Techniques for Predicting Student's Academic Performance". *International Journal of Research and Scientific Innovation (IJRSI), Vol. VII, Issue VIII, August 2020, pp. 97-101. ISSN 2321-2705.*

Porta M., Dondi, P., Zangrandi, N. & Lombardi L. (2021). "Gaze-Based Biometrics from Free Observation of Moving Elements". *IEEE Transactions on Biometrics, Behavior, and Identity Science*, doi: 10.1109/TBIOM.2021.3130798.

Qasim, M., Nawaz, S., Hussian, S. & Habib, T. (2017). "Urdu Speech Recognition System for District Names of Pakistan: Development, Challenges and Solutions". *2016 Conference of the Oriental Chapter of International Speech Databases and Assessment Technique (O-COCOSDA),* 26-28 October 2016, Bali, Indonesia.

Rabiner, L.R. (1989). "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". Proceedings of the IEEE, Vol. 77, No. 2, pp. 257-286.

Rabiner, L.R. & Juang, B.H. (1993). "Fundamentals of Speech Recognition". Englewood Cliffs, NJ: Prentice-Hall.

Rabiner, L.R. & Schafer, R.W. (2007). "Introduction to Digital Speech Processing". Foundations and Trends in Signal Processing, vol. 1, No. 33-35.

Rajasekhar, A. & Hota, M.K. (2018). "A Study of Speech, Speaker and Emotion Recognition using Mel Frequency Cepstrum Coefficients and Support Vector Machines". *International Conference on Communication and Signal Processing (ICCSP),* Chennai, 2018, pp. 0114-0118, doi: 10.1109/ICCSP.2018.8524451.

Ranjan, R. & Thakur, A. (2019). "Analysis of Feature Extraction Techniques for Speech Recognition System". *International Journal of Innovative Technology and Exploring Engineering (IJITEE),* Vol. 7, Issue 7C2, May 2019, ISSN:2278-3075.

Reddy Gade, V. S. & Sumathi, M. (2021). "A Comprehensive Study on Automatic Speaker Recognition by using Deep Learning Techniques". *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI),* 2021, pp. 1591-1597, doi: 10.1109/ICOEI51242.2021.9452885.

Reynolds, D., Andrews, W., Campbell, J., Navratil, J., Peskin, B., Adami, A., Jin, Q., Klusacek, D., Abramson, J., Mihaescu, R., Godfrey, J., Jones, D. & Xiang, B. (2003). "The SuperSID project: exploiting high-level information for high-accuracy speaker recognition". Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on. 4. IV - 784. 10.1109/ICASSP.2003.1202760.

Rituerto-González, E., Mínguez Sánchez, A. & Gallardo-Antolín, A. & Peláez-Moreno, C. (2019). "Data Augmentation for Speaker Identification under Stress Conditions to Combat Gender-Based Violence". Applied Sciences. 9. 2298. 10.3390/app9112298.

Rosdi, F. (2016). "Fuzzy Petri Nets as a Classification Method for Automatic Speech Intelligibility Detection of Children with Speech Impairments". (PhD Thesis).

Rosenberg et al. (2007). "L16: Speaker Recognition". Lecture Slides. Retrieved from: http://research.cs.tamu.edu/prism/lectures/sp/l16.pdf.

Rousan, M. & Intrigila, B. (2020). "A Comparative Analysis of Biometrics Types: Literature Review". *Journal of Computer Science*. 16. 1778-1788. 10.3844/jcssp.2020.1778.1788.

Roy, P. & Das, P. K. (2021). "Review of Language Identification Techniques". *2010 IEEE International Conference on Computational Intelligence and Computing Research*, 2010, pp. 1-4, doi: 10.1109/ICCIC.2010.5705780.

Salim, A. P., Laksitowening, K. A. & Asror, I. (2020) "Time Series Prediction on College Graduation Using KNN Algorithm". *2020 8th International Conference on Information and Communication Technology (ICoICT)*, pp. 1-4, doi: 10.1109/ICoICT49345.2020.9166238.

Salleh, S. S., Bujang, A., Chachil, K. & Wan Ismail, W. A. Z. (2018). "Analysis of Prominent Malay Da'i Voices Frequency and Characteristics". *2018 8th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, 2018, pp. 26-30, doi: 10.1109/ICCSCE.2018.8685010.

Sangwan, P. (2017). "Feature Extraction for Speaker Recognition: A Systematic Study". *Global Journal of Enterprise Information System*, Volume 9, Issue 4, October-December 2017.

Sanjaya, W.S.M., Anggraeni, D. & Santika, I.P. (2018). "Speech Recognition using Linear Predictive Coding (LPC) and Adaptive Neuro-Fuzzy (ANFIS) to

Control 5 DoF Arm Robot". *International Conference on Computation in Science and Engineering*, pp. 1-10.

Sarmah, K. (2017). "Comparison Studies of Speaker Modeling Techniques in Speaker Verification System". *International Journal of Scientific Research in Computer Science and Engineering*, 5(5), pp.75-82.

Saste, S.T. & Jagdale, S.M. (2017). "Comparative Study of Different Techniques in Speaker Recognition: Review". *International Journal of Advanced Engineering, Management and Science (IJAEMS)*, 3(2), pp. 284- 287.

Satriyanto, N., Munir, R. & Harlili. (2019). "Dynamic Background Video Forgery Detection using Gaussian Mixture Model". *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*, 2019, pp. 379-383, doi: 10.1109/ICAIIT.2019.8834463.

Schafer, R.W. & Rabiner, L.R. (1975). "Digital Representations of Speech Signals". Proceedings of the IEEE, Vol. 63, No. 4, April 1975.

Sharma, A. M. (2019). "Speaker Recognition Using Machine Learning Technique". (Master's Projects). Retrieved from https://scholarworks.sjsu.edu/etd_projects/685.

Sharma, A. & Singla, S.K. (2017). "State-of-the-art Modeling Techniques in Speaker Recognition". *International Journal of Electronics Engineering*, 9(2), pp. 186-195.

Sharma, V., & Bansal, P.K. (2013). "A Review on Speaker Recognition Approaches and Challenges". *International Journal of Engineering Research & Technology*. 2(5), pp. 1580-1588.

Shaver, C. D. & Acken, J. M. (2016). "A Brief Review of Speaker Recognition Technology". *Electrical and Computer Engineering Faculty Publications and Presentations*. 350. Retrieved from https://pdxscholar.library.pdx.edu/ece_fac/350.

Shi, X., Yang, H. & Zhou, P. (2016). "Robust speaker recognition based on improved GFCC," *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, Chengdu, 2016, pp. 1927-1931, doi: 10.1109/CompComm.2016.7925037.

Singh, N. (2014). "A Study on Speech and Speaker Recognition Technology and its Challenges". *Proceedings of National Conference on Information Security Challenges*, pp. 34-36.

Singh, N., Agrawal, A. & Ahmad Khan, R. (2015). “A Critical Review on Automatic Speaker Recognition”. *Science Journal of Circuits, Systems and Signal Processing,* 4(2), pp. 14-17.

Singh, N., Agrawal, A. & Khan, R.A. (2018). “The Development of Speaker Recognition Technology”. *International Journal of Advanced Research in Engineering & Technology (IJARET),* 9(3), pp. 8-16.

Singh, S. (2018). “Speaker Recognition by Gaussian Filter Based Feature Extraction and Proposed Fuzzy Vector Quantization Modelling Technique”. *International Journal of Applied Engineering Research,* 13(16), pp. 12798-12804.

Soleymanpour, M. & Marvi, H. (2017). “Text-independent speaker identification based on selection of the most similar feature vectors”. *International Journal of Speech Technology,* vol. 20, pp. 99-108, 2017.

Sturim, D.E., Campbell, W.M. & Reynolds, D.A. (2007). “Classification Methods for Speaker Recognition”. *Proceedings of Speaker Classification I: Fundamentals, Features, and Methods,* pp. 278-297. 10.1007/978-3-540-74200-5_16.

Suchitha, T.R. & Bindu, A.T. (2015). “Feature Extraction using MFCC and Classification using GMM”. *International Journal for Scientific Research & Development (IJSRD),* 3(5), pp. 1278-1283.

Sugan, N., Sai Srinivas, N. S., Kar N., Kumar L. S., Nath M. K. & Kanhe A. (2018). "Performance Comparison of Different Cepstral Features for Speech Emotion Recognition". *2018 International CET Conference on Control, Communication, and Computing (IC4),* Thiruvananthapuram, 2018, pp. 266-271, doi: 10.1109/CETIC4.2018.8531065.

Sui, X., Wang, H. & Wang, L. (2014). “A General Framework for Multi-accent Mandarin Speech Recognition using Adaptive Neural Networks," *The 9th International Symposium on Chinese Spoken Language Processing*, pp. 118-122, doi: 10.1109/ISCSLP.2014.6936621.

Sujiya, S. & Chandra, E. (2017). “A Review on Speaker Recognition”. *International Journal of Engineering and Technology (IJET),* 9(3), pp. 1592-1598.

Sun, B-Y. & Huang, D-S. (2003). “Support Vector Clustering for Multiclass Classification Problems," *The 2003 Congress on Evolutionary Computation, 2003. CEC '03.*, Canberra, ACT, Australia, 2003, pp. 1480-1485 Vol.2.

Tamazin, M., Gouda, A. & Khedr, M. (2019). "Enhanced Automatic Speech Recognition System Based on Enhancing Power-Normalized Cepstral Coefficients". Applied Sciences, pp. 1-13. Retrieved from: www.mdpi.cpm.

Tanaka, F., Isshiki, K. & Takahashi, F. (2015). "Pepper Learns Together with Children: Development of an Educational Application". I*EEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, November 3-5, Seoul, Korea.

Tazi, E.B. (2016). "A Robust Speaker Identification System based on the combination of GFCC and MFCC Method". *5th International Conference on Multimedia Computing & System (ICMCS)*, Marrakech, 2016, pp. 54-58, doi: 10.1109/ICMCS.2016.7905654.

Tazi, E. B., Benabbou A. & Harti, M. (2012). "Efficient Text Independent Speaker Identification based on GFCC and CMN Methods," *2012 International Conference on Multimedia Computing and Systems*, 2012, pp. 90-95, doi: 10.1109/ICMCS.2012.6320152.

Tolba, H. (2011). " A high-performance text-independent speaker identification of Arabic speakers using a CHMM-based approach". *Alexandria Engineering Journal*, 50, pp. 43-47.

Taunk, K., De, S., Verma S. & Swetapadma, A. (2019). "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification". *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, pp. 1255-1260, doi: 10.1109/ICCS45141.2019.9065747.

Upadhyay, R. & Lui, S. (2018). "Foreign English Accent Classification Using Deep Belief Networks". *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pp. 290-293, doi: 10.1109/ICSC.2018.00053.

Zhang, Z. (2016). "Introduction to Machine Learning: K-Nearest Neighbors". Ann Transl Med 2016; 4(11):218. doi: 10.21037/atm.2016.03.37.

Zhang, N. & Yao, Y. (2020). "Speaker Recognition based on Dynamic Time Warping and Gaussian Mixture Model". *2020 39th Chinese Control Conference (CCC)*, 2020, pp. 1174-1177, doi: 10.23919/CCC50068.2020.9188632.

Zhang, Y. & Ni, L. (2017). "Feature Extraction Algorithm using GFCC and Phase Information". *2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 2017, pp. 1163-1167, doi: 10.1109/IAEAC.2017.8054196.

Zhaoyan, Z. (2016). “Mechanics of Human Voice Production and Control”. *The Journal of the Acoustical Society of America*, *140*(4), 2614. doi:10.1121/1.4964509.

Zheng, T.F & Li, L. (2017). “Robustness-Related Issues in Speaker Recognition”. Springer.

Zhu, W., Zeng, N. & Wang, N. (2010). “Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS Implementation.” *Health Care and Life Sciences,* NESUG 2010.