Detection of the Uniqueness of a Human Voice: Towards Machine Learning for Improved Data Efficiency

Saritha Kinkiri

A thesis submitted in partial fulfilment of the
requirements of the University of
Greenwich for the Degree of Doctor of
Philosophy

June 2021

I would like to dedicate this thesis to my loving parents, sister and brother ...

Declaration

I certify that the work contained in this thesis, or any part of it, has not been accepted in substance for any previous degree awarded to me or nay other person, and is not concurrently being submitted for any other degree which has been studied at the University of Greenwich, London, UK.

I also declare that the work contained in this thesis is the result of my own investigations, except where otherwise identified and acknowledged by references. I further declare that no aspects of the contents of this thesis are the outcome of any form of research misconduct.

I declare any personal, sensitive or confidential information/data has been removed or participants have been anonymised. I further declare that where any questionnaires, survey answers or other qualitative responses of participants are recorded/included in the appendices, all personal information has been removed or anonymised. Where University forms (such as those from the Research Ethics Committee) have been included in appendices, all handwritten/scanned signatures have been removed.

Saritha Kinkiri
June 2021

Acknowledgements

The work would not be carried out so smoothly without the help, assistance and support from many people. I would like to thank the following people:

Supervisor, Prof. Simeon Keates, for sharing knowledge without reservation, providing feedback for my thesis and patient guidance. Co-supervisor, Prof. Phil Cox and Dr. Stuart Ashenden, for always having a moment to spare, as well as inspiring and motivating me throughout my project period.

I would like to say thanks to all participants, for their great help and support on my speech database recording work. I would like to say special thanks to a few of my friends who gathered many UnderGraduate and PhD students, for participating as my recording subjects.

My Friends, Arwa and Kousalya, for discussing various issues and idea sharing. They made the hard working period interesting and relaxing. Special thank goes to Arwa for her proofreading and inspiration. All the people who supported my speech database building work, for their kindness, their voice messages and time dedication.

Abstract

The aim of this thesis is to characterise voice characteristics that can establish the identity of the person who is speaking, independent of the language used. The fundamental goal of the work is to understand how humans recognise a speaker. The voice parameters such as: speech rate, natural pauses & intended or unintended speaker pauses, fundamental frequencies, phoneme generation, volume etc. since the combination of all the voice parameters cannot be easily imitated by another person. It is an assumption that different speakers speak differently, however, it is important to understand and remember that the same speaker's voice will change over time. For example, the speaker cannot speak/talk/say the same thing in exactly the same way time after time. However, these differences/variations in speech can be audible and measured by using combinations of voice parameters.

The aim is to eliminate a speaker whom we are not looking for. Individuals use words to communicate with others and the same method to communicate with machines too. Humans successfully use speech software (which is speech to text) to talk to telephones instead of tapping words on the keyboard. But machines are proven to be good at converting speech to text, although not at identifying who is speaking.

Problems remain in recognising an individual from their speech whilst proving reliable, repeatable & robust otherwise the speaker could, for example, find themselves locked out of their online voice accessed. For example, the risks are asymmetric - if one in 100 people is locked out of an account that is not too serious, as customer services will ask for answers to security questions. However, if one in 100 people get into bank account fraudulently this is a bigger problem.

A speaker's voice varies in frequency, tone, and volume sufficiently enough to uniquely identify an individual. However, other factors can contribute to this uniqueness: the size and shape of the mouth, throat, nose, and vocal cords. Sound is produced by air passing from the lungs through the throat, vocal cords and then mouth. A voice makes different sounds based on the position of mouth and throat. It is the variation of these attributes that allows for identification.

Speaker recognition systems are already available, but their overall accuracy is limited because of several issues such as extracted features based on very short time window of

speech and models fail to capture useful information of a speaker since current speech recognition systems and extracted features are language-dependent. By using the voice parameters,the work here was able to eliminate 80 percent of population to be able to identify a person. Recognising 1 out 100 is difficult, but identifying 1 out 5 is comparatively easy.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

As humans, we have several modes of communication available to us, such as speech, gestures, text, drawing, etc. Speech is one of the most efficient ways of communication [1, 2]. It has various characteristics that help us to identify not only words but also the gender, attitude, health, and often even the identity of the speaker. The human voice is the most powerful model of communication and individuals use their voice to communicate with machines surrounding us, too. Identifying a sound from different sources such as sound from animals, musical instruments, vehicles, etc, seems easy for humans [3, 4], but, it is difficult for machines, for example, recognition/identification of a sound from musical instruments, etc [5].

Human capabilities are being complemented increasingly by the advancement of speech recognition systems, artificial intelligence, neural networks and the processing power of a machine is often achieved, simply via a voice command [3]. These voice assistants can support interaction live from anywhere in the world via smartphones, digital wrist phones, etc. Thus, there is a trend of voice-enabled computing and its opportunities, implementing many applications in the real world [6]. Voice assistants are already a part of the daily routine for millions of people.

There is a demand in the market for speech-based biometric systems to improve the securing of technologies including an increasing number of voice control systems such as Siri, Alexa, and Google Assistant, etc. These are the main devices that are expected to drive the growth of speech-enabled technologies in the real world. However, these systems are often inefficient because there is a lack of training data, an increase in population and change of environment, etc. There are lots of industries, academic institutions, and commercial companies, trying to use voice as authentication for many applications such as unlocking smartphones, operating electronic devices by a user's voice, and online banking, etc.

Unfortunately, the human voice is incredibly difficult to analyse and it is not as easy to be understood and then decoded for usage in the machines. With advancing technologies, machines can now drive cars, enter phone numbers from a user speech, predict stock prices, detect a disease in its initial stage, etc. However, machines still struggle to understand human speech and they are unable to communicate and chat with humans the way we converse with our neighbours and friends, etc. The question arises, how do humans communicate? How do we listen, understand, remember, and then recognise?

Listening:

Communication is natural for humans, but it is difficult for a machine. Humans process sound signals and remove background noises by themselves, and then concentrate on the way a speaker is pronouncing words (accent) and replies to the receiver. This process is referred to as speech recognition for human-machine interaction.

Understanding:

Humans can listen and understand a conversation. With the help of memory and recognition, sometimes mispronunciation of a word can still be understood by an individual, that is, by processing the information given before and after the word and also taking into consideration the context of the topic being discussed. However, machines convert that word into the most closely sounding word, unlike humans, who substitute (or process) the word that is the most relevant to the topic. Thus, machines can produce errors such as changing the meaning of the sentences or generating a nonsense sentence.

Importance of Context:

Consideration of context is also a challenge. It includes many factors such as: What has been said in the previous conversation, relationship with a speaker, situational context, etc. Machines still struggle with this context concept and they often fail to understand the context.

The overall production of the complete human voice is a combination of the soundbox, physical characteristics such as weight of the body, height, etc., a measurement from other physical characteristics e.g.: shape and size of the nasal cavity, chest, etc., which creates a unique feature that helps to be used as biometry, the same way that fingerprint of unique to an individual.

Biometric Authentication (BA) aims to use a person's unique characteristics to identify them. BA is a technology that helps to reduce fraud cases since every person's biometric information is unique to her/himself. The word biometric originates from Greek, "bio" refers

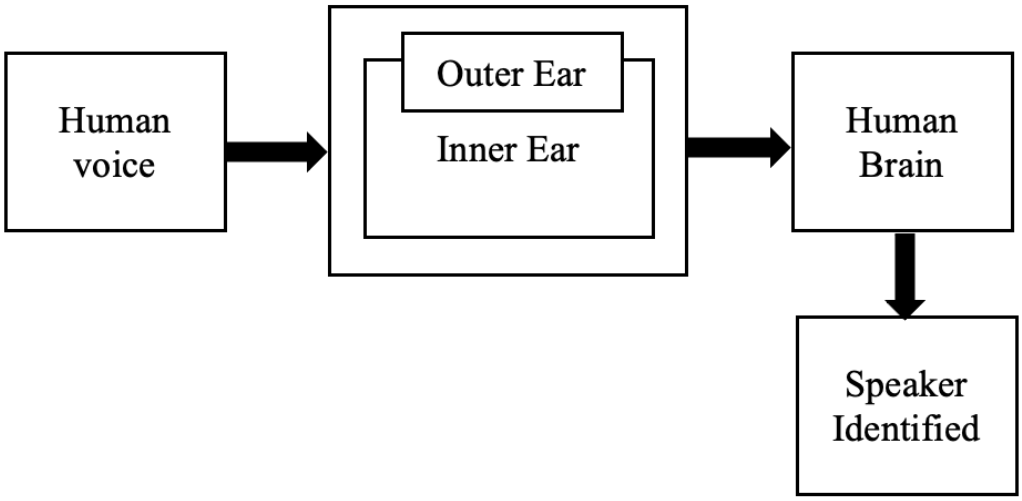

Fig. 1.1 How Human Process a Voice to Identify a Speaker

to life, and "metric" means to measure, when combined they describe how one "measures a person's life" [7]. Biometric technology is considered to have two types of categories: physical and/or behavioural. Physical characteristics include DeoxyriboNucleic Acid (DNA), fingerprints, facial recognition, and iris/retina scan, while behavioural characteristics include voice, handwriting, and signature [8–10].

Biometric voice recognition systems focus on identifying the unique characteristics of a voice and store those in a database for future use. To identify a speaker, a voice recognition system needs to understand the characteristics of the voice, which includes both physical and behavioural characteristics [11]. One of the problems is that most of the characteristics are physical, which means they cannot measure themselves. For instance, one cannot measure the length of the mouth or nose cavities, the weight of the person's head, etc. Therefore, a machine needs to determine and understand how to identify these characteristics from the voice signal itself, since that is the only available data to measure technique. Human physical characteristics would not change when they talk in different languages. Then the question arises, can a machine identify a speaker when they talk in another language that is not English? For example, can a system that can recognise the identity of a voice of a native English speaker identify the same voice when speaking e.g.: Dutch/French language.

Furthermore, the system should be able to deal with challenges such as voice imitation, which can be particularly challenging, because now the standard physical characteristics are adapted on purpose to create a specific output. Using a human's voice profile data could also be useful for speech recognition tools if one can automatically detect who is speaking and then adjust the speech recognition profile to improve overall accuracy.

1.1 Aim

This research is aimed to investigate the advantages and drawbacks of current methodologies and propose a method to identify a speaker independent of language used.

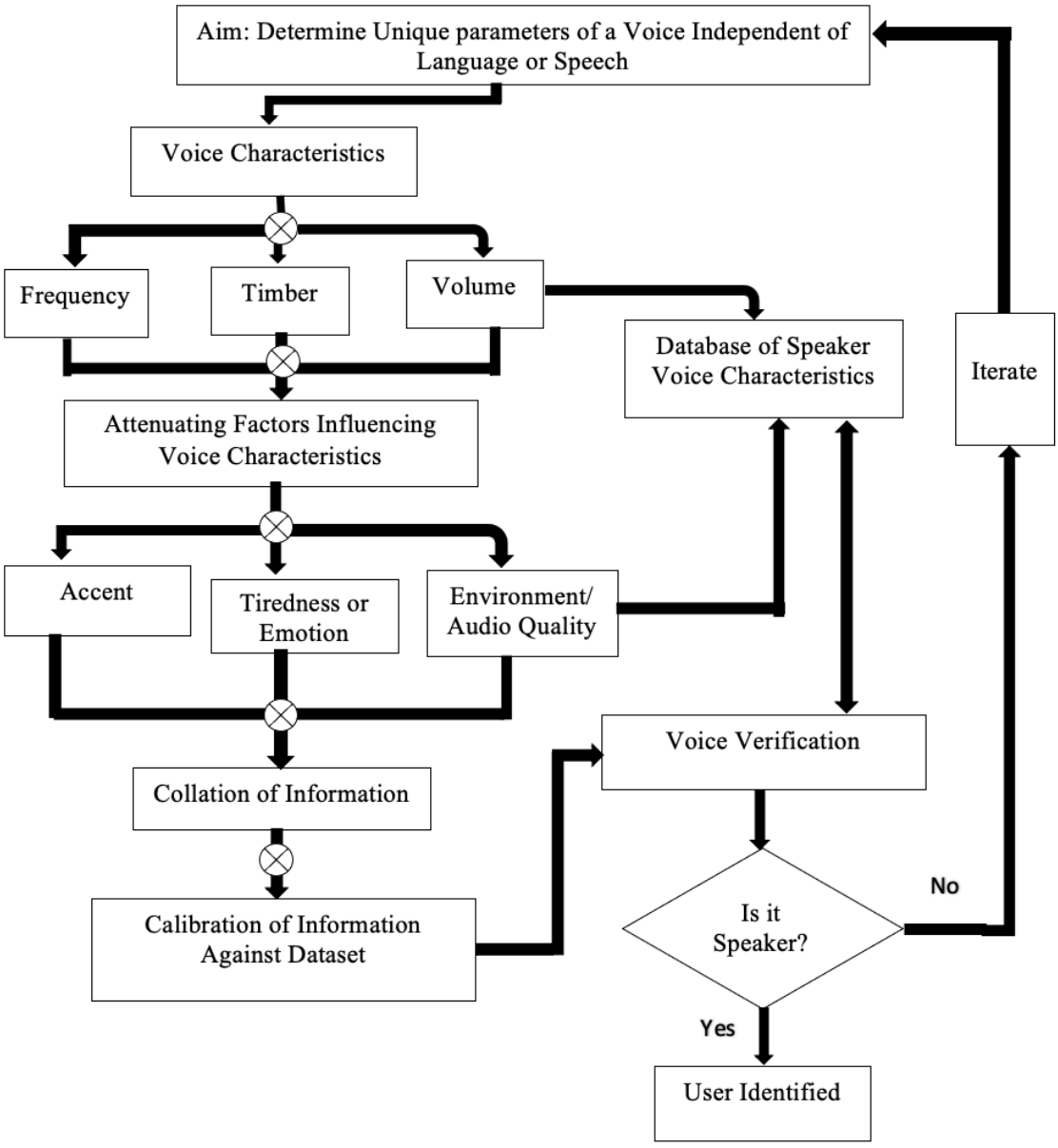

Fig. 1.2 Aim of the Research

1.2 Objectives

1. In depth knowledge of existing methodologies.
2. Understanding how humans learn and how they can apply their learning skill to identify an object/person/sound ?
3. Design and conduct a survey to see how human beings recognise a speaker from their voice.
4. Analyse the characteristics of human voices.
5. Design a framework to identify a speaker based on characteristics of their voice.
6. Validation of the framework by comparing with objective 1 analysis.

1.3 Thesis Outline

The present study is designed for identifying speakers independent of the language used in the speech. This is being done through proving acoustically that each individual is different from another according to the fundamental frequency, rate of speaking; and also through testifying the listener's perceptual abilities in perceiving and differentiating different speaking rates in different languages, pauses, accent, and pronunciation.

Therefore; the entire structure of the presented thesis is based upon voice characteristics Nevertheless; the thesis is divided into two main parts: the first the theoretical part that represents the history, definitions, and problems; the second is an experimental part to illuminate some of the theoretical problems. The theoretical part consists of two sections. The first section exposes a general introduction about how humans can learn and use that knowledge to identify a speaker, and how humans can identify a voice that is familiar and unfamiliar when the used language is known to them. Language usage is important because of the several levels of information it can reveal. Then, a clear distinction must be established between three terms; speaker recognition, identification, and verification. The second section of the theoretical part deals with the survey of participants recognising a speaker when they speak/talk in a different language that is familiar/unfamiliar to the participants.

The experimental part consists also of two sections: the first section involves the methodology and procedures of the main experiments; starting with gathering the data, the number of participants involved, the way of recordings, data analyses, the steps of analyses, the measurements, and finally the test procedures. The second section of the experimental part contains all the results of the main experiment; the results and measurements with their

statistical representations, for all of the involved informants, in addition to the perceptual results of the naïve listeners who joined in the experiment.

The underlying approach of the research is to observe and analyse language-independent speaker identification based on the fundamental characteristics of human voices. The aim is to explore speech parameters in both controlled and uncontrolled tasks, such as free speech or reading a script respectively. For example, this research will explore how many people out of a sample population of 100 participants can be excluded through using a combination of simple voice characteristics, such as dominant frequencies and pauses. The collection of data shows that a number of principal voice characteristics are independent of the language being spoken. The training and testing of the recogniser plays a major role in identifying a speaker, but in this thesis, the work will concentrate on the elimination of a speaker from a pool of potential candidates using the fastest possible means for the least amount of data training. There is other research available on recognising individuals from lots of data training, but this research is focused on trying to make the lightest weight system possible and to explore how much a security system can be enhanced for very little data training.

1.4 Publications

1. Saritha Kinkiri, Wim J.C Melis: 'Reducing Data Storage Requirements for Machine Learning Algorithms Using Principle Component analysis'; 1st International Conference on Applied System Innovation (ICASI) , on 22 to 25th of May 2016, Okinawa, Japan and Published on IEEE (DOI: 10.1109/ICASI.2016.7539804).

2. Saritha Kinkiri, Wim J.C Melis and Simeon Keates: 'Creating Patterns for Machine Learning Using Multiple Alignment Making'; 1st International Conference of Human Brain Project (HBP), on 6 to 8th of February 2017, Vienna, Austria (DOI: 10.3389/978-2-88945-421).

3. Saritha Kinkiri, Wim J.C Melis and Simeon Keates: 'Machine Learning for Voice Recognition'; Second Medway Engineering Conference on Systems on 6th June 2017, London, United kingdom.

4. Saritha Kinkiri and Simeon Keates: 'Identification of a Speaker from Familiar and Unfamiliar voices'; 5th International Conference on Robotics and Artificial Intelligence, on 22 to 24th of November, 2019, Singapore ACM (DOI:10.1145/3373724.3373742).

5. Saritha Kinkiri and Simeon Keates: ‘Characteristics of a Human Voice’; 2nd International Conference on Advance in Signal Processing and Artificial Intelligence on 30 June -2 July , 2020, Berlin, Germany.

6. Saritha Kinkiri and Simeon Keates: ‘Phonemes: An Explanatory study Applied to Identify a Speaker’ ; 2nd International Conference on Machine Learning, Image processing, Network Security and Data Science on 18th -19th June 2020, Silchar, India Published on Volume 1241 of the Communications in Computer and Information Science series (DOI: 10.1007/978-981-15-6318-8-6).

7. Saritha and Simeon Keates: ‘ Speaker Identification: Variations of a Human Voice’; 6th International Conference on Advances in Computing and Communication Engineering on 22 to 25th July 2020, Las Vegas, USA, publsihed on IEEE (DOI: 10.1109/ICACCE49060.2020.9154998).

8. Saritha Kinkiri and Simeon Keates: ‘Applications of Speaker Identification for Universal Access’; 22nd International Conference on Human Computer Interaction on 19 -24 July Copenhagen, Denmark, published on Volume 12189 of the Lecture Notes in Computer Science series (DOI: 10.1007/978-3-030-49108-6-40).

Chapter 2

Background

2.1 Introduction to Speech Recognition Systems

In recent years, an increasing number of applications are being developed to improve the interaction between humans and machines, supporting a more "natural" interaction between them [12]. Humans communicate with each other using speech, gestures, writing text, drawings, facial expressions, and body and sign language. One of the modes of interaction between people is verbal communication, but machines still face certain challenges when using verbal communication to interact with individuals, and/or to identify a person. There are two ways humans can communicate with machines, that is, through speech recognition, and voice recognition. Currently, speech recognition systems can recognize human spoken words with 95 percent accuracy in the English language, which is similar to humans [13].

2.1.1 Elementary Concepts of Speech Recognition Systems

A speech recognition system converts speech to text as shown in Figure 2.1. Speech recognition is language-dependent and aims to recognise what was spoken, independent of factors such as accents and emotion [14]. Yet, speech recognition has some drawbacks and problems when converting speech to text, such as a speaker's accent. A machine needs more computational power and time to be trained for different languages and accents from the same person, requiring more data storage, etc.

A speech recognition system aims to not depend on the physical characteristics of the body, because one wants to understand the speech and to recognise the word, regardless of who is speaking. One factor that affects speech recognition is language and most often, the person's first language. If someone speaks a foreign language, their accent tends to be related to the person's first language. Emotions can also have a significant impact on

speech production. For example, tiredness may cause a person to mumble words instead of pronouncing them properly, which is still understood by others since we tend to combine contextual information with a prediction to derive meaning. However, current artificial systems are unable to derive context in the speech and/or benefit from prediction.

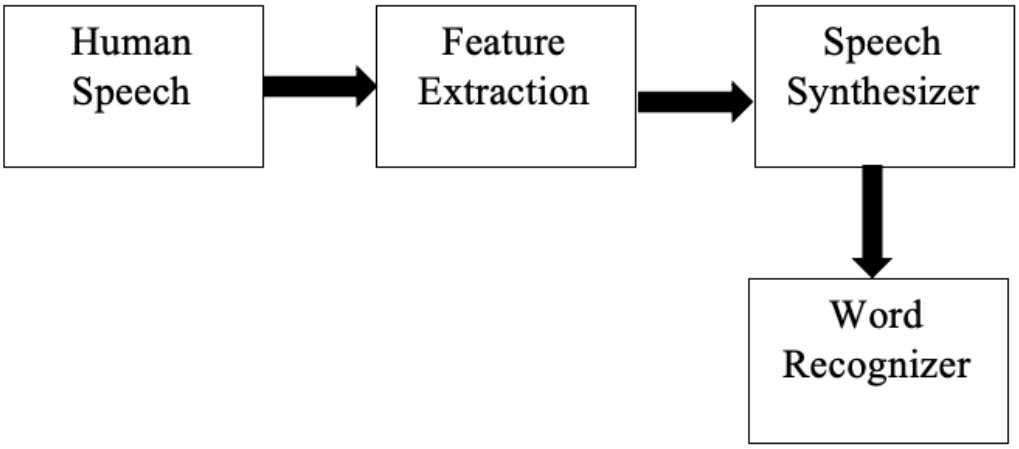

Fig. 2.1 Block Diagram of Speech Recognition System

Speech recognition technology can also be used to identify a speaker, which is known as voice recondition. It can be divided into two categories which are closed and open set recognition. In the closed set task, a speaker is determined from data that already exists. On the other side, open set task identification, the speaker needs to be identified from a database, where the target speaker does not already exist in the database.

Humans are good at speech recognition, whereas making a machine to this accomplish efficiently by itself is a difficult task [15]. Researchers have been able to achieve accuracy in some systems, where speech can be converted into text, but the sources are limited such as some sources are based on grammar, some on vocabulary, and some on knowledge of speech, etc. To make a computer learn and improvise speech recognition, the "meaning of the speech" would be more helpful than anything else. To get "meaningful speech', one needs to apply knowledge resources to speech recognition. To achieve sources from knowledge, two things are essential in speech recognition which is: searching and matching. In these two essentials, knowledge sources such as syntax, sequence of words to be used in searching and can only be verified by matching the words with the context of the storyline.

Researchers have developed numerous speech recognition systems. Dragon has improved the most and achieved an accuracy of 95 percent, which is similar to humans. Instead of

using a keyboard and mouse, humans can use Dragon to convert speech to text [8, 13]. Human speech is mainly based on context, such as situation and conversation about the topic, etc. Nowadays, machines are good at predicting, which word comes next if they come to a cross partial sequence of words. For example, when human types "I am from XXX country", which is wrong. However, the machine is good at predicting the word after and corrects it automatically saying that "I am from XXX country". Presently, humans interact and communicate with machines more than they do with their fellow beings. For example, Alexa and Siri, have become an integral part of our daily lives, assisting us in our day-to-day activities, be it setting up our calendar, or providing a weather report, etc.

During the last four decades, a variety of speech recognition technologies have been proposed, demonstrated, and implemented using different algorithms. Machines can understand and identify a speaker through speech with the help of speech recognition systems. Verbalised words are digitised to make patterns and then compared with codes in the dictionary for identification.

The speech recognition technologies can be differentiated by the following considerations.

1. Does a machine need more speakers to train to be able to identify speech patterns ?
2. Can a machine recognise continuous speech or can it recognise only discrete words ?
3. Does the capability of a machine recognition system depend upon vocabulary? That is, can it identify a speaker with the help of limited vocabulary or does it require a larger range of vocabulary to do so ?

A variety of speech recognition systems are available in the market. Some of them are speaker-dependent and some of them are discrete. Humans have started using these speech recognition systems more, as compared to using a keyboard. Speech recognition systems use syllables as their basic unit. The limitations of syllables lie in factors such as homophones, where, groups of letters can have similar pronunciation, but quite different meanings (Homonyms), making recognition challenging. For example, ‘their’ and ‘there’, share a common group of letters, sound familiar (homophone), but the actual choice of which word to use, requires e.g. contextual information. Consequently, if a machine was to distinguish more details by understanding the sounds when similar words are pronounced by the same person, then it would have achieved better accuracy.

2.1.2 History & Use of Speech Recognition System

The concept of machine recognition of human speech came in the early 1920s. The first machine to recognise speech was named and manufactured in 1920 [16]. Later on, research

on speech technology was started at Bell Labs in 1926 [17, 18]. Researchers have been working on fundamental ideas of acoustic phonetics and some early attempts at speech recognition by machine were made in the 1950's [19, 20].

Later on, at Bell Laboratories, Davis, Biddulph and Balashek worked and developed a system that could recognise digits for a single speaker in 1952 [21–23]. Olson and Belar at RCA laboratories were able to recognise ten distinct syllables of a single speaker in 1956 [24, 25]. At, University College of England in 1959, Fry and Denes tried to build a system that could recognise four vowels and nine consonants based on phonemes. This system used a spectrum analyser and a pattern matcher to make decisions on recognition [26–28]. The phoneme recogniser allowed a sequence of phonemes in English to improve overall phoneme accuracy for words that have more than two phonemes. During the same period, Forgie was able to recognise 10 vowels embedded in a /b/-vowel/t/ were recognised [29, 30].

In the 1960s, Suzuki and Nakara of the Radio Research Lab in Tokyo, Japan developed hardware that could recognise a vowel. At this time, computers were not still good enough in terms of hardware. However, the Japanese system was able to build a vowel decision circuit by using a spectrum analyser and was able to recognise what vowel was spoken by a speaker [31, 32]. The second hardware phoneme recognise was built by Sakai and Doshita of Kyoto University in 1962, Japan. In 1963, again Japan developed the digit recogniser with the help of Nagata and researchers at NEC Laboratories. This was the initial attempt made for speech recognition at NEC and then led to a productive research program. One of the problems of speech recognition systems was variations of speech in time scale. To rectify this problem, three research projects were initiated towards the development of speech recognition. In 1960, Martin and his colleagues at RCA laboratories developed a system that could detect the start and end of the speech. At the same time, Vintsyuk suggested the use of Dynamic Time Warping (DTW) which was developed for connected word recognition. However, the concept of connected word recognition did not come to light until the 1980s.

The area of isolated word or discrete utterance recognition systems was developed by Velichko and Zagorukyo in Russia, Sakoe and Chiba in Japan, Itakura in the United States, and usable technology in the 1970's [33]. The Japanese research helped to determine how dynamic methods could be used in speech recognition and Russia and the United States helped the use of pattern recognition ideas in speech recognition. A large group of people at IBM, developed a speech recognition system using large vocabulary [34]. Over two decades, researchers studied three tasks which are Ner Raleigh language, the laser patent text language, and Tagore.

Researchers at AT&T Bell labs, conducted initial experiments to make a speech recognition system that was speaker-independent. To achieve this, researchers started collecting

a large dataset that had variations of different words from a range of speakers to be able to see different patterns among speakers. This research was continued for a decade and they developed the techniques for creating speaker-independent patterns. Then they finalised the project and it was funded by the Défense Advanced Research Projects Agencies (DARPA).

The speech recognition system could recognise the speech using a vocabulary of 1011 words in 1973 by CMU using a system called Harpy system. Then, the goal was the research was to develop a system that should be capable of recognising spoken words based on pattern matching of individual words [35]. Moshey J. Lasry developed a speech recognition system where he talked about spectrums of digits and letters, but the results were inaccurate. In the 1980s, speech research took off as a result of a shift in technology from template-based approaches to statistical modeling methods. The Hidden Markov Model approach could recognise thousands of words [36]. In 1990, Dragon launched a system called Dragon & Dictate which could recognise 100 words with 45 minutes of training time [37, 38]. Bell South developed a voice recognition system that produced information about what the speaker said through a telephone in 1996 [39]. The recognition system achieved 80 percent accuracy in 2001. After a decade, Google launched a speech search system that was built with 230 billion words from actual users, and after 2015, they released "Google Voice". Various technologies have been developed and released in the market, used by people in their daily lives. Speech recognition was proven and it achieved accuracy comparable to humans. However, these systems are good at recognising what has been said rather than identifying who is speaking [40, 41].

2.2 Basic Concepts of Voice Recognition

The air from a person's lungs passes through vocal cords to produce a human voice out of the mouth. That includes the lips, tongue, mouth, palate, etc. The following shows the human vocal cord production in detail as shown in Figure 2.2. The air comes from the lungs and then creates a flow through the larynx and pharynx. The larynx is considered as an energy provider for vocal folds to make fluctuations in the air pressure called sound waves and the volume of air determines an amplitude of a sound wave. These sound waves travel through & over the shape and position of a tongue, lips, palate and other human speech organs [42]. Every sound wave has several features, because of the changes in vibration of the vocal cords. The sound wave goes through the mouth and nasal cavities to produce speech as shown in Figure 2.2. Vocal folds create different types of the human voice, which are voiced speech, unvoiced speech (voiceless), and whisper. Humans use all these types to listen, understand and recognise a speaker.

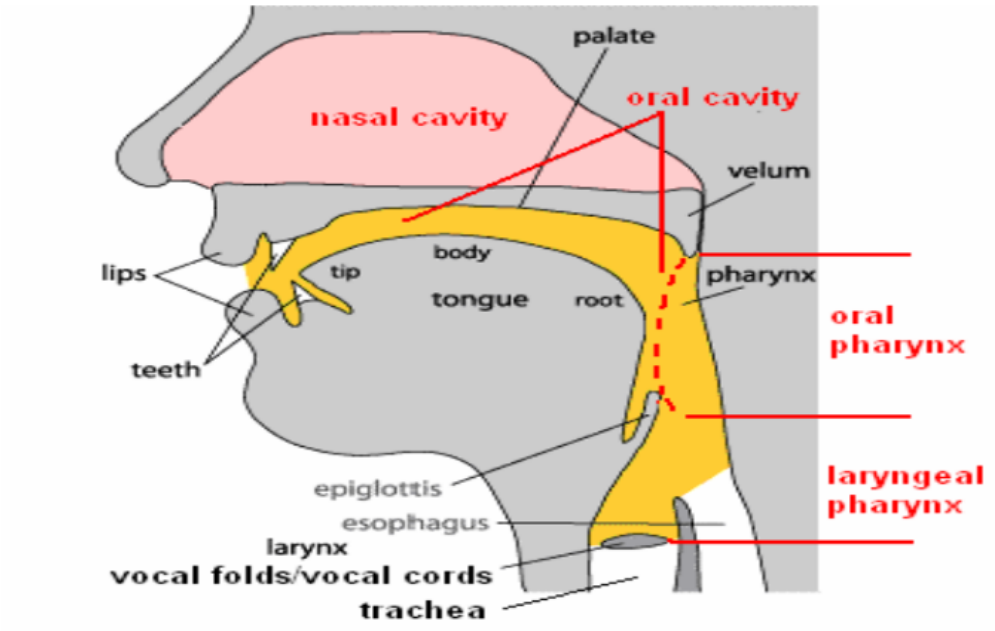

Fig. 2.2 Human Voice Production System
[31]

The vocal tract is one of the most important things in the human voice production system. Human speech conveys information in terms of pitch, which is the fundamental frequency. Female and male voices have different frequency ranges, and it varies in vocal tract length [43]. Normally, women have higher pitch when compared to men. However, it is possible that a person with a higher pitch can be a male, and for a person with a lower pitch to be a female. Based on vocal tract length, humans can predict a listener's body size as well.

Every individual speaker has particular uniqueness in their voice, which helps identify them. The uniqueness of a human voice not only depends on the vocal tract length and physical features but also depends on the speaker's ability to control organs in the vocal tract. However, it is not easy to change physical features, but it is possible with ageing. Physical features of a human voice include, vocal tract length, size of tongue and teeth, etc [44]. The analysis of human speech as shown in Figure 2.3

2.2.1 Speaker Identification and Speaker Verification

There are two types of voice recognition systems i.e.: speaker identification and speaker verification [45, 46]. Speaker identification systems can give two outputs which are: no identification claim; and identity claimed [47, 48]. The system identifies the best match

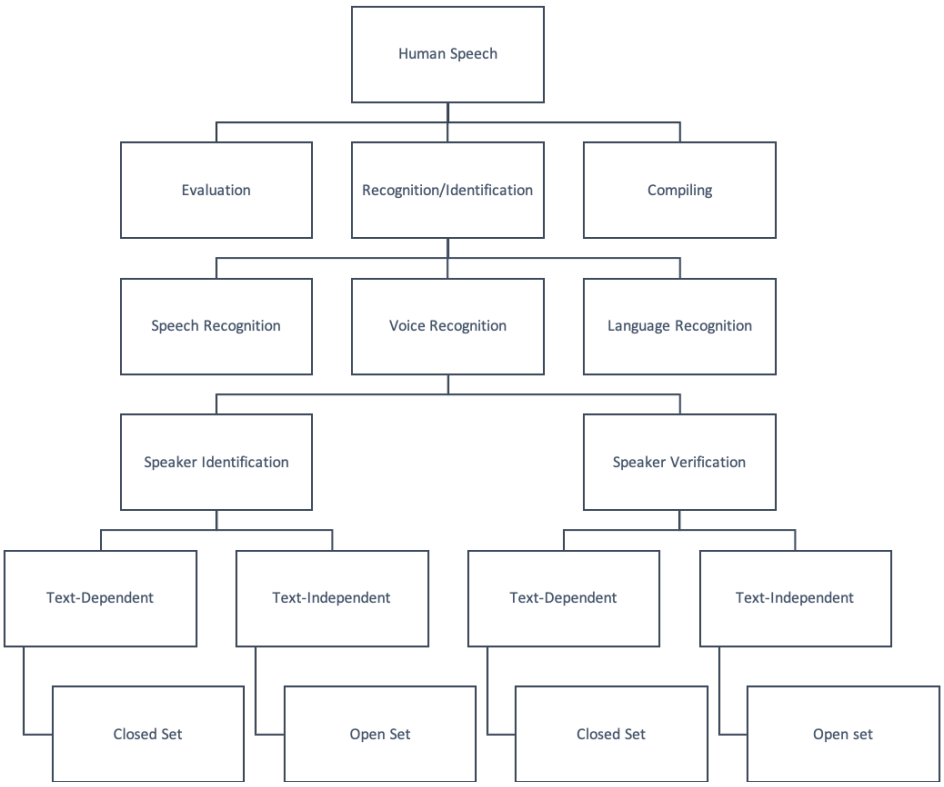

Fig. 2.3 Approach to Human Speech Analysis

when compared with a test sample, as shown in Figure 2.4. Speaker Verification involves two outputs, which are accepting or rejecting a speaker, and it distinguishes if the speaker voice matches with a voice already stored in the database [43, 49, 50]. The result is mainly dependent upon the probability of a voice match.

Speaker identification is considered to be a difficult task when compared with speaker verification [51, 52]. The reason behind this is that, as the number of speakers increases, the probability of making the wrong decision to identify a speaker also increases. On the flip side, speaker verification is easy, because systems will be having only two speakers for comparison at any stage as shown in Figure 2.5.

2.2.2 Open-Set and Closed-Set Identification

Speaker identification is further divided into closed-set and open-set identification. In closed-set, a speaker is identified from a set of already enrolled speakers. On the other hand, in the open-set identification, the speaker can be either be registered or the speaker may not be in the database, which means a test voice sample has not been registered in the past. A closed-set system is used to identify the best match to the test speech sample. Then, verification is used

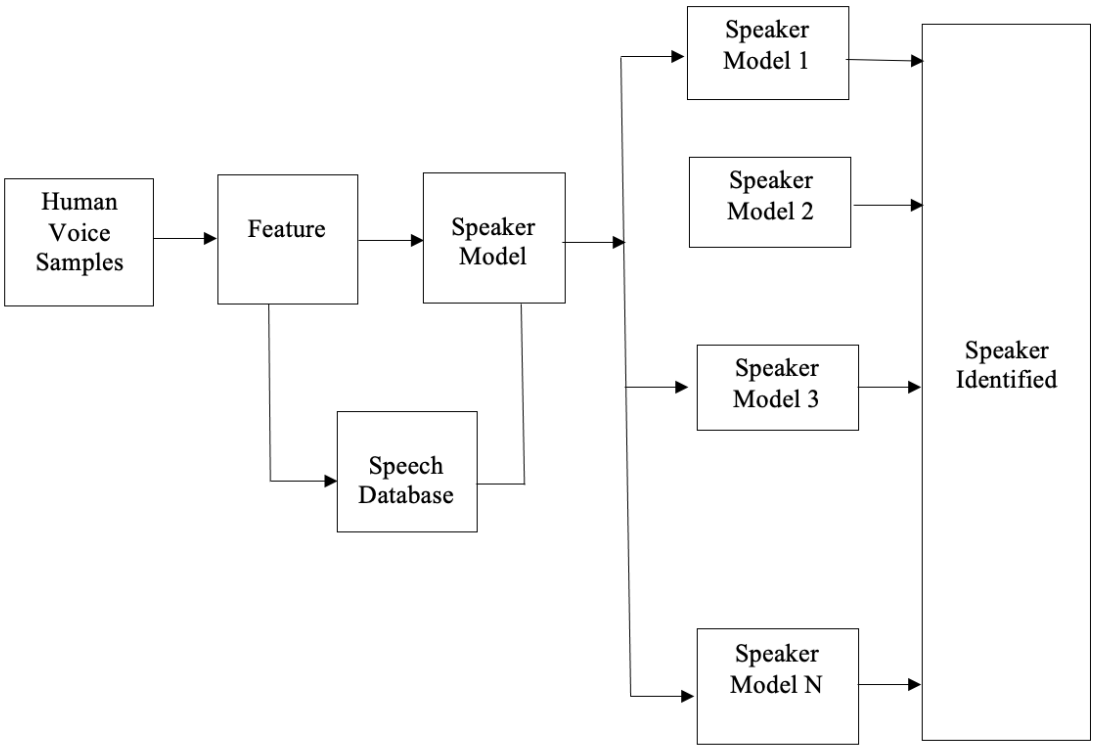

Fig. 2.4 Testing Phase of a Speaker Identification System

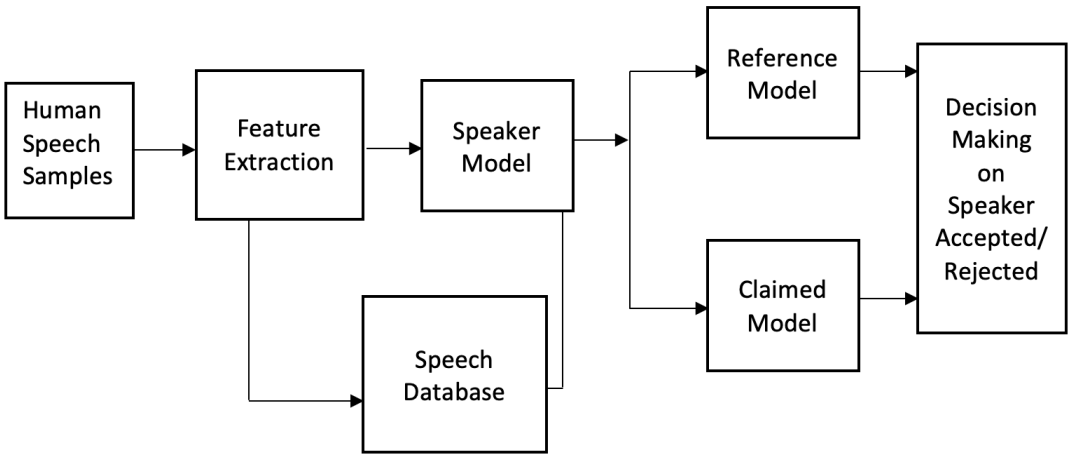

Fig. 2.5 Testing Phase of a Speaker Verification System

to compare the distance of the speaker with a chosen threshold to make a decision. At the end of the comparison, the system can identify a speaker or have no match as a result. The decision is made purely based on choosing the best matching voice sample from a database, despite the level of accuracy of the result. In the open-set identification, there should be a predetermined threshold so that the similarity degree between the unknown speaker and the best matching speaker is within the threshold level.

2.2.3 Text-dependent and Text-Independent Tasks

There are two modes of operation in speaker recognition which are text-dependent and text-independent [53]. In text-dependent speaker identification, a speaker is used to read/speak the same text or number for both the training and testing phase. During the recognition phase, the speaker is asked to read or speak the same text. Whereas, in text-dependent verification, speech samples used in training would be the same, but different for every verification task. A speaker is asked to read/speak words or digits randomly selected by a system and previously saved in the database during the testing phase. The advantage of using this system is, it will help eliminate any errors caused by knowing the speech sample beforehand.

In text-independent systems, the speaker does not need to speak/read the same words or numbers both in the training and testing phase [53, 54]. That means speech samples used during enrolment and testing are different. This type of system requires more training data in terms of speech samples and speakers need to talk for a longer time as well. In this case, enrolment can happen without speaker knowledge or permission [55, 56]. As a result, text-dependent recognition achieves more accuracy when compared to text-independent recognition.

2.3 Feature Extraction Of a Speech

Theoretically, it is possible to identify a speaker from a speech waveform. However, there is a large amount of variability in human speech because of several things. So, it is better to extract features that would be helpful for identification.

Feature analysis is a technique that achieves speaker-independent voice recognition. Feature analysis does not try to find an exact or the best match between input voice and a reference voice from a database. In this technique, the first step is to apply Fourier Transform on input voice to convert from the time domain to the frequency domain. The computer tries to find similar characteristics between the expected input and the digitised input voice. These characteristics will be present in every speaker, and so the system does not need to be trained

for every speaker. These kinds of systems are speaker-independent and characteristics include accents, pitch, volume, and speech rate. Speaker independent systems have proven to be ineffective in identifying a speaker. One of the hardest parts is to tell us what characteristics are unique to a particular speaker since, for example, a speaker fluent in multiple languages would use different types of accents and pronunciations.

Feature analysis is a technique that can aid speaker-independent voice recognition. Feature analysis does not try to find an exact or best match between input voice and a reference voice from a database. In this technique, the first step is to apply a Fourier Transform on the input voice to convert from the time domain to the frequency domain. The computer tries to find similar characteristics between the expected input and the digitised input voice [57].

2.4 Feature Matching Techniques for Speaker Identification

While many researchers aim to better understand how the brain works at the lowest level and how it provides for its learning functionalities, it may be that more suitable answers need to be searched in how the brain converts information into patterns, as it seems to be those patterns that lie at the basis of most, if not all, of our learned information. For example, if someone asks you to explain the structure of your home, you will first think about where to start from, kitchen or cellar, and from there you will work your way methodologically through the remainder. So even though all information is there, you will try to prioritise and then explain to your friend following a particular, most often logical, pattern.

The recent popularity of deep learning has raised the significance of using hierarchies within the models that lie at the basis of most artificial brain architectures. This is also in line with the human brain's multi-level hierarchical structure for processing information. However, while in many cases the underlying models are now becoming hierarchical, the feature sets used during learning are often fixed or have limited flexibility once learning has started. The Simplicity and Powerful (SP) theory is a method of learning in which features are combined in various ways depending on the requirements to allow for suitable multiple alignments to be made. This approach comes from bio-informatics where it is found in the context of e.g. DNA sequence alignments. To achieve these alignments, similarities are identified within each provided pattern during the learning phase, which tends to lead to overall data compression. Each unique pattern is then saved, and so when a new pattern is presented, SP theory can be used to check whether there is any similarity with any of the

already saved patterns and will continue to add new patterns to its learned information. The issue then becomes how this machine saves and retrieves information.

Generally, the human brain retrieves information from its "memory", which for the brain is a set of interconnected neurons. While neurons are quite fast in comparison to the transistors used in current computers, their functionality is quite different. For example, if you want to catch a ball, you need to estimate the trajectory of the ball to catch it, which happens automatically in the brain through a derivative pattern that aligns with previously learned patterns influenced by certain parameters, such as the estimated weight of the ball, the force of throwing and environmental conditions such as wind, etc. On the other hand, computers would need to calculate every step to ensure that a robot catches the same ball. An additional difference between computers and the brain lies in the fact that a computer has separate memory in the form of memory cards and hard drives, which is not stored automatically, while the brain seems to be one large pattern-focused memory that stores/adjust information continuously.

Speaker identification systems started in the late 1980s following the improvement of speech recognition systems. The improvements were made in feature extraction methods and classification methods in the early stages.

At the initial stages of speaker recognition systems, there were only text-dependent systems. Dynamic Time Warping (DTW) and template matching techniques were used. These techniques work only when the same text is spoken by an individual in both training and testing data. However, if the speaker changes her/his word at the testing stage, the system failed to identify a speaker.

2.4.1 Acoustic-Phonetic Approach

The acoustic-phonetic approach has been developed to recognise spoken words by using phonemes. This method had been used for more than 40 years. Phonemes are distinctive and characterised by a set of properties that occur in human speech, that can be changed into a speech signal over time. Every language has its phonemes and unique way of pronouncing. However, the English language has 44 phonemes that do not sound the same in all cases, i.e.: the same phoneme can sound different in different words. The Phonetical approach was the earliest method of recognising words and then later used to identify a language spoken by a speaker. There are 3 steps that were followed/involved in this approach as shown in the Figure 2.6.

Feature Extraction:

Spectral analysis was applied to a speech signal to be able to extract features from a speech signal.

Segmentation and Labelling of Phonemes:

Each phoneme was labelled with segmentation of speech signal.

Recognition of words:

Combination of phonemes labels helped to recognise words.

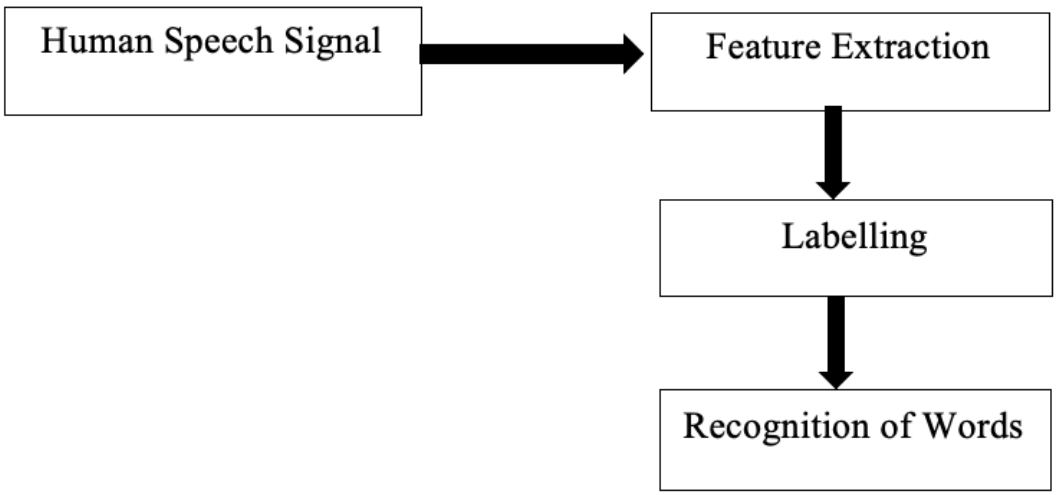

Fig. 2.6 Block Diagram of the Acoustic Phonetic Approach for Speech Recognition

2.4.2 Pattern Recognition Approach

Pattern recognition is a mathematical framework and has been developed over the past two decades. This can be applied to a sound that is smaller than a word or a sentence [56]. There are two steps in this approach, which are: pattern training, and pattern comparison. A speech template was developed in the training phase and then two unknown speech samples would be compared in the comparison phase, with patterns which were learned in the training phase [58, 59], as shown in the Figure 2.7

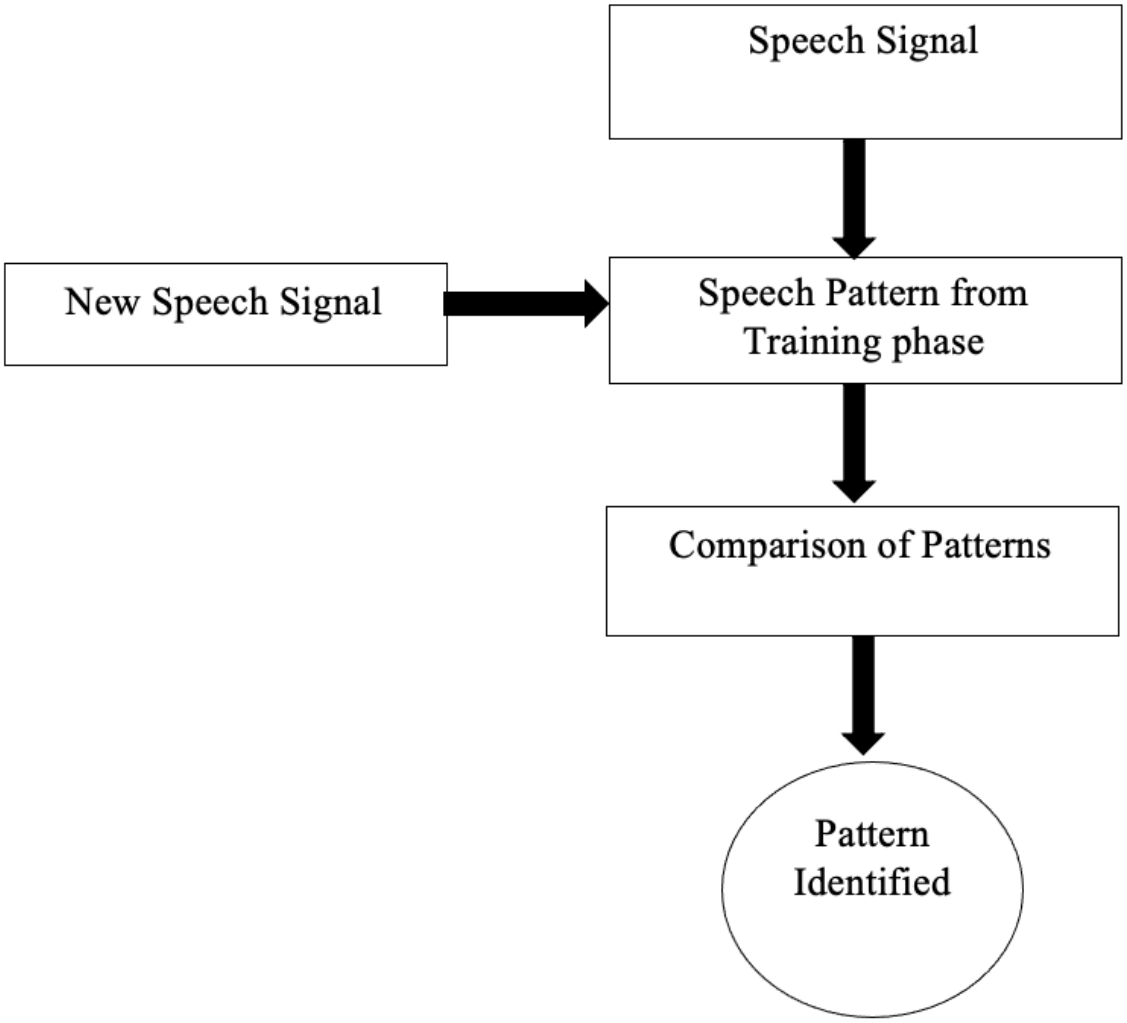

Fig. 2.7 Block Diagram of Pattern Recognition Approach

2.4.3 Template Matching Approach

Template matching (TM) techniques are based on an algorithm that uses words to recognise a speaker. In TM, the speaker was asked to read a word or sentence and which was then digitised and stored in a database as a reference template. During the test phase, the computer attempted to compare the input voice with a reference from the database. The computer then tried to find the best match between the two reference templates, as shown in Figure 2.8. These systems are known to be speaker-dependent and 98 percent accuracy has been achieved. However, the expression of words might differ during the testing as compared

to the training due to factors such as tiredness or stress. The drawback of using the TM approach is, the pronunciation may change because of the previous phoneme. A Speaker's voice may change over time and affects such as speaking rate.

As the technology for speaker identification evolved, the focus has become for systems to be text-independent thus there was no place for template matching techniques of the early 2000s.

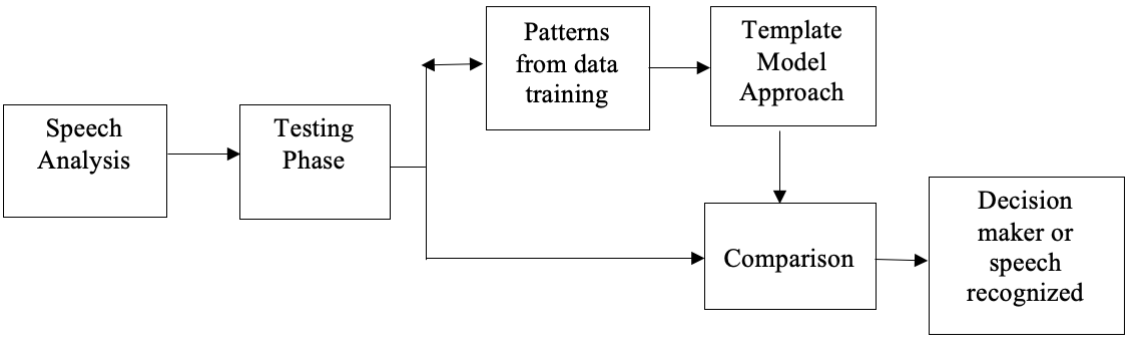

Fig. 2.8 Block Diagram of Template Matching Approach

2.4.4 Vector Quantization Approach

Vector Quantization (VQ) is used to reduce the data required for a speech recognition system. It is a technique of dividing a large number of data set points (which are called vectors in this approach) into smaller groups [60, 61]. Each group is called a cluster and can be represented by its centroid point. The collection of these points or code-words is called a code-book. Each codebook contains several vectors, which are stored in an individual speaker database. In this approach, the distance would be measured between the training frames for two speakers as shown in Figure 2.9.

2.4.5 Dynamic Time Warping

Dynamic Time Warping (DTW) is a method to determine the similarity between two feature vectors, which varies in time or speed. DTW would help the machine to find out the best match between the two patterns as shown in Figure 2.10. DTW has been applied in audio, video, etc. DTW has achieved better accuracy in word recognition. DTW was originally developed for speech recognition, but later on, researchers started using it for speaker identification [62]. However, DTW is only good for a small number of speakers (templates).

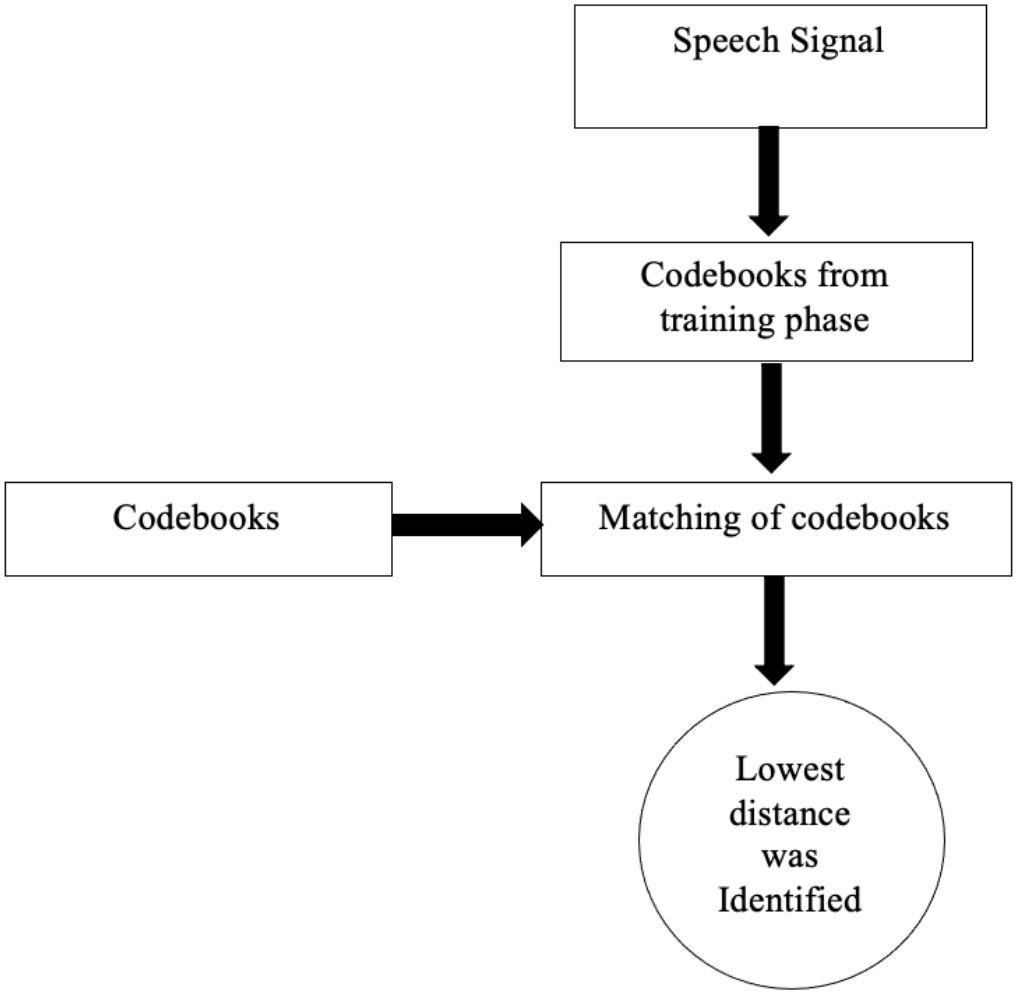

Fig. 2.9 Block Diagram of Vector Quantization Approach

Another drawback of DTW is that words need to be recognised first before the identification of a speaker can proceed. One positive is that identifying a speaker is language-independent.

2.4.6 Statistical Based Approach

Variation within human speech depends on several reasons such as a combination of different sounds, speaker variability, etc. This type of approach depends on the characteristics of the input. This approach has been proven to be the best probabilistic model for speech recognition. Hidden Markov Model (HMM) is the principal technique for probabilistic modeling and it is efficient for speech recognition. HMM, the model is a technique where

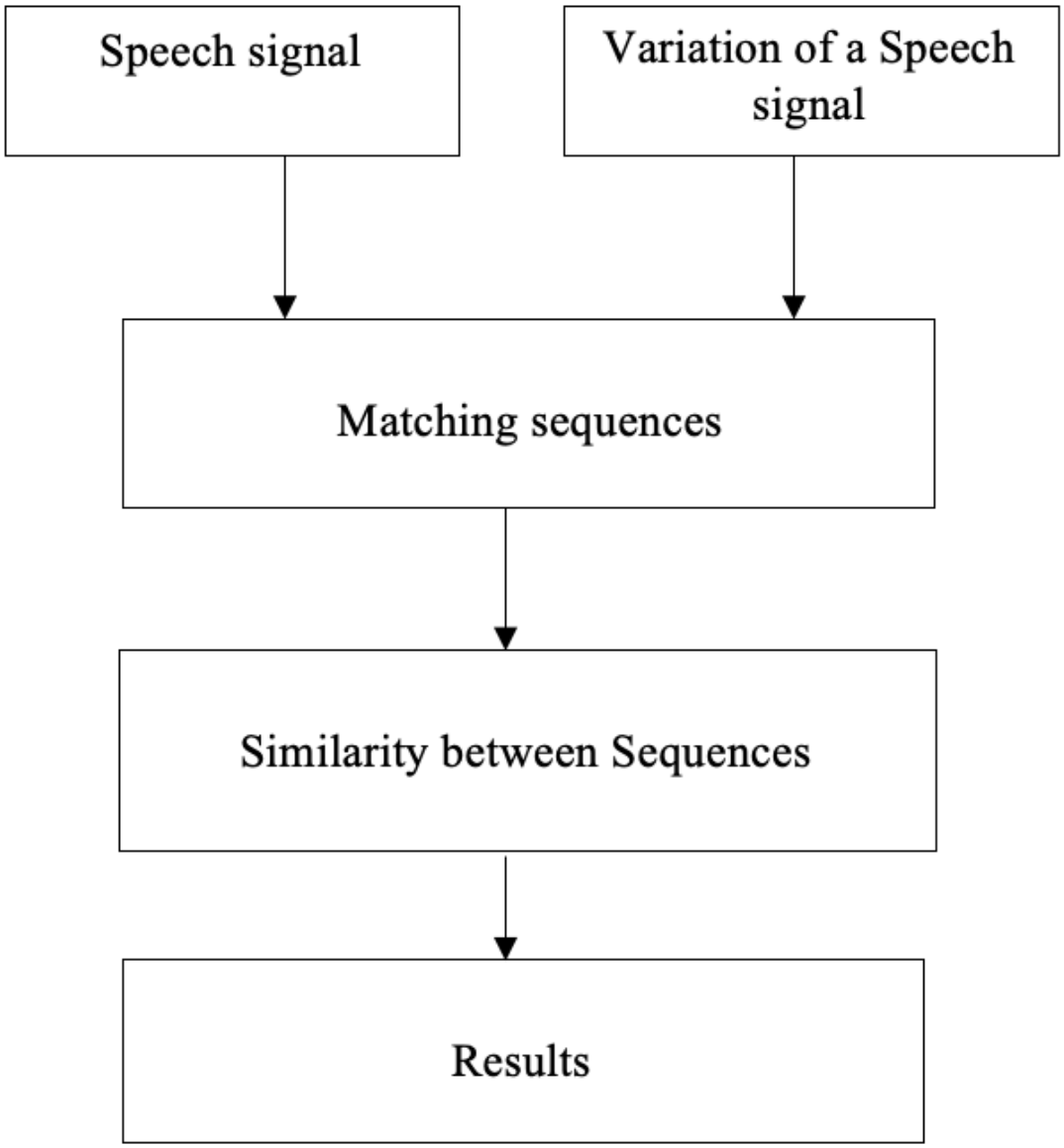

Fig. 2.10 Block Diagram of Dynamic Time Warping

speech is generated from several states for each HMM model. Each model has different output distribution and the HMM model is a combination of words and each word is trained individually [63–65], as shown in the Figure 2.11

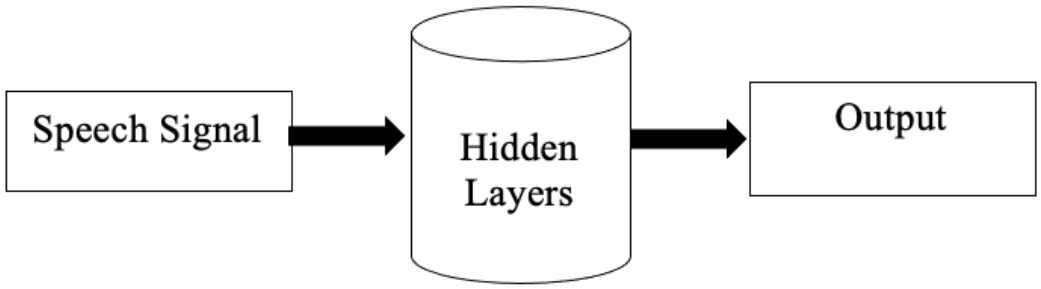

Fig. 2.11 Block Diagram of Hidden Markov Model

2.4.7 Artificial Neural Network Based Approach

In this approach, where intelligence' is involved to analyse and visualise the speech signal to extract features. This approach depends on a person who coordinates and designs it for recognition. This approach is a knowledge-based system, where knowledge is extracted from experts of the contribution of a person who designs it [66].

This type of approach network included several neurons. Each neuron computer's nonlinear weight of inputs and broadcast results to the outgoing units, training sets are used for assigning pattern of values to input and output neurons, training set determines the weight of strength of each pattern as shown in Figure 2.12.

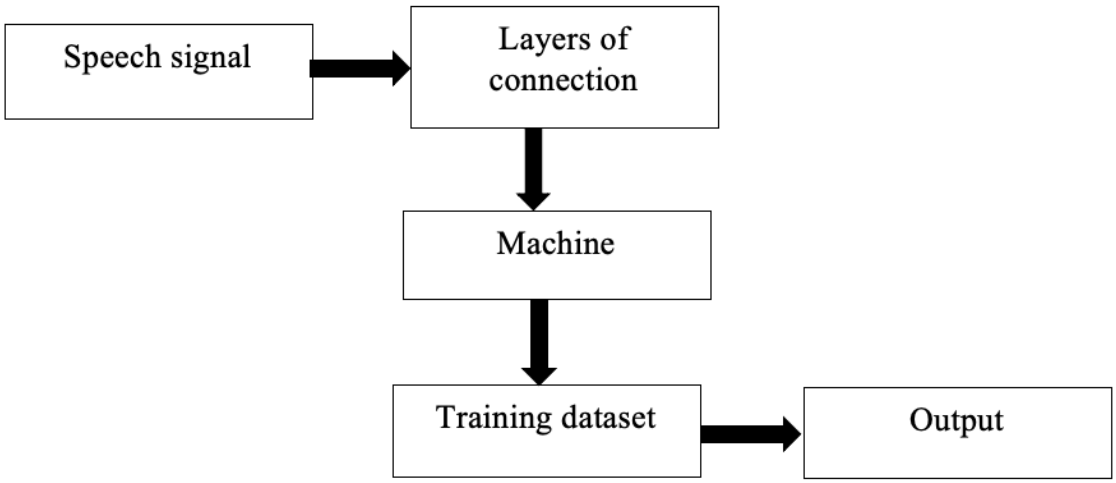

Fig. 2.12 Block Diagram of Artificial Neural Network Based Approach

2.4.8 Comparative Study of Approaches

The advantages and disadvantages of approaches are summarised in the Table 2.1.

Table 2.1 Comparative Study of Speech Recognition System Approaches

Number	Techniques	words	Limitations
1	Acoustic Phonetic Recognition	System takes less processing time connected words	System takes longer time to execute each word
2	Dynamic Time wrapping	Easy to find match between two sequences	Difficult to find a match if there is variation in speech System takes longer time for computational work
3	Pattern Recognition Approach	Pattern matching is easy and quick in between two words	System cannot recognise if there a variation in a pattern Its applicable only for word to word match System needs to more time to process
4	Vector Quantization Approach	Useful to reduce data	It is text dependent
5	Template Base Approach	It better for small vocabulary	Not applicable for larger Vocabulary Difficult to find similar patterns
6	Artificial Neural Network Approach	Useful for larger vocabulary and it can train larger data as well Easy to implement and can change the size of training data easily Achieve recognition rates accurately	Required larger amount of data for training System need more computation power

2.5 Factors Affected in Speaker Recognition System

The performance of the current speaker recognition system is affected by several factors. The quality of the voice is one of the factors on which the speaker recognition system is mostly dependent. If the quality of the human voice recording is not good/clear enough, it would be very difficult to identify a speaker [67]. For example, humans take a longer time than supposed to, to identify a speaker if the speaker's voice is not clear enough to hear.

The other factor is noise. The background noise is one of the most aspects of speaker recognition where systems accuracy gets affected. Clean samples help systems get better accuracy than noisy samples.

2.6 Research Gap

Human voice or speech signals contain information about an individual such as speaker identity, speaker emotion, speaker message content, language, etc. Speaker identification is a technique for recognizing an individual by her/his voice. Research in this area is continuing and various developments have been done, but still, accuracy needs to be improved. Researchers have been trying to increase the accuracy of Speaker Recognition systems. An accent is one of the features that can help to identify a speaker in only one language. However,

it is one of the limitations of SR is because different people can speak with different accents and that it can be challenging for a machine to recognise the speaker as such.

A speaker recognition system needs to learn voice patterns that should be able to identify a person. Since voice has the characteristics of both physical and behavioural features, feature extraction is a method of converting speech into features that contains the characteristic information of a speaker. The current features that have been used in the speaker recognition systems are language-dependent and accuracy is affected when they speak in other languages. Therefore, there is a research gap in feature extraction approaches for automatic speaker recognition systems. The proposed method for the development of an accurate speaker recognition system is extracting features from speech signals which should be language-independent, applicable to both text-dependent and text-independent speaker recognition systems.

Identifying language-independent features of a voice is key to investigating the unique characteristics of a speaker's voice. To be able to identify the language-independent parameters, one should understand firstly how human speech works [12, 68]. There are two levels in human speech: primary level (low level), speech conveys a message through words. A person listens to her/his conversation, which then helps to analyse their accent to be able to identify a person. One can design a machine to learn a person's accent to identify a speaker. However, classification boundaries learned by a system for a particular accent do not work for other accents. The second level, speech carries specific information about a speaker for recognition by extracting features from voice characteristics such as frequency, volume, and timbre.

Humans can recognise a speaker by just listening to a few words such as: "How are you?", "Hello" and their response to identifying a speaker is a few seconds. Sometimes, humans can predict a speaker's age, gender, and emotion, just by listening to their voice. The following questions have been raised and answered in the following chapters.

1. How can humans learn, understand, remember and then recognise?
2. How long does humans take to identify familiar and unfamiliar voices?
3. Do humans need to be familiar and/or understand the language to identify a speaker?
4. What are the parameters that would help to identify a speaker?
5. Do phonemes have impact on identification?
6. How much data do we need to recognise a person?
7. Where can we implement speaker identification technology in real world?

Chapter 3

Identification of a Speaker: Familiar and Unfamiliar Voices

3.1 Introduction

Learning is a necessity that helps in day-to-day life and also prepares us for a better future. For a person, learning is the most important process to acquire knowledge and improve intelligence [69]. It is also the main feature of machine learning, which attempts to build on the learning principle of the human brain and to develop computer intelligence. Machine learning and human learning have several basic similarities, but the mechanisms of machine learning can still be improved substantially. For instance, people can learn from very limited amounts of data when compared with machines and are very adept at inferring patterns in data or completing missing data. Currently, most machine learning algorithms have been inspired by certain mechanisms of human learning [70, 71].

In today's world, machines are continuously being developed to make human life easier. people are often disappointed when machines do not perform the functions, that humans expect them to do [72]. This is one of the reasons why machines need to be advanced, smarter, and user-friendlier. Some people believe that humans should make them more like people, which involves allowing them to learn and respond like humans do [73, 74]. However firstly, one needs to understand how a human being thinks and how their brain works. A simple example is the working of a modern computer; a computer takes an input and produces an output, however, a human brain is much more complicated and complex, including the process of creating and storing memories, since there are still unknown aspects about its actual mechanism of action & even how memories are created and stored [75].

A human brain is built up of neurons, which are combined into a network that can interpret information received from the environment. Neurons have a structure called synapses, which conduct electrical impulses and chemical signals from one neuron to another. These synapses are responsible for a brain's potential to think and sustain its consciousness [76]. Humans are good learners, they can learn by themselves own self, e.g. from their own experiences. The human neural network system as shown in Figure 3.1.

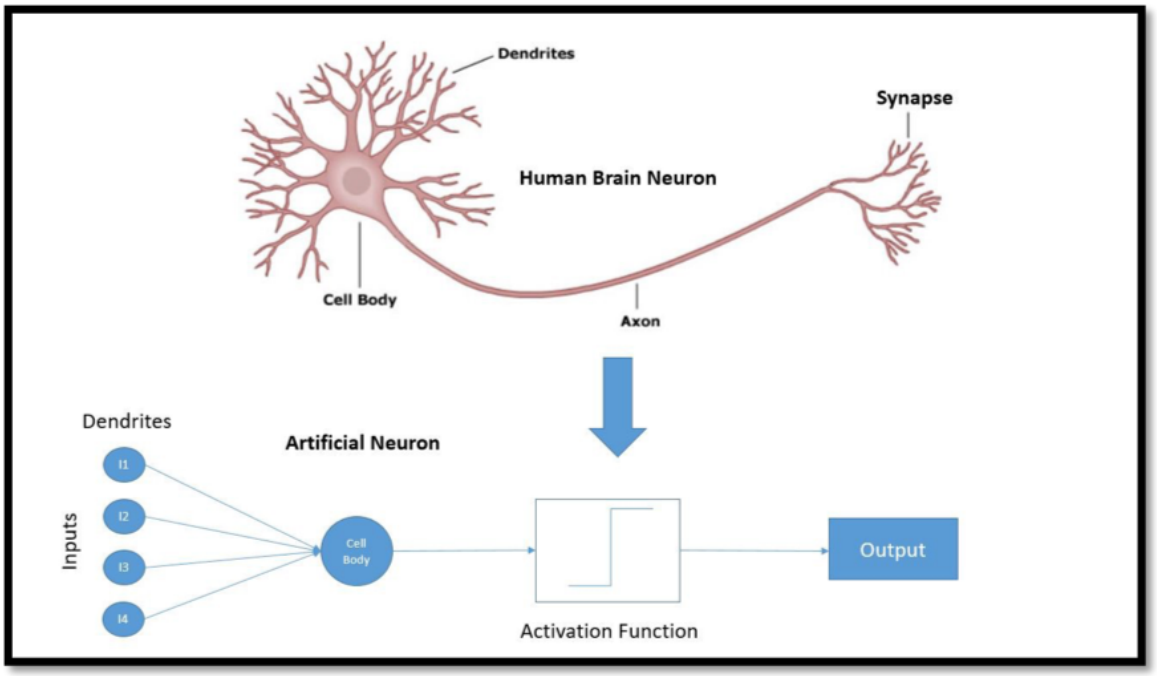

Fig. 3.1 Human Ways of Learning
[77]

Currently, learning is one of the major differences between machines and humans. Improvement in the learning ability of a machine can produce functions and tasks, similar to humans. Developing a better understanding of human learning should help achieve enhanced machine learning. Humans have features that allow them to use all their senses to improve their knowledge. They not only sense the environment, such as light, touch, sound, but they also sense and feel emotions, such as anger, hunger, and tiredness. People also have limitations to their abilities though. For example, they have a hearing range from 20Hz to 20 kHz while dogs have a hearing range of 40Hz to 60 kHz. This chapter will explore the ability of humans to learn, remember and identify a speaker who is familiar and unfamiliar, based on their voice.

Every person has a unique and different voice when compared to other people, which helps us identify a speaker based on their voice. Some important questions here are, how to do humans:

1. Recognise and comprehend a particular voice and correlate it to a specific person?
2. Remember the voice of a person they meet after a long time?
3. Differentiate between voices of people they meet on a regular basis?

3.2 Methodology

To find out how the human brain processes and recognizes different voices, two experiments were carried out with the help of both male and female participants and a survey was conducted based on the results obtained, to conclude.

In the first experiment, famous movie artist's (English) voices were downloaded from YouTube. Participants were asked to listen to the audio clip and asked if participants can recognise them or not. The overall view of the first experiment is shown in Figure 3.2

In the second experiment, participants were requested to read a few sentences in English, at different distances while keeping the microphone in one place.

3.3 Experiment 1: How People Recognise Voices

The first experiment is divided into 2 parts; taking into consideration that the time taken to identify a voice is recorded and compared, the first part of the experiment is based on the participant's familiarity with the voice (of a movie artist), and the second part of the experiment is based on the participant's familiarity to the language being spoken. There were 100 participants. All participants were over 18 and the range of age lies between 18 to 50 years. 35 Participants lived in the UK and 25 participants lived in India, but English is not their native language. 50 Participants lived in the UK and had English as their mother tongue.

3.3.1 Identification of a Speaker: Familiar and Unfamiliar Voices in Known Languages

Before the actual test starts, participants were asked to listen to the audio files from YouTube and ensure whether the voices were familiar to them or not. The reason for doing this was to

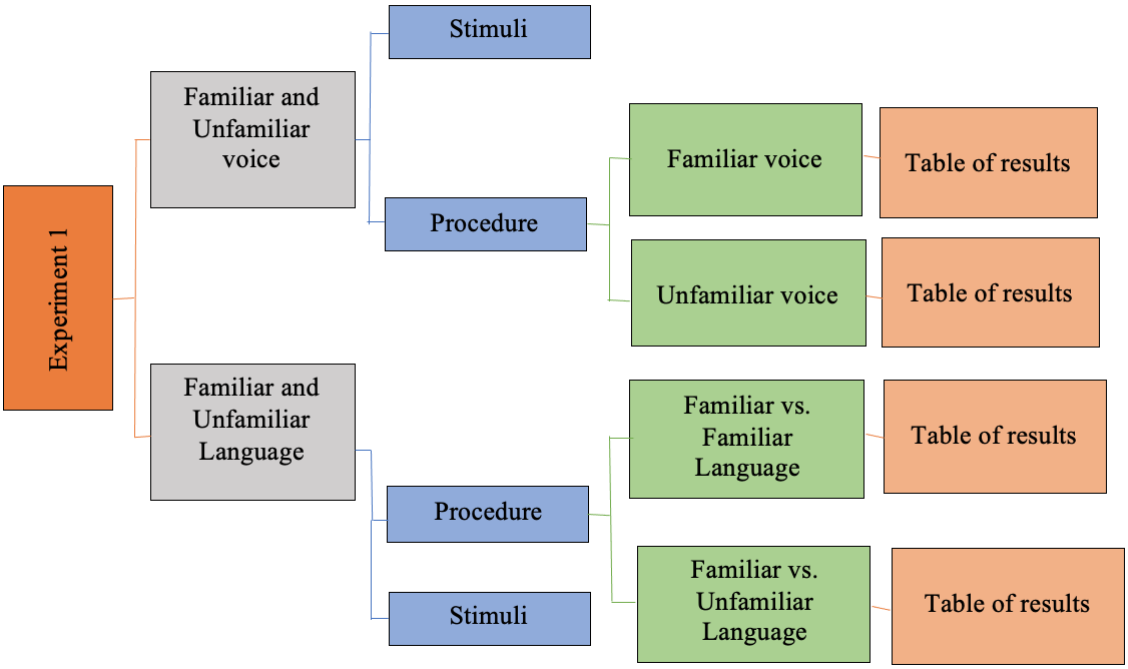

Fig. 3.2 Overview of Experiment 1

compare how humans can recognise a person's voice with which they are already familiar or unfamiliar with.

This experiment was performed in two parts. The first part of the experiment was used to derive and analyse data on familiar voices. Participants were requested to listen to a familiar movie artist's voice through YouTube recordings, which was the training data set. Then they were asked to listen to a different recording of the same movie artist, and identify if they were the same movie artist, or not.

The second part of the experiment was to analyse the data on unfamiliar voices. Participants were asked to listen to unfamiliar movie artists' voices from YouTube recordings and memorise the speaker. Once the Participants had listened to the recording, they were able to recognize whether the movie artist was female or male. Subsequently, they were asked to listen to a different recording of the same movie artist, and identify if they were the same movie artist, or not, as shown in Figure 3.3

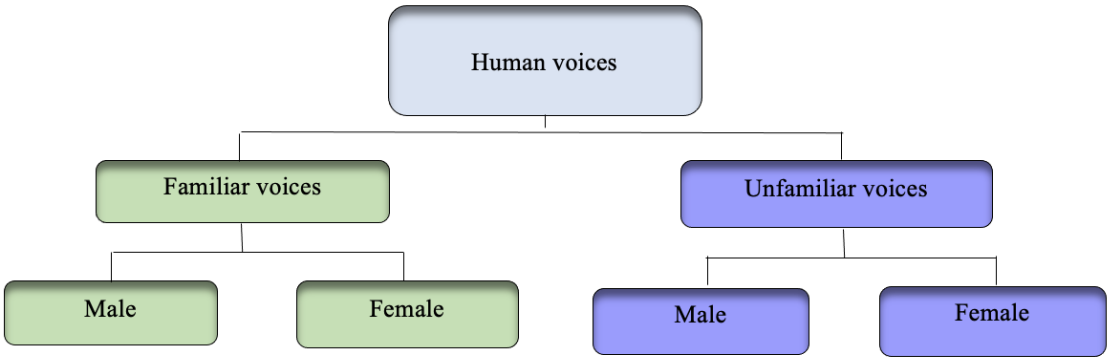

Fig. 3.3 Identification of Familiar and Unfamiliar Voice

Stimuli

Ten male and ten female movie artists' voice samples were downloaded from YouTube. The voices were recorded from multiple channels such as iPhone and Mac-Book. Voice samples were divided into a small window size which is 10, 20, 30, 40 sec, and so on.

Procedure

Ten audio files were downloaded from YouTube for each movie artist and the Audio files were each 60 seconds long. The files contained recordings of movie artists are: Trevor Howard, Tommy Cooper, Tanner Cruz, James Earl Jones, Windsor Davies, Billie Piper, Julie Dawn Cole, Morgan Freeman, Tim Curry, Kristen Schaal, Fran Drescher, Holly Hunter, Scralett Johansson, Mariska Hargitay, James Wood, Jessica Lange, Emma Stone, Kathleen Turner, Lauren Bacall, Emily Blunt, Carey Mulligan, Helen Mirren, Vera Farmiga, Catherine Zeta-Jones, Rikcy Gervais, Nina Dorbev, Sara Wayne Callies and Victoria Pedretti and Elizabeth Lail.

Participants listened to the files and were asked the following questions:

1. Was the person male or female?
2. Did you recognise the person?
3. Can you recall the persons image when you hear their voice?

Recognition of a Familiar voice

Fifty participants were asked to listen to an audio clip of an artist for example Julie Dawn Cole, Morgan Freeman, etc., they were familiar with, for 60 secs. Then they were asked to listen to another audio clip of the same artist, and the time taken in seconds for the participant to recognize the voice was measured in Table A.1.

Table 3.1 The Time Taken by Participants to Recognise a Familiar Voice from a Second Audio Clip

Participant	**Recognition of a Familiar Voice Audio clip of 60 seconds**	
	Measure Time in seconds	
	Female	**Male**
1	20	10
2	20	10
3	10	10
4	20	10
5	10	20
6	20	20
7	20	10
8	10	10
9	20	10
10	20	30
11	20	10
12	10	20
13	20	20
14	20	10
15	20	20
16	20	10
17	20	20
18	20	10
19	10	10
20	30	20
21	20	30
22	30	20
23	20	20

Table 3.1 continued from previous page

24	20	10
25	10	20
26	20	30
27	10	20
28	20	20
29	10	20
30	20	10
31	10	30
32	20	10
33	10	10
34	20	10
35	20	30
36	20	10
37	20	10
38	20	10
39	20	20
40	10	10
41	10	10
42	10	20
43	10	10
44	20	10
45	10	20
46	20	10
47	30	30
48	10	30
49	20	10
50	20	10

According to Table A.1, 65 % of the participants had taken 20 seconds, 32 % had taken 10 seconds and 3 % had taken 30 seconds to identify a female movie artist where they are already familiar with. On the flip side, 37 % of the participants had taken 20 seconds, 51 % had taken 10 seconds and 12 % had taken 30 seconds to identify male movie artists.

On average, to identify a female movie artist, a participant took 17.1 seconds, and the time is taken by all participants to identify the artist ranged between 10 to 30 seconds. Whereas

on average to identify a male movie artist, a participant took 16.1 seconds, and the time is taken by all participants, to identify the artist ranged between 10 to 30 seconds.

Recognition of an Unfamiliar voice

Next, the participants were asked to listen to an audio clip of an artist they were unfamiliar with, for 60 secs. Then they were asked to listen to another audio clip of the same artist, and the time taken in seconds for the participant to recognize the voice was measured.

Table 3.2 The Time Taken by Participants to Recognise an Unfamiliar Voice from a Second Audio Clip

Participant	**Recognition of a Familiar Voice Audio clip of 60 seconds**	
	Measure Time in seconds	
	Female	**Male**
1	40	60
2	100	50
3	40	50
4	30	40
5	20	50
6	40	20
7	30	30
8	50	20
9	60	30
10	30	20
11	100	70
12	50	40
13	30	20
14	40	40
15	50	60
16	40	20
17	120	50
18	40	20
19	100	50
20	80	40
21	50	20

22	20	20
23	10	30
24	40	30
25	30	50
26	30	10
27	40	30
28	20	20
29	10	40
30	60	20
31	50	40
32	40	20
33	50	60
34	30	30
35	10	20
36	50	20
37	50	60
38	50	20
39	30	20
40	30	50
41	40	30
42	50	60
43	20	20
44	20	30
45	20	10
46	10	20
47	30	30
48	40	50
49	40	20
50	70	60

According to Table 3.2, On average, participants have taken 37.7 seconds, time range lies between 10 to 120 seconds and 33.2 seconds, range 10 to 70 seconds to identify an unfamiliar voice of movie artist female and male respectively. 24 % of the participants have taken 40 seconds, 20 % of them have taken 30 seconds, 8 % 10seconds, 18 % 50, 3 % 60, 03 % 100, 02 % 80, 1 % 70 to 100 seconds to identify a female movie artists. On the other hand, 26 %

20, 23 % 30, 20 % 40, 09 % 10, 11 % 50, 10 % 60 and 1 % of the participants have taken 70 seconds to identify a male artists respectively, which they are unfamiliar with.

3.3.2 Identification of a Speaker: Familiar and Unfamiliar Voices in Unknown Languages

The purpose of this experiment was to observe, how much data and time people need to recognise a person both in familiar and unfamiliar languages?

This experiment was two-fold; first, participants were asked to listen to YouTube recordings of a movie artist who spoke in a language familiar to the participant, as a training data set. Then they were asked to listen to a different recording of the same artist in the same language and time taken to identify if they were the same movie artist or not, is measured. In this experiment, a hundred candidates have participated. All candidates were over 18 and the range of candidate's ages lies between 18 to 50 years old. 40 candidates lived in India and 60 candidates lived in the UK, but their mother tongue is not English.

In the second part of the experiment, time is taken to recognise an unfamiliar language was measured. Participants were initially asked to listen to movie artist's voices speaking in languages familiar to the participant, as training data. Then they were asked to listen to an unfamiliar language from the same movie artist and the time take for them to identify whether it is the same speaker or not, is measured. In this experiment, there were 100 people. All participants were over 18 and the range of people's age lies between 18 to 50 years old. 30 people lived in India and 40 people lived in the UK, but their mother tongue is not English. 30 people lived in the UK and had English as their mother tongue.

Stimuli

Both male and female movie artist's voice samples were downloaded from YouTube. Ten male and ten female movie artist voices were recorded from multiple channels such as iPhone and MacBook. Voice samples were divided into small window sizes of 10, 20, 30, 40 sec, and so on.

Procedure

Ten audio files were downloaded from YouTube for each movie artist. The Audio files were 60 seconds long. The files contained recordings of the movie starts such as Kamal Hassan, Rajni Kanth, Sai Pallavi, Samantha, Raj Shekhar, SP Bala Subramanyam, Chinamayi, Dhanush, Vijay Devarakonda, Surya, Srinivas Murthy, Vikram, Hrithik, Arijit, Deepa Venkat, Devi, Katrina Kaif, Naziya, Rashmika Mandanna.

a) **Familiar Person vs Familiar Language**

In the first part of the experiment, candidates were asked to listen to one movie artist speaking in 2 different languages that the candidates are familiar with; the first language is the training data and the second language is the testing data.

The first participant listens to the training data for 60 secs and then, she/he is asked to listen to the testing data. Simultaneously, the time taken for the participant to recognise whether it is the same artist in the testing data or not is measured and noted in the Table 3.3.

Table 3.3 Time Taken to Identify a Speaker Whose Language is Familiar

Candidate	**Familiar Language Audio clip of 60 seconds**	
	Time taken to recognise a speaker in an unfamiliar language	
	Female	**Male**
1	40	50
2	40	30
3	50	40
4	60	40
5	40	30
6	30	30
7	60	50
8	40	40
9	20	20
10	10	30
11	10	30
12	40	40
13	30	20
14	20	10
15	10	20
16	10	30
17	10	40
18	40	50
19	30	20
20	60	50
21	40	30
22	30	30

Table 3.3 continued from previous page

23	10	30
24	20	40
25	40	30
26	30	20
27	40	50
28	10	30
29	50	60
30	60	40
31	70	30
32	10	40
33	50	30
34	40	20
35	10	30
36	50	40
37	40	20
38	30	10
39	50	30
40	60	30
41	40	20
42	30	10
43	60	30
44	50	60
45	40	20
46	20	40
47	10	30
48	10	30
49	40	10
50	30	10

According to Table 3.3, for an artist whose language the candidates are familiar with, on average a candidate has taken 34.3 seconds to recognise a female artist, and for all these candidates the range lies between 10 to 80 seconds and on an average, a candidate had taken 29.1 seconds to identify a male artist and the range lies between 10 to 70 seconds, whose language is familiar with. 24 % of people have taken 40 seconds, 21 % 10 seconds, 19 % 30 seconds, 10 % 20 and 60 seconds, 03 % and 1 % have taken 70 and 80 seconds respectively

to identify a female movie artist. On the other side, 28 % of people have taken 30 seconds, 20 % 10sec, 20 % 10, 19 % 40 sec, 8 % 50 sec, 4 % 60 and 1 % have taken 70 seconds to identify a male movie artists.

b) **Familiar Person vs Unfamiliar Language**

In the second part of the experiment, participants were asked to listen to two voice recordings in two different languages; familiar languages were used for training data and an unfamiliar language was used for testing data, both from the same movie artist. It was ensured beforehand that all participants were unfamiliar with the language spoken in the testing data.

Each participant listened to the training data for 60 sec, then they were asked to listen to a new recording. Simultaneously, the time taken for the person to recognise whether it was the same artist in the testing data or not, was measured and noted in the Table 3.4. Surprisingly, even though participants did not have any prior knowledge about the language used in testing, all of them were still able to recognise the speaker.

Table 3.4 Time Taken to Identify a Speaker Whose Language is Unfamiliar

Participant	**Familiar Language Audio clip of 60 seconds**	
	Time taken to recognise a speaker in an unfamiliar language	
	Female	**Male**
1	100	80
2	120	90
3	120	170
4	110	80
5	20	60
6	130	70
7	120	100
8	60	70
9	50	80
10	50	60
11	40	80
12	120	170
13	130	100
14	120	80
15	140	90

Table 3.4 continued from previous page

16	80	70
17	70	40
18	120	90
19	90	120
20	130	80
21	120	170
22	60	130
23	170	100
24	40	70
25	40	100
26	20	30
27	40	30
28	20	20
29	40	50
30	50	50
31	110	70
32	70	40
33	90	70
34	130	100
35	50	30
36	60	30
37	70	100
38	90	120
39	100	100
40	60	50
41	70	40
42	80	80
43	60	70
44	50	90
45	100	110
46	130	120
47	40	30
48	50	30
49	20	10

Table 3.4 continued from previous page

50	20	10

Table 3.4 shows the time takes to identify an artist whose language is unfamiliar to the participant. According to Table 3.4, the range for the time taken to identify a female artist lies between 10 to 70 seconds and on average it takes a participant 71.3 seconds to identify the female artist. The range for the time taken to identify a male artist lies between 10 to 70 seconds and on average it takes a participant 71.7 seconds to identify the male artist.

3.4 Experiment 2: Analysis of Variations of Distance and Volume of a Speaker

The second experiment was conducted by asking participants to read a given script, at different positions while keeping the microphone at one place, to observe the volume and time taken to identify a speaker. This experiment aimed to find out whether the volume of a person speaking into recording equipment affects the time taken to identify the speaker.

3.4.1 Equipment

Table 3.5 Experimental Conditions

Language	English and speaker's familiar Language
Recording Equipment	Audacity, Scarlett 2i2 studio, Anechoic Chamber
Operating System	MacBook Pro
Programming Language	Python
Sampling rate	44100
Headset or Headphone	Participant Choice

The selected programming language was Python, free to use and widely compatible on any of the major operating systems such as Windows, iOS, etc. The initial implemented code was evaluated and compared with other programming tools to check whether Python was providing the correct results or not. Participants were asked to read the following sentences in English:

1. The boys enjoyed playing dodge ball every Wednesday.

2. Please give me a call in ten minutes.

3. I love toast and orange juice for breakfast.
4. There is heavy traffic on the highway.
5. If you listen closely, you will hear the birds.
6. My father is my inspiration for success.
7. I will be in the office in 10 minutes.
8. I will go to India to meet my parents.
9. Turn the music down in your headphones.
10. It all happened suddenly.

3.4.2 Procedure

The experiment was performed to find out whether the distance and volume of a speaker affected the time taken in the identification of the speaker. The distance the participant was sitting at, from the recording equipment controls the changes in the volume of a participant recorded. Hence, the participant was asked to sit at 5m, 10m, and 15m away from the equipment. At each position, the variations in amplitude and frequency were measured, which in turn, influenced the time taken to identify the speaker. The amount of time taken by participants is shown in Figure 3.4 .

3.5 Results

Participants recognise female/male voices from the recordings provided. For example, a female voice sounds different as compared with a male voice. Distinguishing female and male voices helped participants to reduce the candidate population and achieve the highest probability to identify a speaker. Visual and audio representation provides the human brain with a similar pattern as seeing and hearing a person in reality. An audio-visual combination provides information required to identify a speaker within a limited period. For example, a child can identify their mother on a phone call by listening to her voice. The audio call alerts a visual part of the brain. Hence, the audio-visual combination makes it easier to identify the person.

Participants were not sure about the speaker since it was the first time they were listening to the voices. Participants requested to hear the audio files a couple of times before they can

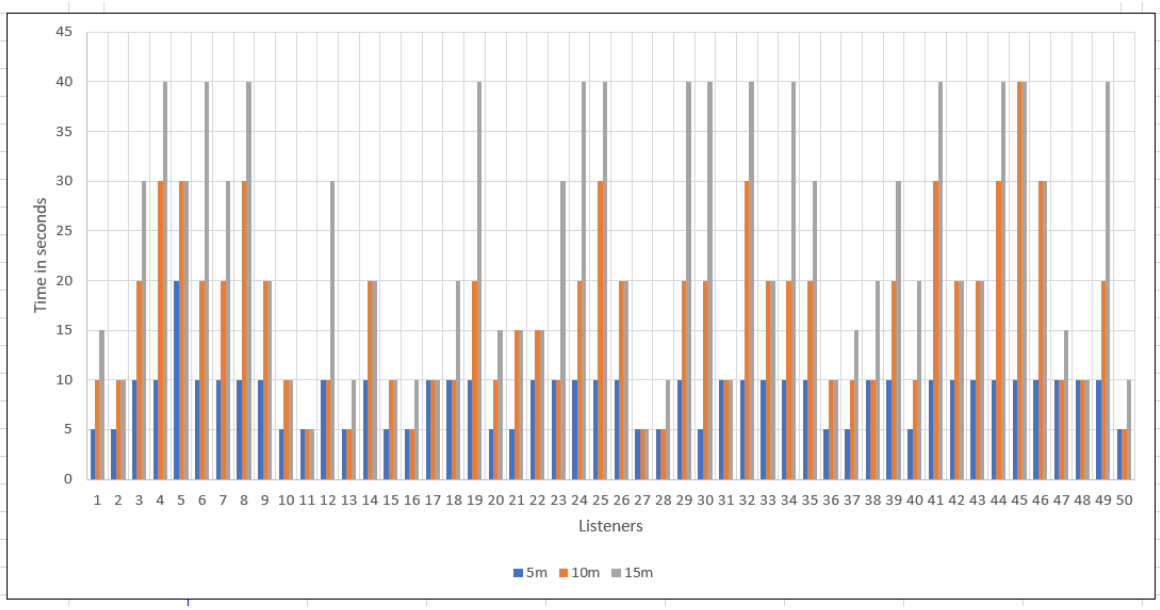

Fig. 3.4 Time Taken to Recognise Who is Speaking

identify a speaker. More than half of the participants were not able to identify a speaker even after listening to it a couple of times. However, participants can be distinguishing between female or male voices: Female voices are often softer than male voices. Females also tend to have a higher pitch and male voices have a lower pitch.

Researchers asked participants which factors helped them to identify a speaker. There were several reasons given by the participants to identify a movie artist:

1. Some movie artists have a unique way of saying a few words in different movies.
2. Some movie artists have a unique accent in all the movies.
3. Certain movie artists have a distinctive voice; for example, deep voice which makes it easier to recognise the artist.
4. Several participants correlate the voice of a movie artist with their faces, since they have already seen and heard the artists in movies.
5. Although some participants were unable to explain how they identified a familiar artist, the rest of them provided the following reasons.

The experiment on different languages proved that human does not need to understand or be familiar with the language, used by a speaker, to be able to identify them.

Distance does have an impact to identify a person. Participants had taken less time to identify a person when the speaker is at 5m away. The results showed that volume and

distance are not dependent on each other in all cases. Some of the participants had taken the same amount of time to identify a speaker independent of distance, whereas, others have taken a long time to identify a speaker when they are at 15m away.

3.6 Summary

To understand how humans are identifying a person, several movie artist's voice samples were collected and a database of 100 voluntary participants was collected, consisting of friends, family, university staff, and students. Each participant was asked to listen to an audio clip of a movie artist and then asked who do they think of talking to. The recording of the first audio clip was used as the training set, and the second audio clip was the test set. The survey took place in a normal office room, using a normal microphone for the recordings. Slight echoing and background noise were present in the samples arising from the computer fans and surroundings. It was observed that the movie artists were listed one by one, female followed by male or male followed by a female in such a way, as to distract the participant's attention from the gender of the movie artist. The results indicate that participants took less time to identify a speaker with who they are familiar, as compared to unfamiliar voices. However, there was not much significant difference in terms of time taken to recognise them. The sound of a speaker's voice is efficient for the listener to identify a person.

Humans have capabilities in their auditory system that are extreme and exceptional in terms of identifying voices. For example, "birth babies" can already recognise the voices of their mothers and a mother can understand what her baby is trying to tell her by listening to the sound which they make, to convey a message. The reason behind this is that humans have enough sensory memory, which gives them the ability to listen and recall from speech, and that includes contextual information about how they expressing the speech. Current technologies can capture a large amount of data in terms of speech, which can be used for speech recognition, but not for speaker identification. So far, only humans can identify a speaker based on their voice with almost 100 percent accuracy.

Chapter 4

Characteristics of a Voice to Identify a Speaker

4.1 Introduction

Speech is a unique mode of communication among humans. Speech is a complex method of communication systems when compared with other methods. As humans also use non-speech, which is non-verbal communication to convey information [78]. Nonverbal communication not only accentuates the meaning of words but also provides information such as, what kind of emotional state the person is in. Non-verbal communication provides a higher level of information, which includes characteristics of a human voice and this chapter will show how humans can use these characteristics to identify a person.

The human voice is extremely difficult for a computer to analyze and recognize [79]. There are two components in human voices: verbal and non-verbal. Human life starts with non-verbal communication with other people. On average, children under the age of two, use the production of sounds instead of words to communicate. However, people who cannot speak use nonverbal communication too. Both children and non-speaking people can communicate efficiently to share information and emotions without using words.

Verbal communication is one of the most common methods used for interpersonal communication. It uses words to convey information to others and conveys information about the speaker. Verbal communication often assists with the identification of the speaker too, but not all the time. Verbal speech includes a speaker's accent, speaking style, and pronunciation, etc [80]. Typically, individuals can identify a familiar speaker with high accuracy, but humans use a combination of parameters to identify a person such as a speaker's accent, speaking style, and pronunciation, etc.

Table 1 provides the variation of human speech and how the human voice changes in a different situation.

Table 4.1 Variations of Human Speech

Variation in speech	**Modulation**
Types of Speech	Reading a book in a Normal/Angry mode. Giving a lecture in a classroom
Effects of Audience	With Whom They are Communicating With, For Example: Children/Parents/Friends/Lectures
Environments	Noisy place such as: Traffic, Noisy-Classroom
Emotional State	Happy/Sad/Angry/Excited
Life span	Age Gap Differences in Children and Ddults, Teenagers or Elderly people
Types of voices	Rough/Loud/Soft

4.2 Internal Mechanics of Human Voice Production

Identifying language-independent features of a voice is key to investigating the unique characteristics of a speaker's voice. To be able to identify the language-independent parameters, one should understand firstly how human speech works. A voice pattern can be considered as one of the bio-metrics that is unique to an individual in the same way that fingerprints, iris pattern and DNA are [81].

4.2.1 Production of a Human Voice

The input for the human voice is air, which passes through the lungs, then through the vocal folds to produce a sound, as shown in Figure 4.1. This sound is a part of the means of communication, but it does not help us understand what the speaker is trying to convey. Sound is carried through the vocal tract, (combination of mouth, lips, and tongue) which acts as a filter, making sound understandable when it leaves the lips. The average vocal tract length for males is 17cm and 14cm for females [82].

4.2.2 Characteristics of a Human Voice

Humans can identify a speaker in a wide variety of situations. For example, imagine someone is sitting behind you. You can hear, but cannot see, them and cannot understand what they

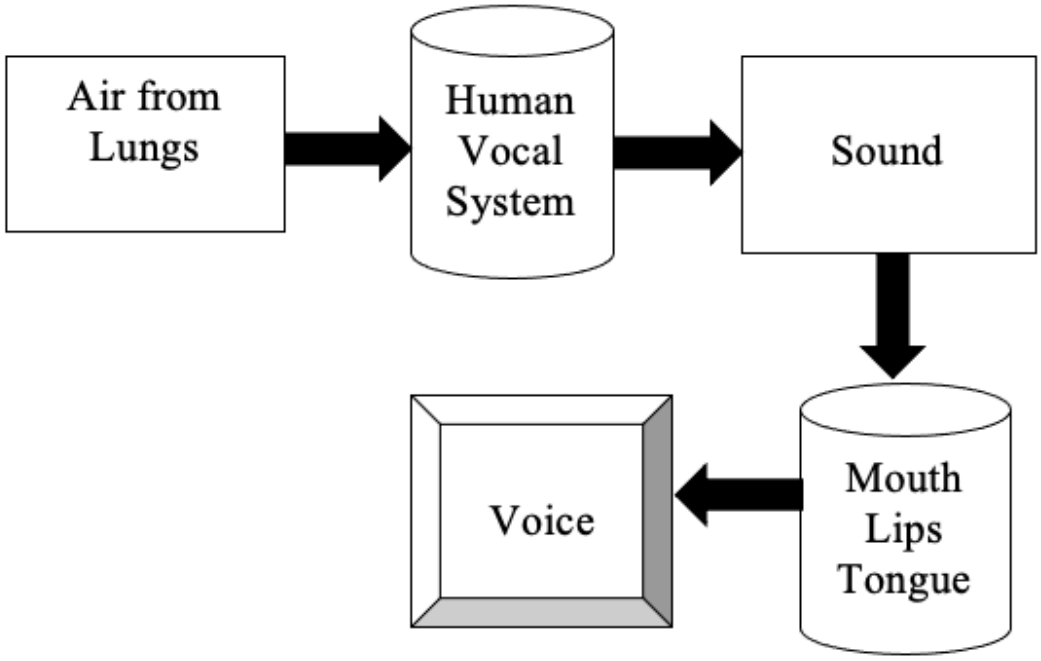

Fig. 4.1 Human Overall Voice Production

are talking about since you do not know the language they are talking in. However, you have enough data to build a picture of the speaker, which includes their gender, approximate age, and even their emotional state. The question is though, what information is required to identify the speaker? To identify a speaker, one should be able to recognize the individual pattern of their voice.

There are three principal characteristics of a human voice: frequency, timbre, and volume, as shown in Figure 4.2. The frequency of a voice depends on the number of vibrations of the vocal cords per second. The vocal cords of men, who are perceived to have a lower number of vibrations per second, normally operate between 100-130 vibrations per second. On the other hand, the vocal cords of women, who are perceived to have a higher number of vibrations per second, normally operate between 180-220 vibrations per second [83, 84]. The second characteristic, the timbre, distinguishes sounds that have the same frequency and loudness (volume). Timbre is also called tone colour or tone quality. For example, each musical instrument has a different timbre, which is represented by comparing harmonics that are present besides the fundamental frequency [85]. Lastly, the volume or amplitude of a voice is the vibrations that affect loudness [86]. The higher the amplitude of the vibrations, the larger the amount of energy carried by the wave & thus the louder it is. The units of volume are measured in decibels (dB). Volume relates to how the waves, produced by the

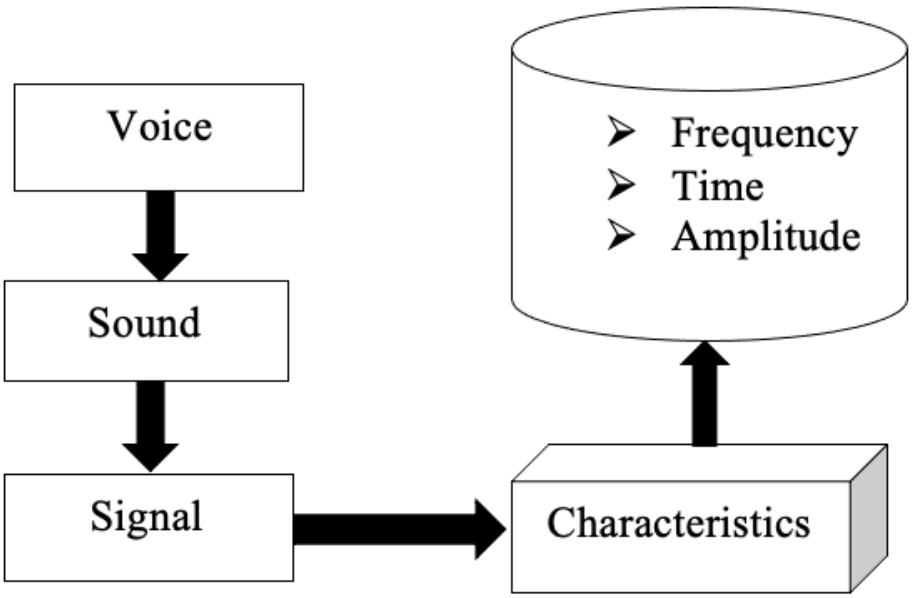

Fig. 4.2 Characteristics of Human Voice

vocal cords, are amplified within the body based on factors such as the speakers' mood, with whom the person is conversing, the context of the conversation, how much physical effort the person is putting into it and so on [87, 88].

4.3 A Preliminary Study of Human Voice Characteristics

The experiment was conducted and 100 participants were involved; 35 female and 65 male, ages ranging from 20 to 40 years old. 30 participants are native English speakers and others are from different countries namely Egypt, India, Germany, France, Ethiopia, Saudi Arabia, Sri Lanka, etc. A script was developed for participants to read a list of sentences.

The script below shows a sample of what participants were asked to read, which was recorded for the study.

1. The boys enjoyed playing dodge ball every Wednesday.
2. Please give me a call in ten minutes.
3. I love toast and orange juice for breakfast.
4. There is heavy traffic on the highway.

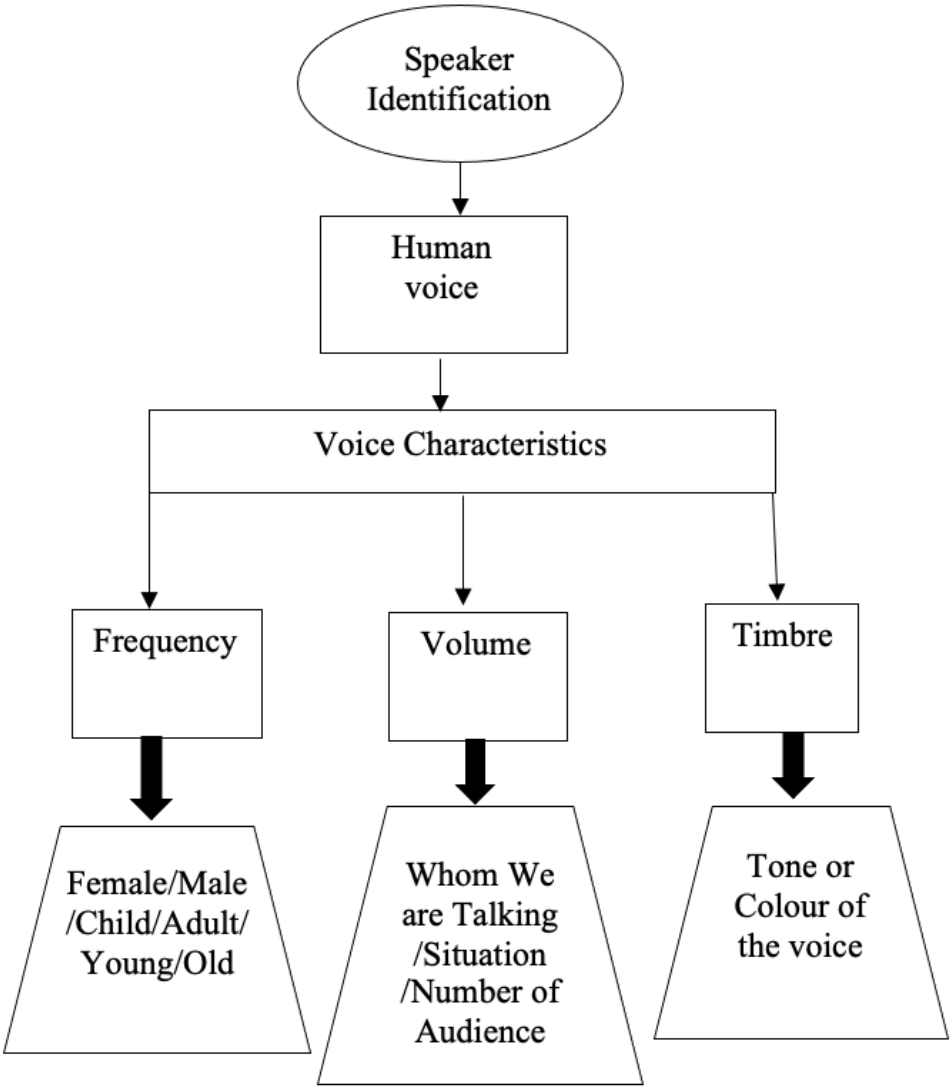

Fig. 4.3 Voice Characteristics

5. If you listen closely, you will hear the birds.
6. My father is my inspiration for success.
7. I will be in the office in 10 minutes.
8. I will go to India to meet my parents.
9. Turn the music down in your headphones.
10. It all happened suddenly.

4.3.1 Initial Analysis

An ideal voice recognition system should aim to generate voice patterns that are independent of the language spoken. Only the participant's voice should be required to provide input to the voice recognition system for testing and development purposes, i.e. no other constraints such as a specified language or content. A consent form was prepared for the participants, explaining the purpose of the research, and participants were asked to go through the form before recording was started. All participants were older than 20 years of age and understood the English language. Participants were asked to read out a prepared script, which consisted of ten sentences that included all phonemes in the English language.

4.3.2 Frequency Analysis

Spectrum analysis transforms a sound wave into the frequency domain. The sound of a voice is created from vibrations produced by a person's vocal folds. But, the voice from vocal folds needs to be filtered to be understandable. The filters in the voice production are nothing but vocal tract/resonators. The sound from the vocal folds is had to pass through by vocal tract, or else humans can't hear the sounds from the vocal colds on their own. The resonators are responsible for producing a unique voice for every individual. By applying Fast Fourier transform (FFT) to a participant's voice recording, the fundamental frequency has been observed for each participant and noted in the Table B.1.

Table 4.2 Analysis of Fundamental Frequency of People's Voices

Participant	**Mean. Freq (Hz)**	**Median. Freq (Hz)**	**Min Freq (Hz)**	**Max Freq (Hz)**
1	223.16	231	192	239
2	580.83	587	520	604
3	533	533	515	558
4	441	441	434	448
5	128.83	128.83	121	142
6	118.16	120	109	126
7	136.5	136.5	133	139
8	130.33	129	123	139
9	571.66	575	534	616
10	213.66	213.66	179	235
11	162.66	163	156	167
12	214	220	184	225

13	119.83	119.83	101	141
14	120.16	120.16	110	130
15	138.33	140	110	155
16	452	452	403	479
17	221.83	223	200	237
18	265.33	259	243	293
19	225.33	225.33	203	251
20	224.16	224.16	199	240
21	227.16	227.16	191	249
22	261.16	259	252	275
23	177.16	177.16	143	223
24	144.5	144	140	156
25	240.33	240.33	228	252
26	258.66	249	225	339
27	262.16	262.16	245	286
28	111	113	90	119
29	141.33	142	119	160
30	126.83	126.83	110	142
31	335.5	335.5	311	369
32	335.16	335.16	314	378
33	376.83	376.83	330	402
34	241.5	241.5	220	261
35	251.16	251.16	227	285
36	226.83	226.83	201	256
37	224.83	224.83	191	268
38	149.5	137	113	260
39	431.33	408	403	486
40	129.66	127	139	123
41	180.66	170	142	223
42	142.66	145	109	164
43	163.66	163.66	131	198
44	247.83	247	217	288
45	166	160	149	196
46	430.66	430.66	420	440

47	518.83	520	472	552
48	545	545	504	591
49	255.33	255	247	266
50	421	453.5	421	488

4.4 Potential Characteristics for Speaker Recognition

So far researchers have been explored the possibility of recognising a person from their fundamental frequency, but what if two participants have the same frequency range? What are the other parameters that one has to consider to identify a person?

4.4.1 Fundamental Frequency

Frequency range values have been observed from Spectral analysis. Each person has a specific frequency range for their Fundamental frequency, by looking at the frequency range, one can eliminate people whose Fundamental frequency falls outside of any observed readings.

Minimum and maximum fundamental frequencies for all participants shown in Figure 4.4. For example, let say a participant frequency is 100 Hz, one can eliminate the people who do not fall under the 100 Hz frequency range, with this one can eliminate on average 40 to 50 % of the population from a database.

4.4.2 Speech Rate

Speech rate is another factor to be considered to identify a speaker. People communicate with each other at different speech rates [84]. An experiment was conducted where participants (speaking in the English language) were recorded 6 times, in a noiseless room.

100 participants were requested to read a script as mentioned in the 4.3. Their speech rate was calculated as the number of words per minute, as shown in Table B.2

Table 4.3 Participants Speech Rate was Observed

Participant	Min SR (WPM)	Max SR (WPM)	Mean SR (WPM)	Median SR (WPM)
1	98	110	104	103
2	110	120	113.83	113
3	106	118	113	113

Table 4.3 continued from previous page

4	118	134	126.83	127.5
5	110	135	121.33	120
6	92	100	96.66	97
7	110	135	121.11	120
8	126	135	129.83	128.5
9	100	120	112.5	112.5
10	140	142	140.83	140.5
11	110	135	111.83	111
12	108	120	113.83	112.5
13	106	118	111.66	111
14	125	134	129.83	129
15	110	135	121.33	120
16	126	134	129.5	128.5
17	145	150	148.83	150
18	135	140	137.16	136.5
19	125	126	125.16	125
20	115	130	121.33	120
21	90	95	92	91
22	135	138	136.66	136.5
23	126	150	144.83	149
24	128	132	129.66	130
25	140	140	140	140
26	90	98	93.33	93.5
27	115	120	117.66	116.5
28	128	132	130	130
29	100	106	101.83	100
30	110	115	111.83	111.5
31	140	145	141.5	141
32	124	135	129.33	129.5
33	120	140	128.83	127.5
34	90	100	94.66	95
35	110	140	121.16	117.5
36	100	105	101.5	101
37	145	150	148.16	149.5

Table 4.3 continued from previous page

38	110	118	114.16	115
39	130	140	136.66	137.5
40	125	135	129.33	128.5
41	120	133	127.66	129
42	100	140	110	100
43	124	129	125.5	125
44	130	138	132.83	133
45	125	130	127.83	128.5
46	130	137	132.5	131.5
47	140	143	141.5	141.5
48	90	96	93.16	93.5
49	121	130	125.33	125
50	110	120	115.33	115

Speech rate involves both physical and psychological characteristics of a person, such as their: gender, age, emotional state, and movement of lips, and tongue, etc [17]. Speakers can change their speaking rate if they would like to do so. However, changes in speech rate can happen without a speaker's knowledge, because speakers cannot always control the way they are speaking. The following factors impact the speech rate of a speaker and perception of a listener as shown in Figure 4.5.

Natural (relaxed) Speaking Rate:

This is the rate of speech that people use to communicate with their family, close friends, and people with whom they spend more time. Culture plays an important role and it is where a person's natural speaking rate develops. Even geographical locations can have a major impact on the speaking rate. For example, different locations within the same country often have different speaking rates.

Impact of Behaviour:

The most common impact of behaviour on speech rate is when strangers communicate with each other. Individuals present emotions such as nervousness, and reluctance when they converse with unfamiliar people. For example, presenting in front of an audience for the first time is always nerve-wracking, causing speech rate to be faster or slower rate than usual.

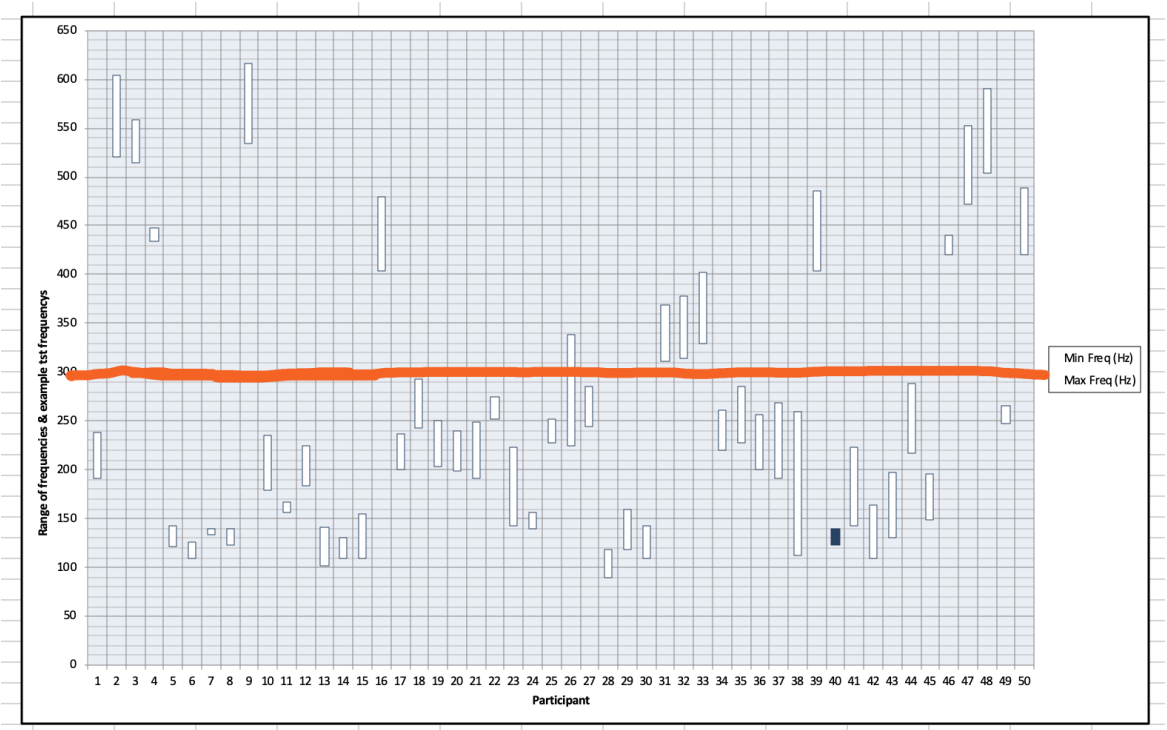

Fig. 4.4 Elimination of Possible List People From a Database

At a Work Place:

Workplaces usually involve working at a fast pace to produce quicker results, which undoubtedly causes stress. If a person is unable to work under pressure, they might be mentally processing information at a slower rate, which can cause them to talk slowly too, thus, reducing their speaking rate. However, they talk faster to keep up with the fast work pace.

Speeches:

During a speech in front of an audience, the speaker would normally take more pauses than usual to gain maximum attention from the listeners & also to allow them time to pick & choose their words carefully. This is usually beneficial to convey a message or gain support from the audience. Such practices are most commonly seen by leaders, politicians, and professional speakers.

Emergency Situation:

People will talk faster when they are in an emergency so that they can convey their problems to the listener as soon as possible. Normally, this is observed in situations where help is required such as an emergency call for an ambulance, or the police.

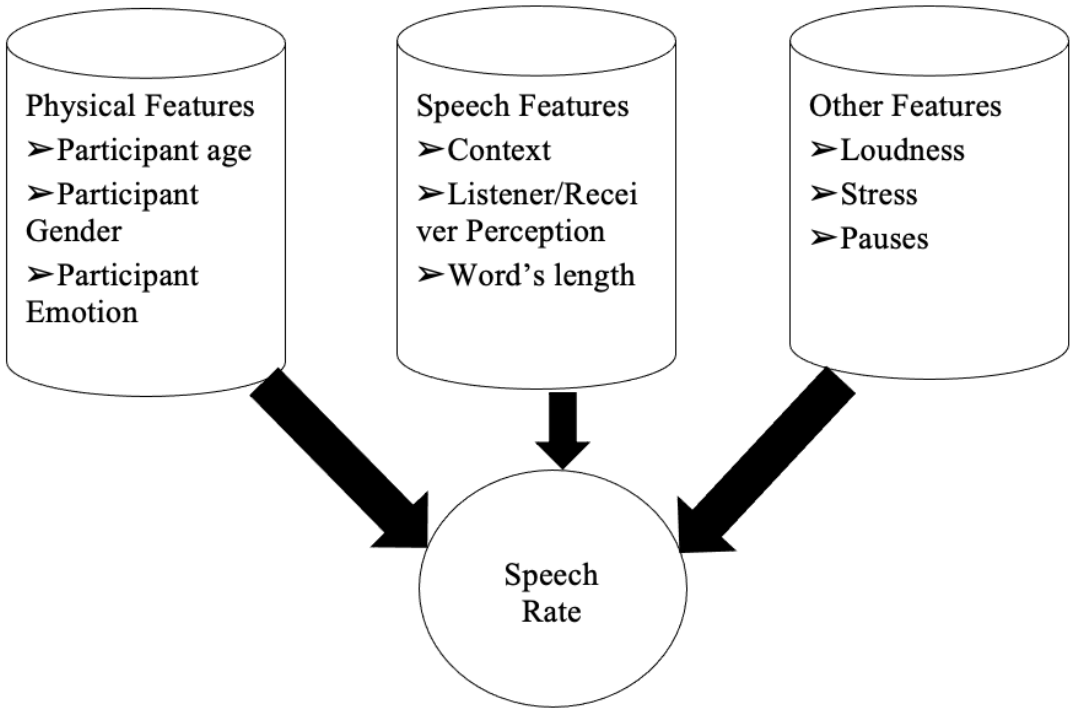

Fig. 4.5 Causes of Variations of Speech Rate

Contexts of a Speech:

Sometimes people either talk slowly or quickly, depending upon their knowledge of what they are talking about. If their understanding of the subject is clear and thorough, they might talk comparatively faster than normal. Equally, if they are unsure, they may talk more slowly.

Vocabulary:

Generally, if sentences have longer words and are difficult to pronounce, speakers take a longer time than normal to finish their sentences.

People have different speech rates based on the above-mentioned factors as shown in Figure 4.6, but still, speech rate, may be specific to an individual speaker. Speech rate may, therefore, possibly be used to identify a speaker.

4.4.3 Articulation Rate

Articulation Rate (AR) is defined as the number of speech units delivered per second. The speech units can be syllables or words [19]. AR is similar to SR, but the main difference

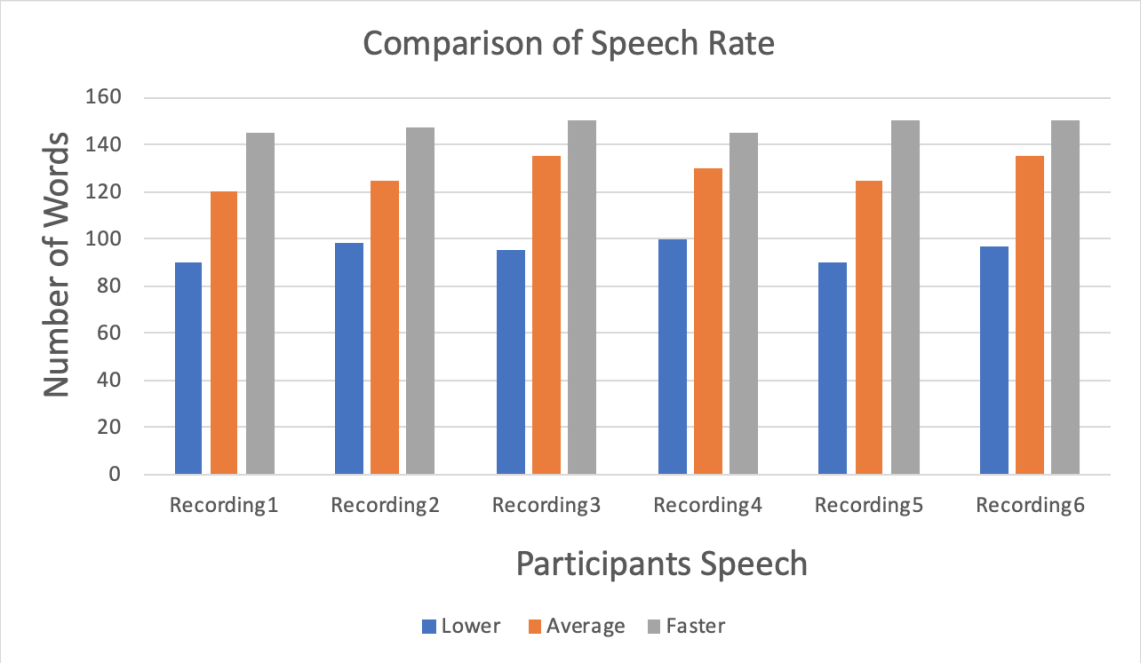

Fig. 4.6 Comparison of Speech Rates for One Participant Recorded 6 times

between them is that SR includes pauses, whereas, AR does not. The speed of speaking can either be defined as SR or AR as "speech rate". AR and SR depend on the continuity of the speech. For example, both SR & AR can be fast speaking rates when speech is fluent.

In summary, AR & SR are the same if the speaker has no pauses at all. SR is the mean of the words or syllables per minute including pauses. AR is the mean of words or syllables per minute recording between pauses. Thus AR will always $\geq$ SR.

4.4.4 Accent

An accent is one of the keys to human speech to identify their locality. An accent provides various details about a speaker, such as an ethnicity, social status, and first language.

People often tend to mimic the other person's accent subconsciously when they are conversing. Everyone has an accent in their speech community, and some words are more pronounced than others. Hence, one may be able to use an accent to identify a speaker, however, it will be difficult if all the speakers are from the same locality.

4.4.5 Pause

During speech, there are two types of pauses: intentional (conscious) pauses and Natural (unconscious) pauses.

Intentional pauses may occur when:

1. While giving a presentation, one make sure to give a pause in between our words to ensure that the, audience is listening & make our speech clearer.
2. While discussing a project topic with our supervisor/project leader.
3. Trying to choose words more carefully.

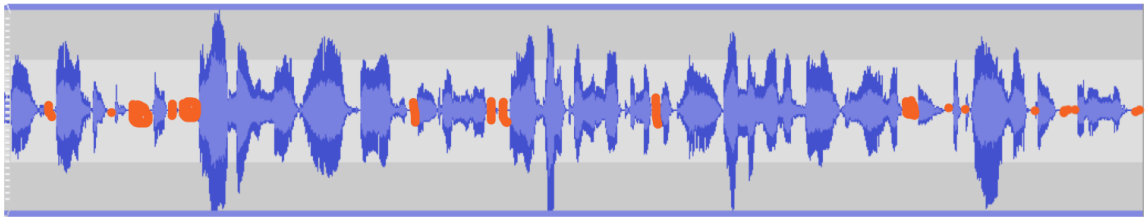

Fig. 4.7 Pauses by a User (Blue Denotes Speech and Red denotes Pause)

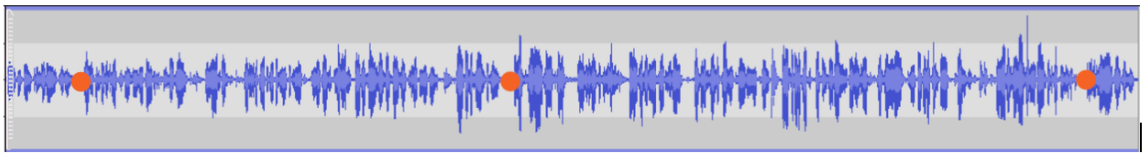

Fig. 4.8 Fewer Pauses by a User (Blue Denotes Speech and Red denotes Pause)

Natural pauses will occur without our knowledge or consciousness, in situations such as:

1. While talking in our first language.
2. Relaxed conversations with parents/family/friends.
3. When giving a speech on something you are very well-versed in.

4.4.6 Speech Variation

Human voices change over time, from birth through puberty & into old age. For example, children/infants sound different compared with adults. Voices sometimes change during day and night. Recordings of people's voices in the morning and evening time show their relative amplitude and frequency values changed based on various reasons.

1. Participants are more active in the morning and they became tired by the night because of work during the daytime.
2. Some participants were active in the evening since they were about to go home.
3. Some participants sounded the same during morning and evening hours.

4.4.7 Impact of Audio Equipment

Human speech communication often takes place in a noisy environment. Background noise comprises unwanted noise from the surroundings and becomes a part of the recording. Different types of recording equipment record the speaker's voice differently. In this project, the choice of microphone is important to capture a participant's voice. Certain microphones are designed for a specific purpose and a specific environment. Background noise and quality of recording equipment systems can create further challenges for voice recognition systems. If testing is done over the phone, the accuracy of the results can be affected by noises such as people talking or driving a car in traffic, etc.

4.5 Results

A speaker's voice varies based on several factors and situations. However, there is a list of parameters that can be used to identify a speaker. The frequency of the highest peak is one of the parameters used to identify a speaker. Female and male participants typically have different frequency ranges. There are two ways of identifying a speaker based on frequency values. Firstly, one has to decide whether a speaker is female or male. Secondly, comparing the frequency value of a person with all the participants. And, finally, eliminating the ones which do not match.

4.6 Summary

The characteristics of the voice have been analysed and observed. A database of 100 participants voice samples was collected, consisting of university staff and students, who live in the UK. Each participant was asked to read the script, which was designed to include all phonemes in different contexts. The recording took place in an anechoic chamber, using a high-quality microphone for the recordings. There was no echo and background noise. The vibration of a speaker's vocal folds, followed by patterns created by the physical components from the human speech is as unique as a fingerprint. Speaker recognition systems capture unique characteristics of a voice, such as tone, frequency, volume to be able to identify a speaker.

Chapter 5

Variations of a Speaker's Voice

5.1 Introduction

Communication is an essential part of human life. People use speech (words) to convey information to each other. Vocal tract characteristics in a human voice help identify a speaker. However, human speech signals are language-independent and information is speaker-dependent. With advanced technology, a person can use their voice for biometric authentication, which is unique for individuals [89, 90]. An individual can use their voice in different applications, such as adding an extra layer of security in online banking, etc. The main aim of this chapter is to identify the uniqueness of a voice, which should be independent of a speakers' language

Humans can communicate with each other in different ways such as speech, gestures, writing text, drawings, facial expressions, body, and sign language. Verbal communication is one of the natural modes of communication. The human voice provides two levels of information. At the primary level, the human voice uses words to convey a message and at the second level, it conveys speaker information about language, emotion, gender, age, and generally, the identity of the speaker. There are two general types of speaker recognition systems: Speaker Verification and Speaker Identification as shown in Figure 5.1.

In speaker verification, the purpose is to check if someone is who they claim to be. Speaker verification is a one-to-one comparison made between a speaker's voice and their reference profile stored in a database, which results in acceptance or rejection of an identity claimed [91]. Speaker verification can be divided into two types, i.e.: Text-Dependent (TD) and Text Independent (TI) [92]. In TD, the speaker must use a specific list of words/sentences for both enrolment and testing, and since there are known to the machine, the overall recognition rate increases. However, in TI, there is no fixed text. For example, a speaker can choose to speak any words/sentences, as the machine has a much wider spectrum of received

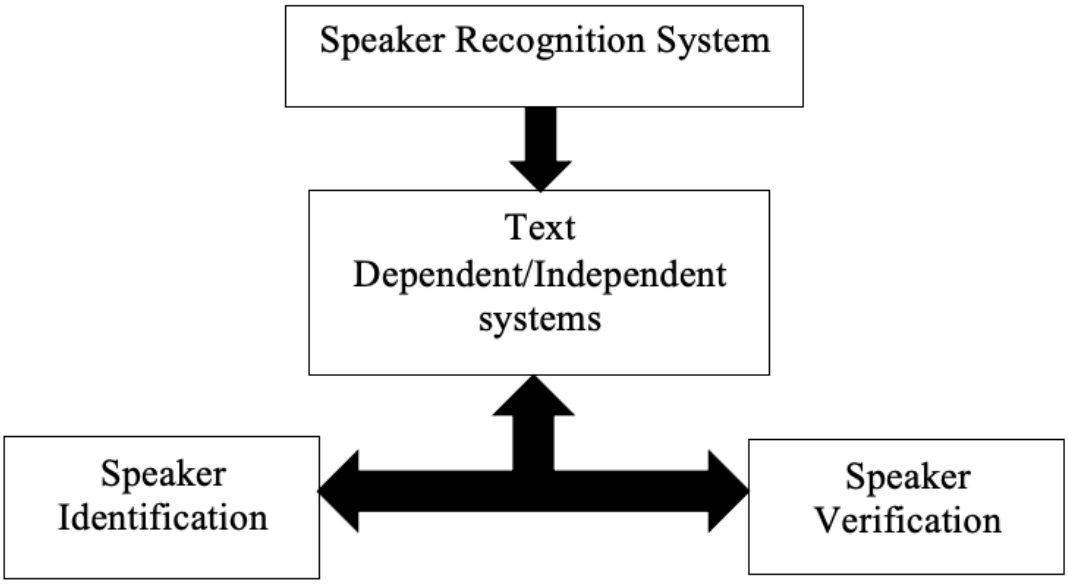

Fig. 5.1 Block Diagram of a Speaker Recognition

input, this approach tends to affect recognition rates, because, both enrolment and testing have different text and the latter will most likely be are unknown to the machine.

In speaker identification, the aim is to identify a person out of a larger group as shown in Figure 5.2, by listening to their voice and checking whether that person's profile is similar to someone in the database. The number of decisions to identify a speaker is approximately equal to the number of participants in the database. Speaker verification becomes challenging if the voice cannot be matched with a voice that has already been stored in the database.

5.2 Background

The problem with human speech is the significant amount of variation that takes place when pronouncing a word. This variation occurs based on stress, environment, recording equipment, etc. Pronunciation of some words is similar in patterns and a machine can easily get confused between them. Furthermore, just like humans, a machine may have a hard time differentiating the same language with a different accent [93]. The speaker identification system's accuracy decreases as the number of speakers increase, whereas, speaker verification accuracy is largely a constant independent of the number of speakers, as it is based on the comparative similarity of two data samples only. Speaker identification, as

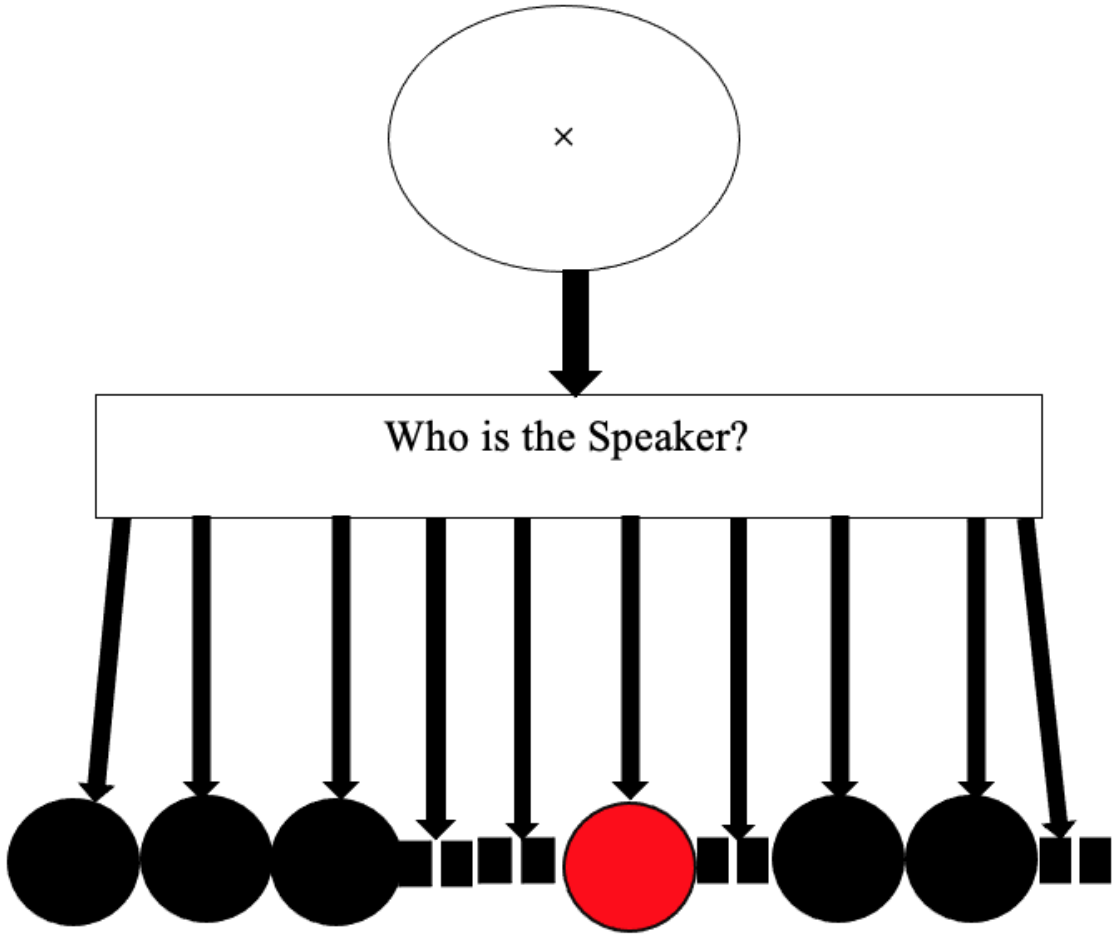

Fig. 5.2 Block Diagram of Speaker Identification

well as verification systems, needs to focus on identifying voice characteristics to uniquely recognize a speaker, independent of what has been said [92].

5.2.1 Feature Extraction

Extracting features is one of the key components in speaker recognition systems. Currently, Mel Frequency Cepstral coefficients (MFCC) are using in speaker recognition systems, have achieved good accuracy in speech recognition [94, 95]. The main problem of MFCC is that coefficients change when a speaker changes her/his language, therefore, it is not recommended for speaker recognition systems. The other problem is when several speakers increase, there is a possibility of extracting likely features from two or more participants who have a similar accent.

The following points should be considered while selecting features to be able to identify a speaker.

1. Features should not be affected by a speaker's aging and health such as: cold, hay-fever, etc.
2. Features should be difficult for others to mimic.
3. Features should be independent of background noise.

However, it is difficult to achieve the above-mentioned characteristics by using an individual feature extraction technique.

5.2.2 Pattern Recognition

The human brain is a complex organ that consists of several systems and subsystems, which have proven to be particularly challenging to discover and understand. So far, research has shown that the brain has different layers with different functionalities to process information from sensory organs and the surrounding environment. The human brain receives input from these sensory organs to learn information and improve overall intelligence. The human brain does not need these sensory organs to be intelligent on its own.

For example, Helen Keller has no sight or no hearing, yet she managed to learn how to write and become one of the best authors in the world. The human brain does not discriminate between hearing, feeling or vision, because it ultimately takes information directly from any sensory organs as patterns. After all, everything in the world around us is based on patterns. For example, the human face is learned as a collection of patterns by combining the patterns of nose, mouth, and eyes [96]. How can machine-learning algorithms learn to acquire information like the human brain?

Humans are good at identifying patterns in, for example, language, music, animals, and people, whereas in artificial intelligence identifying patterns is still a challenging task. The state of the art in pattern recognition techniques used for the voice recognition system are Vector Quantization (VQ), Dynamic Time Warping (DTW), and Hidden Markov Model (HMM).

VQ is a model that represents a larger amount of data into more compact data in the vector space, and each region is called a cluster and then represented as a code-word. The combination of a codeword is called a codebook. In the recognition phase, VQ calculates the code-book with the smallest distortion and then identifies a speaker. If VQ has more code-books, then the system needs more computational power to find the best match/smallest distortion among the speakers [97].

DTW is a method used to identify a similarity between two speech signals. It allows a machine to find out the best match between two signals and then makes that best match as the third signal for those signals. However, a problem with DTW is that it compares speech signals that are based on time. The human voice will change over time due to aging.

HMM is a model that is based on probability and uses a Markov process that generates hidden and unknown parameters, and then uses those parameters for further analysis. HMM is used in speech recognition systems, but is of limited use in speaker recognition [98]. The main disadvantage of HMM is the future state is based on the present one, not on the events that happened before. Voice is not based only on the current situation, because voice changes over time.

A speaker recognition system needs to use the vocal features of the speech to create a voice pattern. Most of the speaker recognition system patterns designed are for one language, usually English. Once a speaker changes her/his language, their patterns of voice may change, and then there is a possibility to reject a person incorrectly. To cover this problem, one needs to develop a speech recognition system, that should be independent of her/his language by extracting features from the voice itself.

5.3 Methodology

This research focuses on establishing how one might go about developing an algorithm to help, build a language-independent, speaker identification system. To achieve this, one needs to understand, how people identify a speaker to then create a technology that is more accurate and intuitive for people to use.

To find out how the human voice changes within the same person when they talk in two languages. An experiment was conducted with the help of both females and males. There were ten participants, 6 females and 4 males. All participants were over 20 years and their age range between 20 to 40 years old. All the participants lived in UK and English is not their mother tongue. For the second task of the experiment, all participants were over 20 years and their age lies between 20 to 40 years old. All the participants lived in India and English is not their mother tongue.

The overall experiment was divided into two sub experiments which were: scripted speech and free speech. In the first experiment, participants were requested to read a few sentences in English and their Native/First language, and in the second experiment, participant's voices were recorded from a TV show where participants talked in both English and their first language. The main aim of this experiment was to observe whether the characteristics of a voice are dependent/independent of the language.

5.3.1 Experiment 1: Voice Characteristics of a Speaker's Voice in Multiple Languages for Scripted Speech

In this study, participants were asked to read/speak a few sentences in English and possibly one other language, which was their first/native language. These recordings were not a test of the participant's knowledge of a language and there was no need to worry about the grammar or structure of his/her language.

The participant's voices were recorded, while they were reading a script in their known/first language. The script had been prepared in a way that, participants could easily translate into their native language. If participants could not do the translation, translating software was provided to support them.

Data Collection

The participant's task was to read a script and this recording was undertaken in a silent/quiet room allocated especially for this research. The script was designed so that it could be easily read by all participants and prevent the usage of foul language as well. The participants were given an option to do some trial recordings before the actual recording to allow them to become more comfortable with the process.

The recording was done in a noiseless room i.e. an anechoic chamber in the Nelson building at the University of Greenwich. The following were used for the recording as shown in Table 5.1.

Table 5.1 Experimental Conditions for Experiment 1

Language	English and Speaker's First/Native Language
Recording Equipment	Audacity, Scarlet 2i2 studio, Anechoic Chamber
Operating System	Mac-Book Pro
Programming Language	Python
Sampling rate	44100
headset or headphone	Participant Choice

Participants were all above 20 years old and were able to read and speak English. The participants were between 20 to 40 years of age that took part in this research. Participants were recruited through direct approaches, such as by emailing university staff and students. Once their voice had been recorded, the recordings were anonymised, and then analysed to identify the unique parameters.

The participant was asked to read the script which was mentioned in Section 4.3 and comprised of a set of sentences, of which some can be found below.

1. I will go to university as I am doing a course in YYY.
2. I have two best friends in my school.
3. I would like to be a teacher at university.
4. I have one sister and two brothers.
5. I would like to sing a song.

Frequency Analysis (Scripted Speech)

Participants were asked to read the above sentences in both English and their native language. The purpose of the recording was to observe and analyse the language-independent parameters of the human voice. The experiment was carried out to observe how frequency and amplitude values changed for a speaker when they talked in two different languages. Participants were asked to read the script in English and their native first language. Fast Fourier Transform (FFT) has been applied to the script to evaluate whether the fundamental frequency changed over the languages. The FFT analysis showed similar waveforms, which suggests that fundamental frequency is independent of the language being spoken, as shown in Table 5.2 and 5.3

Table 5.2 Fundamental Frequency of Participants Observed in English Language (Scripted Speech)

Participant	**Mean Freq (Hz)**	**Median Freq (Hz)**	**Min. Freq (Hz)**	**Max. Freq (Hz)**
1	174.5	175.5	170	177
2	195.66	188.5	180	220
3	156.83	160	135	170
4	137.33	139.5	114	148
5	174.16	164	126	223
6	153	155	140	160
7	219.5	225	187	240
8	181.33	180	170	200
9	134.66	135	125	148
10	226.66	230	210	240

Table 5.3 Fundamental Frequency of Participants was Observed in Native Language (Scripted Speech)

Participant	Mean Freq (Hz)	Median Freq (Hz)	Min. Freq (Hz)	Max. Freq (Hz)
1	175	175.5	170	177
2	195	187.5	189	220
3	155.83	157.5	135	170
4	135.83	136.5	114	148
5	173.16	160	126	223
6	153	155	140	160
7	207.83	200	187	240
8	180.33	180	170	200
9	133.66	131	125	148
10	224.83	220	210	240

The fundamental frequency has been observed both in English and their native language. There was no significant differences between the languages as shown in the Tables 5.2 and 5.3.

Measurement of SR and AR Values (Scripted Speech)

Audacity software was used to cut each speech sample into 40-second snippets. The data was edited and Praat software was used for further analysis. Measurements of SR are calculated by diving the syllables by the whole speaking time and for AR, the time was measured without pauses.

$$Speech\ Rate = \frac{Total\ Number\ of\ Syllables}{Minute} \tag{5.1}$$

$$Articulation\ Rate = \frac{Total\ Number\ of\ Syllables\ Without\ Pauses}{Minute} \tag{5.2}$$

The speech rate for all participants when they talk in English language and their first languages such as: Arab, German, Palestine, Portuguese, French, Sinhalese, Telugu, Creole is shown in Figure 5.3, while articulation rates are shown in Figure 5.4

Participant 7 is the fastest speaker with the highest SR in English but not in the native language. Participant 10 is the fastest speaker with the highest SR value in the native language, but not in the English language. Participant 10 is with highest AR in the native

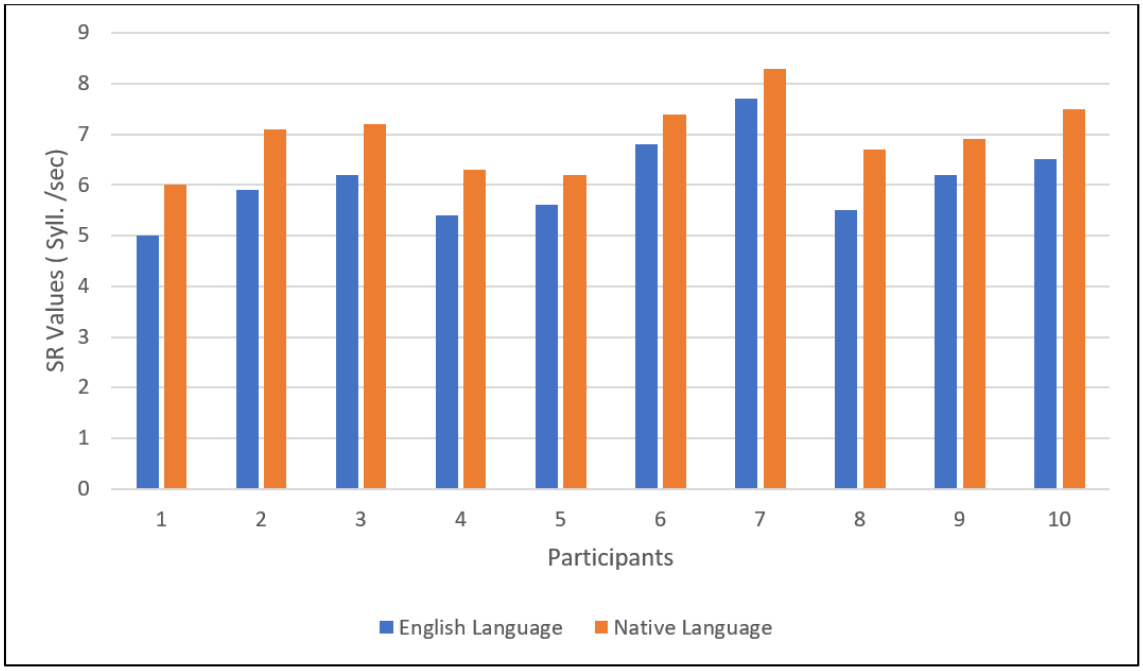

Fig. 5.3 Values of the Speech Rate for Ten Participants in English and Their Native Language was Calculated

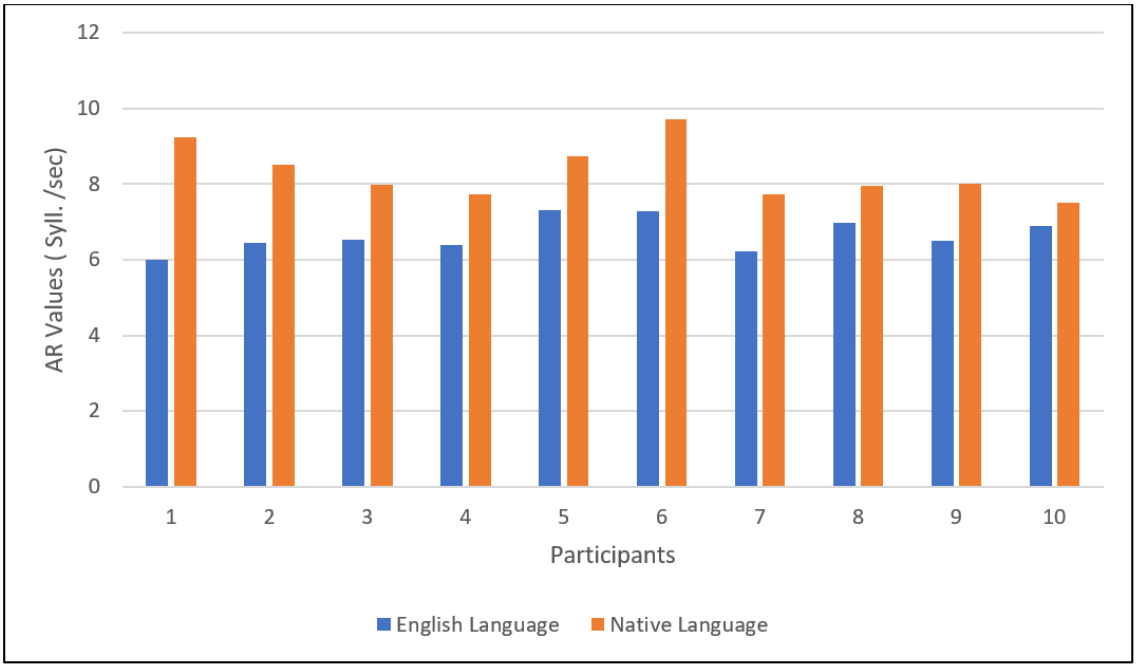

Fig. 5.4 Values of the Articulation Rate for Ten Participants in English and Their Native Language was Calculated

language but not in SR values. Participant 5 is the highest AR value in English but not in the native language. SR and AR not directly proportional to each other.

Estimation/Calculation of Pauses (Scripted Speech)

Each participant has some pauses when they were reading the scripts. Some pauses were silent (where they did not say anything in the speech) and some were filled such as: um, mm, repetition, prolongation, etc. It was necessary to listen to their speech carefully and marked the pauses of both English and their Native Language, as shown in Table 5.4 and Table 5.5

$$Percentage\ of\ Pauses = \frac{Total\ Pauses\ Time}{Minute} \times 100 \tag{5.3}$$

$$Percentage\ of\ filled\ to\ all\ Pauses = \frac{Filled\ Pauses}{All\ pauses} \times 100 \tag{5.4}$$

Table 5.4 Observation of Pauses and with Their Types for Participant 1 (English Language Scripted speech)

Pauses	**Type of Pause**	**Duration of pause (msec)**
Pause 1	Silent	350
Pause 2 [Filled]	Insertion	280
Pause 3 [Filled]	Prolongation	470
Pause 4	Silent	580
Pause 5	Silent	230
Pause 6 [Filled]	Repetition	140
Pause 7	Silent	250

Table 5.5 Observation of Pauses and with Their Types for Participant 1 (Native Language Scripted speech)

Pause	**Type of a Pause**	**Duration of a pause (msec)**
Pause 1	Silent	330
Pause 2 [Filled]	Repetition	410

Table 5.5 continued from previous page

Pause 3 [Filled]	Insertion	290
Pause 4	Silent	350
Pause 5	Silent	420

Volume for all Participants for Experiment 1 (Scripted Speech)

The mean intensity was calculated for all participants by using Praat software as shown in Figure 5.5.

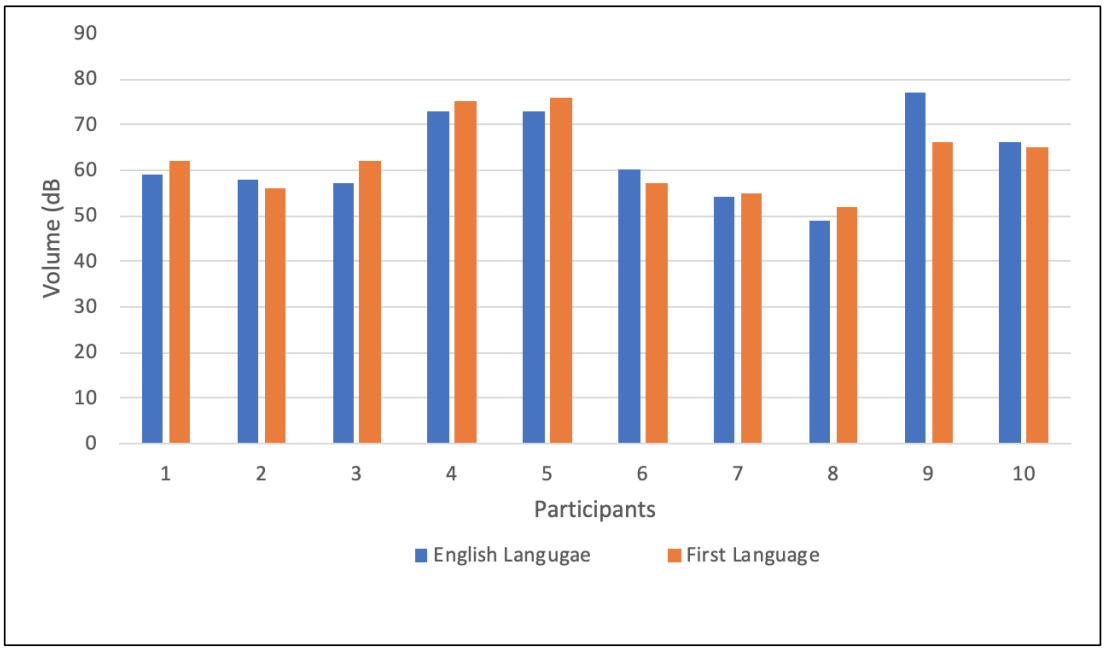

Fig. 5.5 Mean Volume for All the Participants in Experiment 1

Participant 9 had the highest volume in the English language and participant 5 recorded the highest volume in the native language. Individually, there was no significant difference if a participant spoke in either English or Native language. However, it was observed, that participants who recorded the highest volume in one language did not mean that the same participant would maintain the highest volume when they spoke in other languages.

Table 5.6 Mean Attributes for All the Participants

Voice Characteristics	Participants (English and Native Language)									
	1	2	3	4	5	6	7	8	9	10
Frequency (Hz)	116.22 117.78	212.09 212.06	218.35 228.26	193.81 216.80	236.40 239.42	106.25 116.91	225.63 226.45	183.88 194.07	163.68 142.25	205.61 189.45
Volume (dB)	59 52	58 56	57 62	73 75	73 76	60 57	54 55	49 53	77 66	66 65
Speech Rate (Syll./sec)	5.0 6.0	5.9 7.1	6.2 7.2	5.4 6.3	5.6 6.2	6.8 7.4	7.7 8.3	5.5 6.7	6.2 6.9	6.5 7.5
Articulation Rate (Syll./sec)	6.01 9.23	6.45 8.52	6.52 7.98	6.39 7.73	7.31 8.73	7.29 9.71	6.23 7.72	6.98 7.95	6.51 8.01	6.88 9.42
Number of Pauses	07 05	06 04	05 03	06 04	04 03	05 03	06 03	03 02	04 02	05 03
Pause Percentage %	3.83 3.01	3.01 2.01	2.91 2.01	4.01 3.01	2.33 1.83	3.16 2.01	3.50 2.83	2.50 1.66	3.33 2.66	3.66 2.16
Filled Pause Percentage %	42 40	33 50	60 66	50 50	25 33	40 33	66 66	66 50	75 50	80 66
Total Pause Time Sec. /Minute)	2.3 1.8	1.8 1.2	1.75 1.25	2.4 1.8	1.40 1.10	1.90 1.20	2.1 1.7	1.5 1.0	2.0 1.60	2.2 1.3

The mean frequency did not change in both English and Native languages for all participants. Frequency and volume of the participants were independent of the languages. SR values AR values appeared faster in Native language than the English language. More pause in English than native language. Statistically there is no significant difference of these parameters in both language.

5.3.2 Experiment 2: Voice Characteristics of a Speaker's Voice in Multiple languages for Unscripted speech

In this experiment, data was collected from 10 audience –members (5 males and 5 females) of voice English language and Tamil, which is one of the Indian languages. The audience –members did not have any speech disorders. The speakers aged between 20 to 40 years old. The speakers voice was recorded for 40 seconds for each speaker.

Collection of Data

The data was collected through a programme called "Neeya Naana" (Tamil Tv Channel program), which is a show broadcasted on a Sunday night, weekly. The host of the programme is a male who is 40 years old. The host invites people and divides them into two groups such as: college boys vs girls, daughter vs mothers, parents vs children based on the topic being discussed.

The programme gives a topic to debate and allows them to express their views, opinions, suggestions, advice's, anger and frustration etc. The debate topics would include current social affairs, economic and political situation etc. The following are the few examples:

1. Competitive exams
2. Jallikattu
3. Love/arrange marriages
4. Equality of women & men at work place
5. Usage of phone (spending more time on social media), etc.

The host starts with a simple question such as: what is your opinion on "working women"?, "what was your favourite movie"? and so on. Then, people from both groups will start to answer without any preparation, about the questions or topic that a host chooses. The recordings of the TV show were downloaded and the whole duration of each episode was

about 45 to 60 minutes. The recordings were converted into .wav files of 10 minutes duration, then transmitted to MacBook for editing. Audacity software was used for cutting 40 seconds snippets of speech from the audience. The data was edited and used Praat software for further analysis.

Frequency Analysis (Unscripted Speech)

The first step of the experiment was to record all audience –member's voices and apply FFT to the recordings to observe the frequency spectra. Mean, median and participants frequency ranges in English language were observed, noted in Table 5.7 and in Tamil language were in Table 5.8.

The fundamental frequency of participant did not change much when they talk in the two different languages. However, there is a difference, because, the first/native language is more comfortable and easy to converse in when compared to a learned language which was English in this case.The language had more impact on change of a frequency than the familiarity of the script.

Table 5.7 Fundamental Frequency of Audience –Members was Observed in English Language (Unscripted Speech)

Participant	**Mean Freq (Hz)**	**Median Freq (Hz)**	**Min. Freq (Hz)**	**Max. Freq (Hz)**
1	184.16	183	174	200
2	282.66	282.5	275	290
3	190.83	187.5	185	210
4	213.16	213.5	200	220
5	117.5	117	115	121
6	152.33	153	153	160
7	114.5	113.5	110	120
8	237.5	239	230	245
9	179.33	176.5	174	190
10	124.48	125	120	130

Table 5.8 Fundamental Frequency of Audience –Members was Observed in Native Language (Unscripted Speech)

Participant	Mean Freq (Hz)	Median Freq (Hz)	Min. Freq (Hz)	Max. Freq (Hz)
1	185	182.5	175	200
2	284.66	283.5	280	292
3	192.16	188.5	186	213
4	213.5	215	200	222
5	118.5	118.5	116	122
6	153.83	154	145	160
7	115.33	115	110	120
8	239.5	241	230	248
9	180.16	177.5	175	192
10	125.66	125	120	130

The fundamental frequency has been observed both in English and their native language. There was no significant differences between the languages as shown in the Tables 5.7 and 5.8.

Measurement of SR and AR Values for All Audience –Members for Unscripted Speech

The speech rate was calculated as the number of syllables per minute and AR was calculated by dividing the number syllables, but excluding pauses and repetitions, by the articulation time. The data was analysed manually using the Praat software. The following three steps were used for analysis:

1. The first step was to process of 40 second recordings of movie artists, which were then phonetically separated using International Phonetic Alphabet (IPA) and data was transcribed manually by listening the audio files carefully. The problem of doing this manually was counting pronounced syllables a particular time.

2. The second step was dividing the syllables from a speech which was again done manually by a researcher.

3. The last step was to calculate speech and articulation rates from the transcribed speech.

The speech rates observed for all the participants when they spoke in English and their first languages, which was Tamil, are shown in Figure 5.6, while articulation rates are shown in Figure 5.7

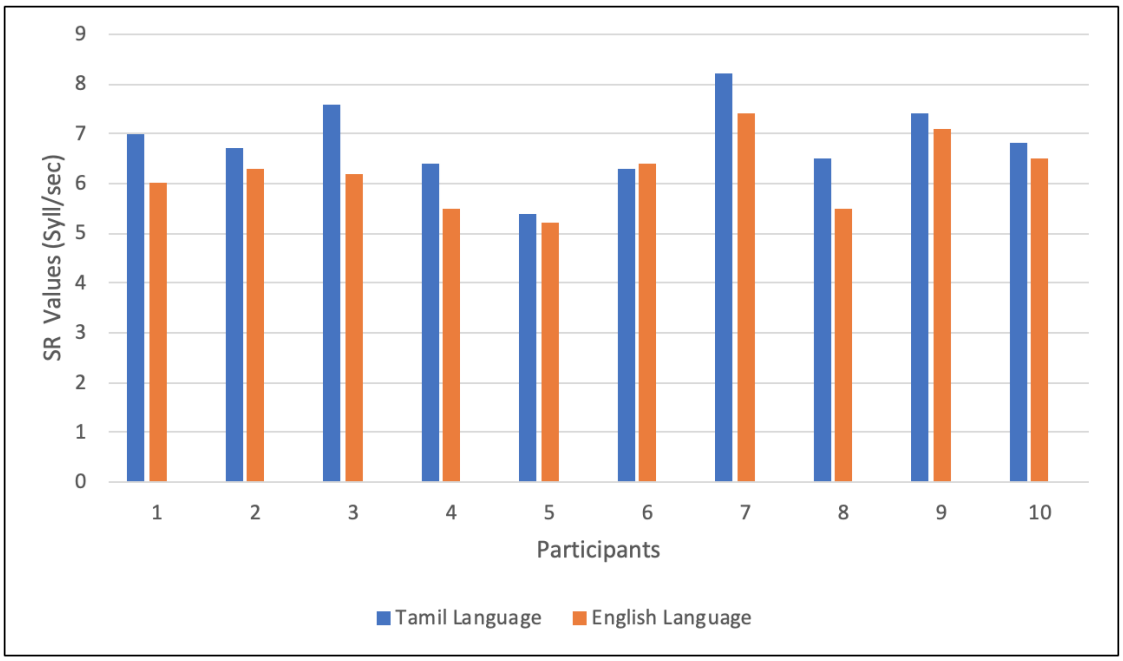

Fig. 5.6 Values of the Speech Rate for All Audience –Members in English and Tamil as Calculated

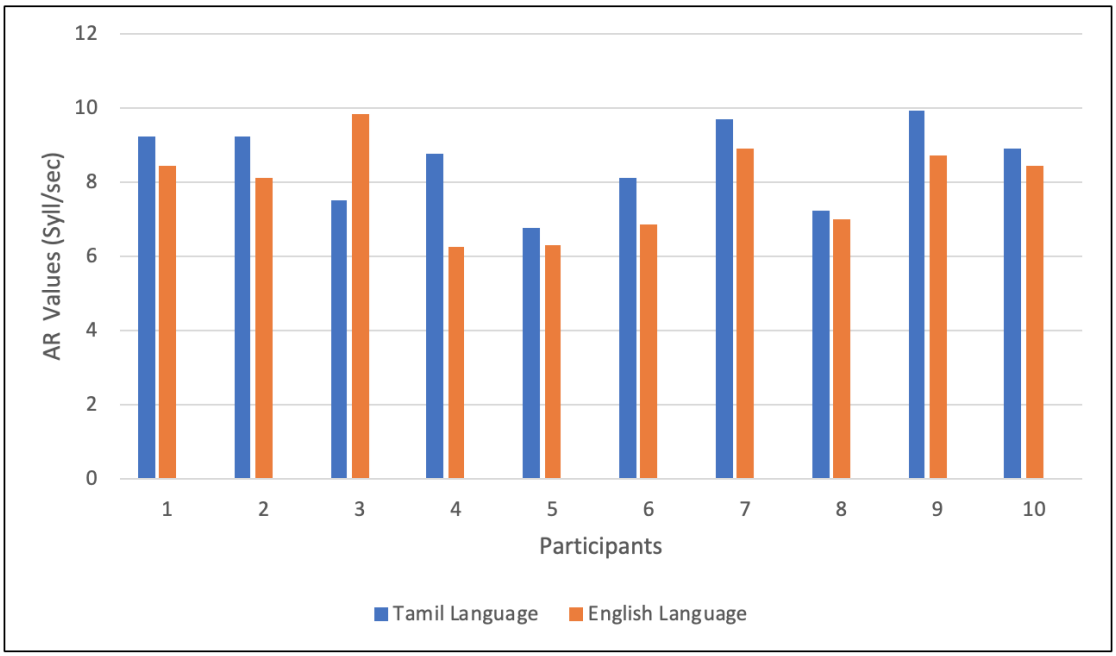

Fig. 5.7 Values of the Articulation Rate for All Audience –Members in English and Tamil as Calculated

Audience –member 7 is the fastest speaker with the highest speech rate in both English and native language, but not the articulation rate in both languages. However, audience –member 7 has the highest AR in English and audience audience –member 3 has the highest AR value , but not SR. It was observed that AR and SR are not dependent on each other in languages.

Estimation/Calculation of Pauses for Unscripted speech

Each participant had some pauses while they were reading the script. Some pauses would be silent (where they did not say anything in the speech) and some are filled such as: um, mm, repetition, prolongation, etc.

1. Silence is where the participants did not say anything, they just breathed in and out.
2. Repetition is where participants said/repeated a word more than one time .
3. Prolongation is where participants extended a word bit longer than it was supposed to be
4. Insertion is where participants said new utterances, such as um, mm. oh etc during the speech.

The number of pauses, duration of pauses and types of pause in English were noted in a Table 5.9 and the native which is Tamil language was noted in Table 5.10.

Table 5.9 Observation of Pauses and with Their Types for Audience–Member 1 (English Language Unscripted Speech)

Pauses	Type of Pause	Duration of a Pause (msec)
Pause 1	Silent	380
Pause 2	Silent	310
Pause 3 [Filled]	Repetition	280
Pause 4 [Filled]	Prolongation	410
Pause 5	Silent	460
Pause 6 [Filled]	Insertion	270
Pause 7	Silent	350

Table 5.9 continued from previous page

Pause 8 [Filled]	Prolongation	540
Pause 9	Silent	320
Pause 10 [Filled]	Insertion	230
Pause 11	Silent	180
Pause 12	Silent	330
Pause 13 [Filled]	Repetition	206
Pause 14 [Filled]	Repetition	268
Pause 15 [Filled]	Insertion	140

Table 5.10 Observation of pauses and with their types for audience –member 1 (Tamil Language Unscripted Speech

Pauses	Type of Pause	Duration of a pause (msec)
Pause 1	Silent	360
Pause 2 [Filled]	Repetition	190
Pause 3	Silent	200
Pause 4	Silent	240
Pause 5 [Filled]	Insertion	350
Pause 6	Silent	460
Pause 7 [Filled]	Insertion	230
Pause 8 [Filled]	Prolongation	250
Pause 9 [Filled]	Insertion	520
Pause 10	Silent	150

Volume for All the Audience –Members (Unscripted Speech)

The mean intensity was calculated for all participants using Praat software through and the mean intensity shown in Figure 5.8.

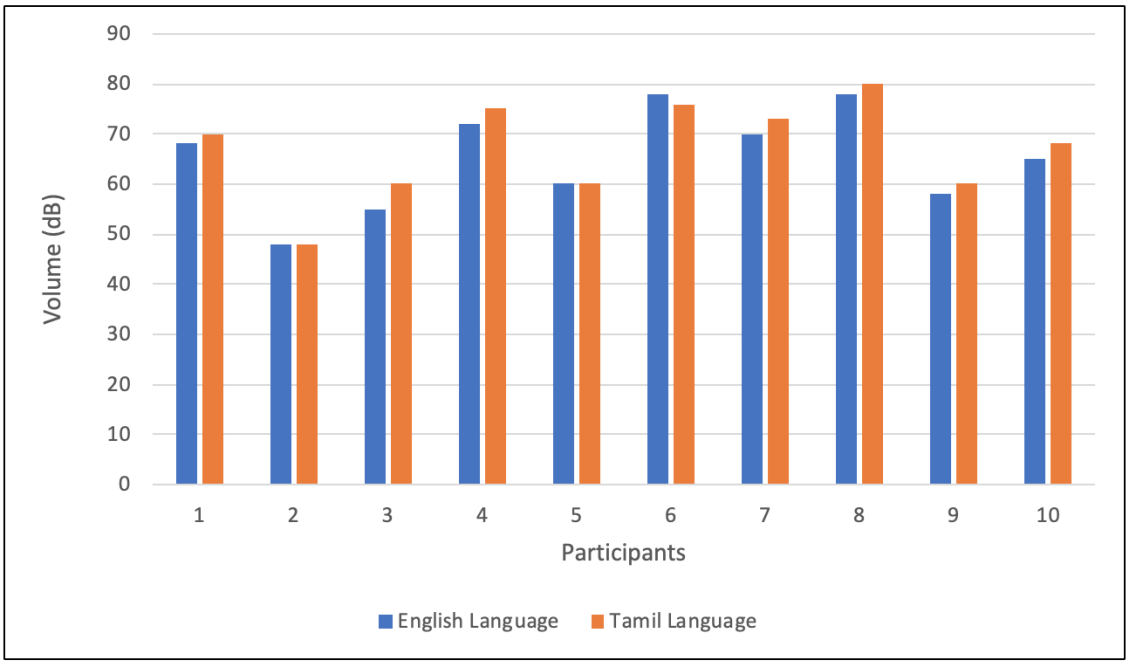

Fig. 5.8 Mean Volume for the Audience –Members in Experiment 2

Audience member 8 was recorded highest volume in both English and Native languages. Audience member 6 was recorded second highest volume in both languages. Participants maintained the same volume in both languages.

Table 5.11 Mean Attributes for All the Audience –Members

Voice Characteristics	Participants (English and Native Language)									
	1	2	3	4	5	6	7	8	9	10
Frequency (Hz)	184.16 185	282.66 284.66	190.83 192.16	213.16 213.5	117.5 118.5	152.33 153.83	114.5 115.33	237.5 239.5	179.33 180.16	124.48 125.66
Volume (dB)	68 70	48 48	55 60	72 75	60 60	78 76	70 73	78 80	58 60	65 68
Speech Rate (Syll./sec)	6.1 7.2	6.3 6.7	6.2 7.6	5.5 6.4	5.2 5.4	6.3 6.4	7.4 8.2	5.5 6.5	7.1 7.4	6.5 6.8
Articulation Rate (Syll./sec)	8.45 9.21	8.12 9.21	7.52 9.82	6.24 8.75	6.31 6.75	6.87 8.12	8.92 9.67	6.99 7.21	8.72 9.92	8.45 8.89
Number of Pauses	15 10	18 10	15 08	16 09	17 09	22 12	20 14	14 07	18 08	22 10
Pause Percentage %	7.7 4.9	8.3 5.3	11.2 7.1	5.2 2.6	8.7 6.1	5.8 3.8	10.3 7.6	9.6 4.8	10.0 7.0	7.9 4.2
Filled Pause Percentage %	53 50	38 30	60 31	62 55	47 44	27 16	40 35	35 42	55 50	40 40
Total Pause Time Sec. /Minute)	4.67 2.95	5.01 3.23	6.75 4.21	3.12 1.56	5.23 3.67	3.51 2.32	6.23 4.57	5.76 2.89	6.01 4.25	4.76 2.52

The mean fundamental frequency was nearly the same for both languages. However, there was a near difference of 2 % between the languages, because members are more comfortable in Tamil than the English language. Members have maintained the same volume, in both languages. Frequency and volume are independent of the languages used. SR and AR values are faster in the native language. However, the fastest participant who has the highest SR values got the lowest AR values. So, SR and AR values are not dependent on each other. For example, a person can have faster SR values but, it does not mean that the same person will have faster AR values. There is no significant difference in terms of frequency, volume, SR, and AR values. However, there is a significant difference in pauses when the audience talked in both languages. There were fewer pauses in native languages cause the audience was more comfortable when they talk in their native language than in the English language. It was observed that pauses are dependent on the language used.

The observations were made, 5.3.1 and 5.3.2, Frequency and volume did not change when participants and members talk in different languages in both scripted and unscripted languages. A number of pauses were less in scripted speech since participants had the script before recording takes place. Participants practiced well enough and they knew what was to bespoke in the recording place. Whereas unscripted members did not have any scripted form to talk, they had to think and then talk, so members had more pauses. SR and AR values both were higher in familiar language/native language when compared to the English language. Pauses are significantly different between scripted and unscripted speech and they will vary, depending upon the situation and topic of conversation.

5.4 Results

Table 5.5 and Table 5.8, the range of frequencies can be used to eliminate the person whose frequency range falls out of any observed readings. It was observed, concluded that participant energy levels vary $\pm$ 3dB when participant talks in their native language. The reason would be, humans are very comfortable with their native language than the learned language.

According to the statistical analysis, in the first experiment, in English, the member with the highest speech rate also had the highest articulation rate, as shown in Figure 5.6 and 5.7. However, a member with the slowest speech rate did not always have the lowest articulation rate. In Tamil, the member with the fastest speech rate had the second-highest articulation rate. The results indicated that speech rate is more prominent and useful to identify a speaker when compared with the articulation rate. However, AR is useful in identifying the slowest and fastest speakers.

AR depends on the variation of phonemes in the words spoken. Whereas, SR mainly depends on the speaker and the situation such as: what is the topic and their opinion on it. The results showed that one couldn't make or assume a direct relationship between AR and SR when a speaker changes the language of speech. However, we can use a combination of speech rate and articulation rate to eliminate a person we are not looking for.

5.5 Summary

In this chapter, two types of voice samples have been collected. For the first experiment, the data was collected from the university staff and students. while for the second experiment, the data was collected from a conversation between two groups of people who have been randomly asked to talk. There are several differences in the second collection of data when compared to the first one. Firstly, data was a collection of argumentation, which includes debating, laughter, pauses, and words like 'hmm', 'mmm', 'aha' etc. Secondly, the data is technical of poor quality as it is recorded in a normal room and using different types of handsets. Finally, there are lots of settings in the room to record speaker voices for a longer-term.

"unvoiced" In the first experiment, a database of 10 participants voice samples was collected whose native language is not the English language, consisting of university students. In the second experiment, a database of 10 audience-members voice samples was downloaded from a TV program and whose native language is not the English language, consisting of random people.

There are other factors, that the system needs to take into consideration to identify a speaker when it comes to different languages. The English language is not phonetical, letters do not sound the same in all cases. For example, "heir" and "hire" contain the same letters, but in the first one, the "h" is silent, whereas in the second one it is not. Even more confusingly, words such as "cow" and "bow" can be pronounced multiple ways even with the same letters. While words such as "no" and "know" sound the same even though they are spelled differently. Individual letters do not sound the same as the k & w are silent in this case.

Chapter 6

Phonemes: An Explanatory Study Applied to Identify a Speaker

6.1 Introduction

Speaker Identification (SI) is a process of identifying a speaker automatically via a machine using the speaker's voice. In SI, one speaker's voice is compared with the n- number of speakers' templates within the reference database to find the best match among the potential candidates. Speakers are capable of changing their voice, though, such as their accent, which makes it more challenging to identify who is talking [99, 100]. In this chapter, the phonemes from a speaker's voice were extracted and investigated with the associated frequencies and amplitudes used to assist in identifying the person who is speaking. This chapter demonstrates the importance of phonemes in both speech and voice recognition systems. The results demonstrate that one can use phonemes to help the machine identify a particular speaker, however, phonemes get better accuracy in speech recognition than speaker identification.

Digital systems need to be given training data, which consists of speech samples to identify a speaker. These speech samples are collected from each person speaking into a microphone and processed by a computer to recognise the voice/speech. Voice characteristics include both physical and behavioral components. The shape of the vocal tract is fundamental in the physiological component. The vocal tract is made up of the mouth, tongue, jaw, pharynx, and larynx which articulate and control speech production by manipulating the airflow generated by the lungs and diaphragm. The behavioral component comprises emotion, accents, rate of speech, and pronunciation. Some elements of speech, such as the ability to roll the letter 'r' are controlled genetically.

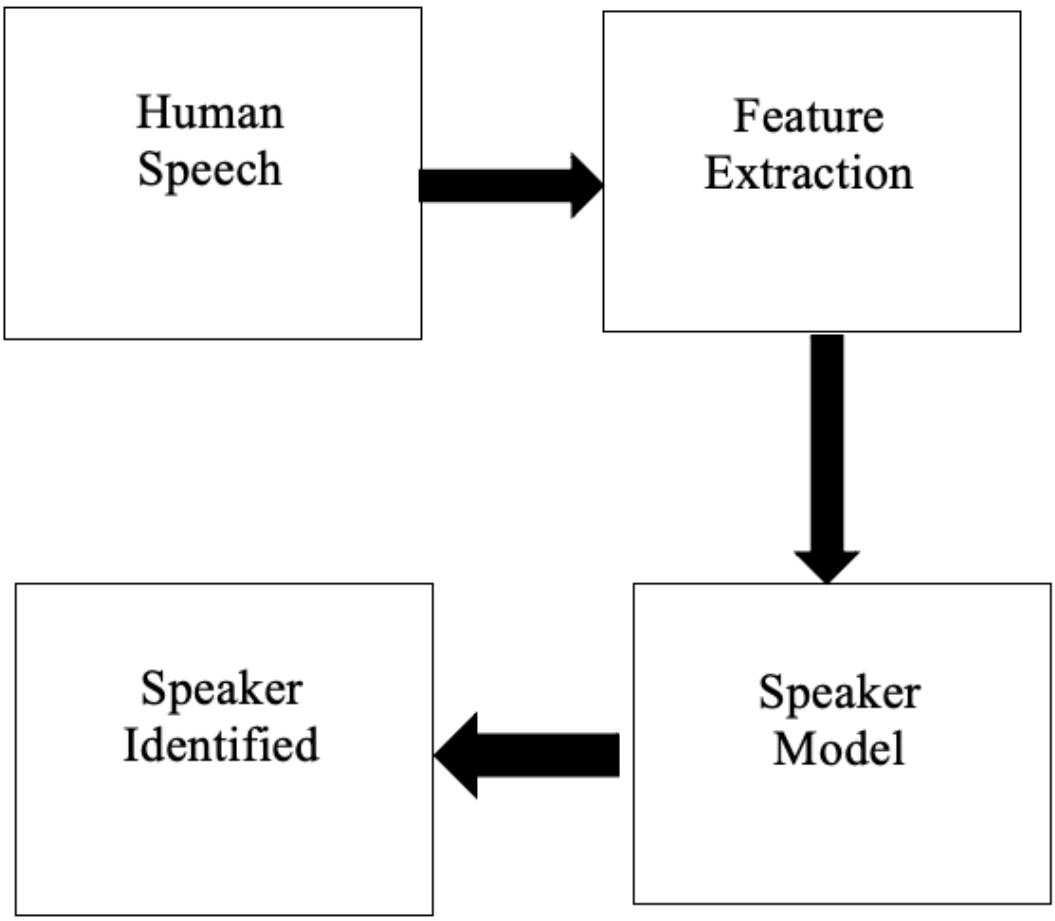

Fig. 6.1 Speaker Identification System

This chapter investigated the differences in the frequencies of phonemes. Hence, an experiment was conducted, which includes collecting voice samples from ten participants and extracted phonemes. This chapter is organized as follows, section 6.2 presents a brief overview of the background of the speaker identification system, followed by results, discussion, and given the conclusion of using phonemes to identify a speaker.

6.2 Background

The sounds of human speech are complex and have been studied for centuries and are still being researched. Research suggests that phonetics has always been an important part of sound production. Phonetics is derived from a Greek word, 'phōnētikós'; where phone means a sound or voice. The small units of sounds are called phonemes, with each language

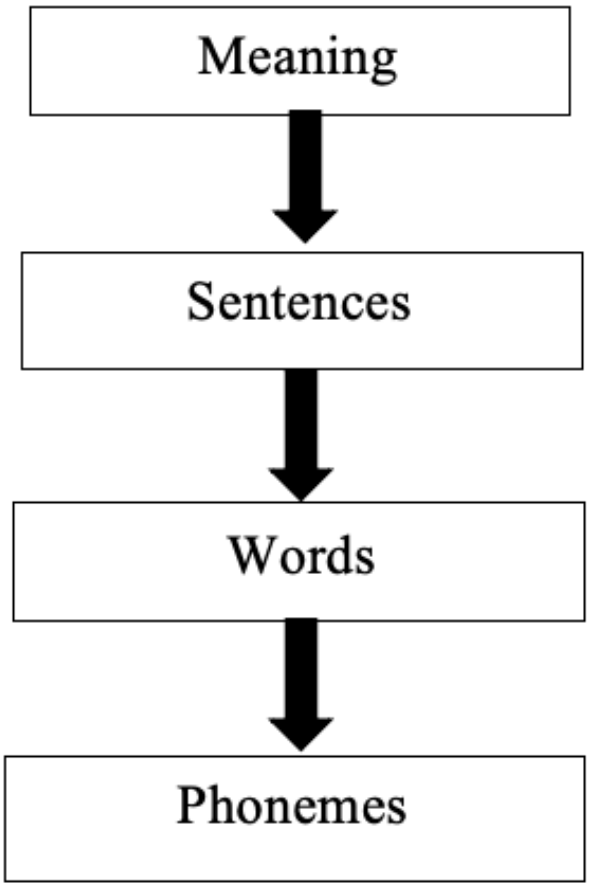

Fig. 6.2 Phonemes are the Basic Building Blocks of Spoken Language

having its own phonetic set [101, 102]. Phonetics has played the main role in learning and understanding a language rather than identifying a speaker. There are 20 letters that are considered to be "voiced," which, in English, include consonants B, D, G, J, L, M, N, NG, R, SZ, TH, V, W, Y, Z, and vowels A, E, I, O and U. There are 8 "unvoiced" sounds: CH, F, K, P, S, SH, T and TH [103].

There are three types of phonetics: acoustic, auditory, and articulatory phonetics. Acoustic phonetics is the physical property of the sounds of a language; that is the volume of sound, frequency of the sound waves, frequency of vibrations, etc. Auditory phonetics is focused on how speakers perceive the sounds of a language, with the help of the ears and the brain. Articulatory phonetics conveys how the vocal tract produces the sounds of a language that is, with the help of moving parts of our mouth and throat, also known as the articulators [104, 105].

Phonetics helps when learning and distinguishing within a language, or between multiple languages [106]. By uttering a sequence of discrete sounds (or phonemes) with the help of our articulators, words are composed. A combination of coherent words leads to a sentence. Phonemes are discrete or different sounds within a particular language, but make up the building blocks of all speech. Thus, all words and sentences are ultimately collections of phonemes [107, 108].

Feature extraction plays a crucial part in speech processing. Features should provide the necessary information to be able to identify a speaker. There are numerous feature extraction methods are available such as: Linear Predictive Codes (LPC), Perceptual Linear Prediction (PLP), Mel Frequency Cepstral Coefficients (MFCC), PLP-RASTA (PLP-Relative Spectra), etc. The most popular feature extraction method is MFCC, but extraction features would be difficult when speaker changes their voice such as: their emotional state, context, with whom they are talking etc[109]. MFCC does not provide enough resolution in frequency regions and a signal can not be reverted from frequency analysis by using MFCC [110].

Phonemes differ across languages; the frequency of the sounds varies in which they occur in words. Some phonemes may not be considered as phonemes in other languages [112, 113]. For example, the Chinese language is tonal, and sounds come from nasal cavities when compared with English. The features will vary while patterns of sound also differ significantly in different languages. The fundamental frequency of “r” is the same for two British speakers. On the other hand, the way of pronouncing r can be used to distinguish between British and non-native speakers.

6.3 Methodology

Phonemes were extracted from a human speech in this chapter. Each phoneme’s amplitude and frequency values were measured and evaluated. There are two tasks in this experiment. In the first task, participants were asked to read phonemes on their own 6 times. For example, participants need to say /p/ 6 times, /b/ 6 times and so on. In task 2, participants were asked to read a list of words that covered phonemes pronounced in task 1. The words were designed to read easily by all participants and prevented the use of foul language as well.

This experiment hypothesized that phonemes would be individual to a speaker, and then one can use phonemes to identify a speaker-independent of a language. Once the experiment was conducted and results were observed, one can use phonemes to identify a speaker, with some limits, as explained in the conclusion.

6.3.1 Task 1

Participants were to asked to read each of the following phonemes on their own.

1. /b/ - **b**ad
2. /p/ - **p**et
3. /th/ - **th**ick
4. /TH/ - **th**is
5. /n/ - pi**n**
6. /ng/ - si**ng**
7. /r/ - th**r**ee
8. /t/ - pa**tt**e
9. /l/ - **l**ong
10. /f/ - **f**our

Measuring Frequency Spectrum of Phonemes

Fast Fourier Transform (FFT), was applied to the voice signal to observe the frequency spectrum. The FFT was applied to phonemes of all 10 participants.

Table 6.1 Fundamental Frequency and Duration of Phoneme \b\ of Participant 1

Mean Freq (Hz)	**Median Freq (Hz)**	**Min Freq (Hz)**	**Max Freq (Hz)**	**Duration (msec)**
255.45	247.94	188.65	370.35	532.5
256.80	249.57	187.92	408.72	591.6
249.14	239.55	188.26	338.37	562.1
253.55	234.92	189.52	353.55	543.7
252.10	239.58	192.64	375.30	539.9
256.50	249.60	186.45	409.23	590.2
253.92	**243.52**	**188.90**	**437.58**	**560**

Table 6.2 Volume of a Phoneme \b\ of a Participant 1

Phoneme \b	**Mean (dB)**	**Min (dB)**	**Max (dB)**
1	68	38	72
2	68	45	71
3	69	49	72
4	66	33	68
5	65	33	69
6	67	34	72

Table 6.3 Fundamental Frequency and Duration of a Phoneme \p\ of Participant 1

Mean Freq (Hz)	**Median Freq (Hz)**	**Min Freq (Hz)**	**Max Freq (Hz)**	**Duration (msec)**
283.26	273.31	254.26	334.09	**473.3**
278.93	269.95	247.12	358.23	**414.1**
264.88	253.23	243.36	325.60	**508.9**
271.15	259.38	241.35	347.53	**443.7**
265.55	250.36	241.86	310.39	**532.5**
245.56	253.15	187.08	290.98	**443.1**
268.22	**259.89**	**235.83**	**327.80**	**469.26**

Table 6.4 Volume of a Phoneme \p\ of a Participant 1

Phoneme \p	**Mean (dB)**	**Min (dB)**	**Max (dB)**
1	65	34	68
2	64	34	69
3	64	36	68
4	64	30	70
5	67	32	70
6	66	33	70

Table 6.5 Fundamental Frequency and Duration a Phoneme \p\ of Participant 2

Mean Freq (Hz)	Median Freq (Hz)	Min Freq (Hz)	Max Freq (Hz)	Duration (msec)
300.42	281.63	277.51	369.83	**528.4**
268.32	263.84	246.22	299.20	**408.4**
256.53	252.15	232.20	290.93	**427.8**
252.05	242.86	231.94	292.78	**538.5**
256.67	244.92	233.51	298.43	**408.4**
226.14	218.46	207.42	267.46	**427.8**
260.02	**250.64**	**238.2**	**303.10**	**456.55**

6.3.2 Task 2

The participants were asked to read the following words:

1. Beep, peep and keep
2. Cat, bat and pat
3. Pad, bad and dad
4. Thick and thin
5. Than, ban and can
6. Cot and dot
7. Pin and spin
8. Fine and pine
9. Sing and ring
10. Tall and ball

After recording the voices of all participants, the next step was to extract phonemes from a script. The phonemes from a speech were extracted manually by a researcher. The observation of, how the frequency and relative amplitude values changed for a specific phoneme. Fast Fourier Transform (FFT), was applied to the voice signal to observe the frequency spectrum. The FFT was applied to the phonemes of all 10 participants.

Measuring Frequency Spectrum of Phonemes

Phonemes play an important role in human speech. Phonemes help us to recognise the sound, in a speaker's language. But all phonemes may not sound similar to how they sound in other words. In this experiment, the frequency of phonemes in words has been observed at various points, such as the position of a phoneme in several words. In this experiment, participants were asked to read phonemes on their own and the words pronounced have been mentioned in the 6.3.2.

Table 6.6 Fundamental Frequency and Duration of an Extracted Phoneme \b\ of Participant 1

\b\	Mean Freq (Hz)	Median Freq (Hz)	Min Freq (Hz)	Max Freq (Hz)	Duration (msec)
Beep	210.78	212.22	182.88	251.59	527.7
Bat	202.05	197.94	181.42	262.30	517.3
Bad	205.60	208.44	182.60	227.63	523.8
Ball	211.97	208.79	182.38	251.27	518.2
Average	**207.6**	**206.84**	**182.32**	**248.19**	**521.75**

Table 6.7 Volume of an Extracted Phoneme \b\ of Participant 1

Phoneme \b\	Mean (dB)	Min (dB)	Max (dB)
1	69	38	72
2	68	45	72
3	68	49	72
4	67	34	70
5	68	38	70
6	69	45	72

Table 6.8 Fundamental Frequency and Duration of an Extracted Phoneme \p\ of Participant 1

\p\	Mean Freq (Hz)	Median Freq (Hz)	Min Freq (Hz)	Max Freq (Hz)	Duration (msec)
Peep	265.46	257.53	254.60	317.68	**137.3**
Pat	241.25	240.91	237.01	246.26	**129.5**
Pad	233.94	231.81	227.64	245.98	**119.1**
Average	**246.88**	**243.41**	**239.75**	**269.97**	**128.63**

Table 6.9 Fundamental Frequency and Duration of an Extracted \th\ of Participant 1

/th/	Mean Freq (Hz)	Median Freq (Hz)	Min Freq (Hz)	Max freq (Hz)	Duration (msec)
Thick	262.08	250.73	243.30	297.46	**145.9**
Thin	250.73	262.08	245.34	297.15	**158.9**
Average	**256.40**	**256.40**	**244.32**	**297.30**	**152.4**

Table 6.10 Fundamental Frequency and Duration of an Extracted Phoneme \b\ of Participant 2

\b\	Mean Freq (Hz)	Median Freq (Hz)	Min Freq (Hz)	Max Freq (Hz)	Duration (msec)
Beep	231.98	239.64	201.84	230.98	125.3
Bat	221.37	223.98	203.94	232.94	167.1
Bad	220.34	220.35	205.97	235.58	188.2
Ball	222.97	230.46	195.01	242.22	229.7
Average	**224.16**	**228.60**	**201.69**	**235.43**	**177.57**

Table 6.11 Fundamental Frequency and Duration of an Extracted Phoneme \p\ of Participant 2

\b\	Mean Freq (Hz)	Median Freq (Hz)	Min Freq (Hz)	Max Freq (Hz)	Duration (msec)
Peep	260.72	257.87	254.89	270.38	**248.8**
Keep	222.68	217.32	194.17	251.32	**256.5**
Pad	263.46	259.01	253.53	281.59	**197.6**
Pin	298.38	297.27	295.33	302.80	**208.8**
Average	**261.31**	**257.86**	**249.48**	**276.52**	**227.92**

The fundamental frequency, volume and, duration of the phonemes have been observed for all the participants. The mean fundamental frequencies in Table 6.1 and in Table 6.6 are not same and there is a difference of 40 %. But, their minimum frequency is nearly the same and their frequencies lie within the range as shown in the Tables. The duration of their phonemes is also not consistent enough. The duration was long when participants pronounced phonemes by themselves, and the duration was less when they pronounced phonemes in words. The frequency range of phoneme p and th was the same when participants pronounced

phoneme on their own and extracted phonemes from the words . So, one can use phonemes p' and th' to identify a participant 1. The volume of the phoneme in Table 6.2 and Table 6.7, are nearly the same. On the other hand, in participant 2, the phonemes p frequency range was consistent enough in both extracted and individual phoneme pronunciation.

Although, their mean fundamental frequency values were changed drastically in 6.3.1 and 6.3.2, which were noted. The hypothesis of this experiment suggested that phonemes would be individual to a speaker, so they could, in turn, be used to identify a speaker-independent of a language. But, after the experiment was conducted and results were observed, it was concluded that phonemes can be used to identify a speaker but with some boundaries, as explained in the conclusion.

People will pronounce some words in a specific way that helps us to identify their origin such as native or non-native or their ethnicity. Native speakers will have a distinguished accent when compared to non-native speakers. For example, the pronunciation of the word "The" can be used to distinguish those who are not British. phonemes can be used to identify those who have a British accent too. On the other side, an accent can also be used to distinguish speakers from the north or south or east or west location of their country. For example, in England, one can easily tell the difference between a speaker living outside or inside London based on their accent, which is purely dependent upon phonemes.

Table 6.12 Attributes of All the Participants

Voice Characteristics	**Extracted phonemes \b\ from the words**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Mean Freq (Hz)	207.6	224.16	247.5	246.5	360.7	417.2	383.2	394.2	372.2	375
Median Freq (Hz)	211.2	234.5	260	242	339	387.5	363	403.5	380	365
Volume (dB)	70	68	65	55	67	72	78	70	69	72
Duration (msec)	524.3	226.4	187.8	202.8	156.8	190.6	202.4	167.9	201.6	247.9

The voiced phonemes are extracted from participants and FFT was applied to observe how relative amplitude and frequency values of a phoneme vary for different words from the same participant. Each phoneme represented a different visual representation of the phonemes of a participant. Once the voiced phonemes of one participant are compared with another participant, it could be observed that some phonemes were very similar to others and some of them were very distinctive. The frequency and relative amplitude values were derived and recorded, from each phoneme.

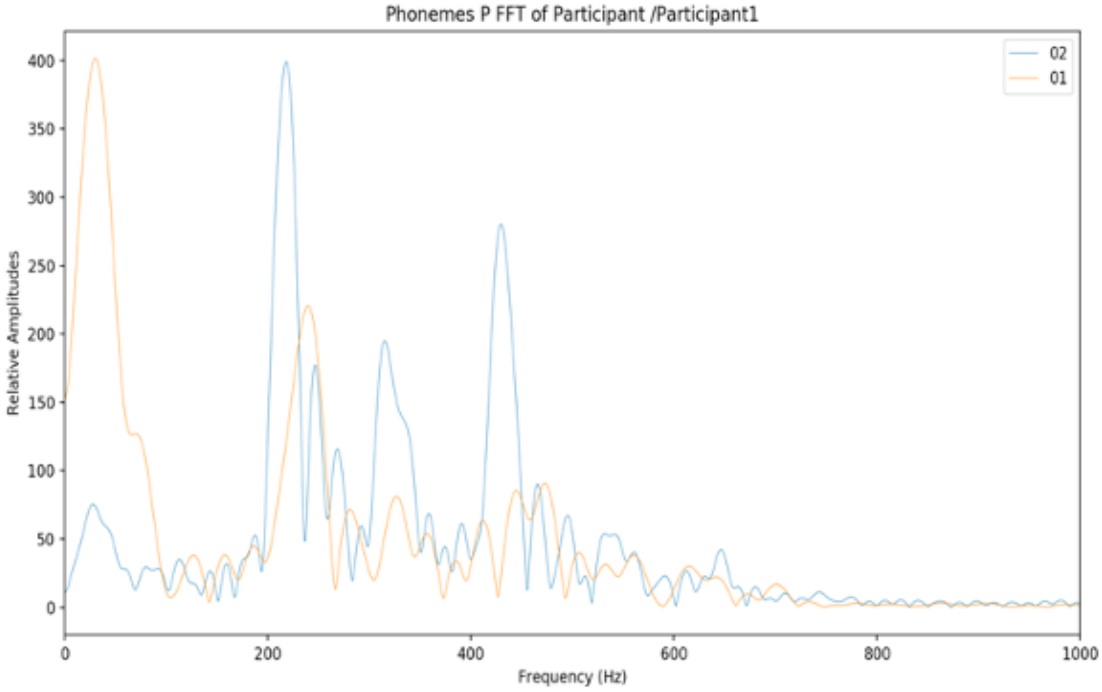

Fig. 6.3 Spectrograph of Phoneme ‘p’ of a Participant 1

Next, voiceless phonemes of all participants were extracted to find out if there was any consistency, to identify a speaker. According to the results, the voiceless phonemes of some participants were sufficiently distinctive to recognize a person. FFT graphs were prepared for both phonemes of all participants and voiced versus voiceless phonemes are compared to conclude.

Parameters of the voice have been observed in both individual phonemes and extracted phonemes from the words. So a question arises here that, is there any pattern in which the combination of individual phonemes frequencies and the frequency of phonemes of a word? The average frequencies of an individual phoneme are equal to a word. For example, frequencies of individual phonemes \b\, \ee\ and \p\ for participant 1 is equal to when the participant pronounces the word "Beep".

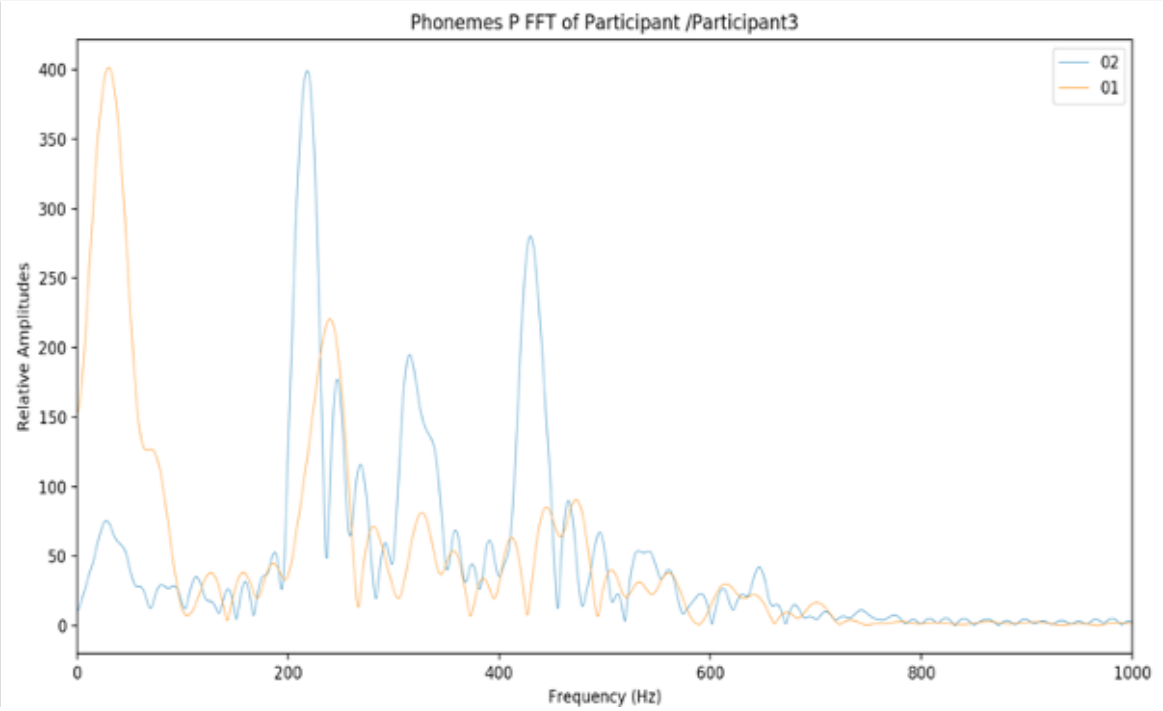

Fig. 6.4 Spectrograph of Phoneme ‘p’ of a Participant 3

\b\+ \ee\+ \p\= {255 + 317 +274}/3
= 282
= Beep

Frequencies of individual phonemes have been added and their average frequency value is equal to the frequency of all phonemes together.

On average, a person can speak 100 words per minute and each word would have four phonemes. In this case, an average person can produce 400 sounds per minute. Each person has a unique way of pronouncing some phonemes, as explained earlier in this chapter. However, the problem is, extracting phonemes manually would be difficult. since one needs to listen to their speech carefully and then extract them. It would be interesting if one can develop an algorithm that can automatically extract phoneme sounds from a participant, and then only phonemes can be observed rather than observing entire words or sentences to identify a person. In this way, one can identify a person with fewer data and within a limited time.

Another interesting factor would be, people from different geographic locations pronounce few words in a very distinctive way. For example, people from /northern Ireland have a very peculiar voice. Their voice is very thick and they emphasise the last letter of the words. So, if one concentrates only on those phonemes, one can easily eliminate a person

who is not from the north for example Belfast. It was concluded that phonemes would be useful to identify a person's geographical location or can be used to identify a person with limited voice samples.

6.4 Results

The voices of participants are used as an initial data set and their phonemes were extracted. Participant 3 and participant 4 have the same similarity when they pronounce the letter "P" as shown in Figure 6.3 and 6.4. On the other hand, participant 6 and participant 9 have a high similarity of producing phonemes "r". Participant 4 is similar to participant 9 when pronouncing the phoneme "th". Lastly, participant 5 is the only one with a distinctive pronunciation of the phoneme "S". Participant 3 and 4 are from Egypt, their pronunciation of the phoneme "p" would be helpful to recognise their nativity. The same phoneme "p" can be used to differentiate German and French speakers.

There are several factors, which make a phoneme sound different and represent different relative amplitude and frequency values. For each participant, a range is set up for dominant frequency, independent of phonemes, meaning he/she can say any phoneme but, the dominant frequency should lie between the range. For several phonemes, like b, n, P, r, and TH, the dominant frequency lies between 245 to 390 Hz. For the phoneme T and V, the dominant frequency lies between 200 to 285 Hz. It was observed that the dominant frequency of phoneme "p" of participants 3 and 4 are the same. The frequency of "S" of participant 1 and 2 are the same but differs in amplitude values. The highest peak of participant 1 of "S" is the same as "w" of participant 2. However, the boundaries of phonemes vary among the languages.

In our daily conversation, a listener can recognize or concentrate on words to understand the meaning, which helps us, to communicate with each other. For example, if someone is continuously saying b, b, b, ... several times and say p 20 times in between and then continue saying b, as humans don't recognize the 'p's and perceive as if the participant said 'b' only.

6.5 Summary

A database of 10 participants' voice samples was collected, consisting of university students and friends. When one considers the Chinese language, it is a tonal language. The method of expressing phonemes would be different to convey the message/information. After observing the data, it is concluded that phonemes will not help us identify a speaker, but instead help

us find out their nativity. Phonemes can play an important role in the linguistic theory of speech. One of the main problems with phonemes is that participants had an influence from their native language on the other familiar language (English) such as; participants can pronounce differently or mispronounce, phonemes in words when they talk in their first/native language. They tend to use their native language phonology skill in other languages that help us recognize their nativity. It would be helpful to understand the language, so one can use it in speech recognition and language identification. Phonemes play an important role to identify non-native speakers. For example, native English speakers can recognise non-native English speakers and non-native English speakers can detect the nationality of another non-native English speaker.

Some of the other factors included are the actual placement of the phoneme in a word; emotions can alter the phonetic emphasis on a word and the context of the word (paint and pain/ sell and cell).

Different letters or combinations of letters may represent the same sound. One should either have complete knowledge of phonemes or they can make use of International Phonetic Alphabet (IPA) to find the phonemes in a word. A letter can produce two separate sounds in two different words. A certain combination of letters. Several words carry sounds of letters that are not present in the word; For example, the word ‘exert’ does not carry the letter ‘Z’ but it is prominent in the word regardless. Phonemes change their frequency based on their position that is in the starting, middle, or end of the sentences. The spectral analysis showed that participant information is non-uniformly distributed. Some of the frequency domains clearly show the differences to be able to identify a speaker. However, the setback is, how can one decide the frequency bands for an individual when other participants also have the same differences, for example, phoneme ‘p’ is nearly the same as shown in Figure 6.3.

Participants have used knowledge of phonemes from their original language that helps us identify their nationality. It is difficult to extract a phoneme, if one does not observe or listen carefully, for example, /p/ in cap and /b/ in a cab. If a system is trained based on phonemes only, without context/situation the system cannot figure out which phoneme is pronounced.

Chapter 7

Applications of Speaker Identification for Universal Access

7.1 Introduction

Speaker Identification is the process through which a machine automatically identifies the speaker, based solely on the voice of the speaker. It is interesting how an individual can recognise a person out of sight, simply their voice, which makes the research challenging. Voice of human bio-metric property and recognition of a particular person's voice can be used in different applications such as: unlocking an office door, marking student or employee attendance, monitoring elderly people's health, online banking services, or helping people with dementia to be able to identify a who is speaking [113]. This chapter explores a range of such applications and discusses how emerging technologies can be used to support a variety of users in a series of different contexts of use.

7.2 Voice Recognition

Verbal speech is one of the most common forms of communication. It uses words to convey information to others. The nature of the sounds produced as part of that speech assists in the identification of the speaker, but it is not always straightforward to recognize the identity of the person. Verbal speech constitutes a speaker's accent, speaking style, and pronunciation, etc., but our ability to recognize who is speaking can be affected by background noise, the emotional state of the person, and even something as simple as whether they have a common cold or a blocked nose [114]. Typically, like people, can identify a speaker with a comparatively high degree of accuracy if they are sufficiently familiar to us. Humans use a

combination of parameters to identify a person such as speaking accent, speaking style, and pronunciation, etc. However, humans may not always use the same parameters every time. If someone has a particularly distinctive characteristic, such as a lisp or a very particular way of saying a common word, humans can use that information to speed up the process of recognizing the speaker [115]. However, training a computer to use similar parameters and shortcuts, to identify a speaker, is not easy. Consequently, one needs to explore more systematic approaches to recognise the speaker.

Automated voice and speaker recognition is typically a two-step process: speaker identification and speaker verification. Speaker identification is the task of identifying who a speaker is from a field of possible candidates, usually by either trying to match the speaker to the closest stored speech template or by trying to iteratively eliminate possible candidates until only a single potential candidate remains [116]. This chapter will discuss the different types of applications that can be used in speaker identification, which is the task of identifying who a speaker is.

Speaker identification can be further subdivided into two types, namely: closed set and open set. The closed set is where the possible range of candidate speakers is defined and the voice to be recognized is from within that set. The open set includes the possibility that the speaker may not be from within the existing set of stored speaker templates.

It is possible to subdivide the recognition one step further into text-dependent or text-independent. Text-dependent is where the speaker is uttering a known (defined) phrase or set of phrases, whereas text-independent is any possible spoken content. Figure 7.1 summarizes these different approaches to identifying the particular person who is speaking.

7.3 Example of Applications of Speaker Recognition in Use

The purpose of this chapter is to explore a range of different potential uses of this new technology to examine how it can support the principles of universal access. Universal Access is typically taken to focus on addressing the needs of those with some form of impairment or functional limitation that may either be innate, e.g. arising from a medical condition or injury, or situational, i.e. where the impairment or limitation arises from the circumstances that someone finds themselves in. Examples of the latter can include while driving or in a very busy environment.

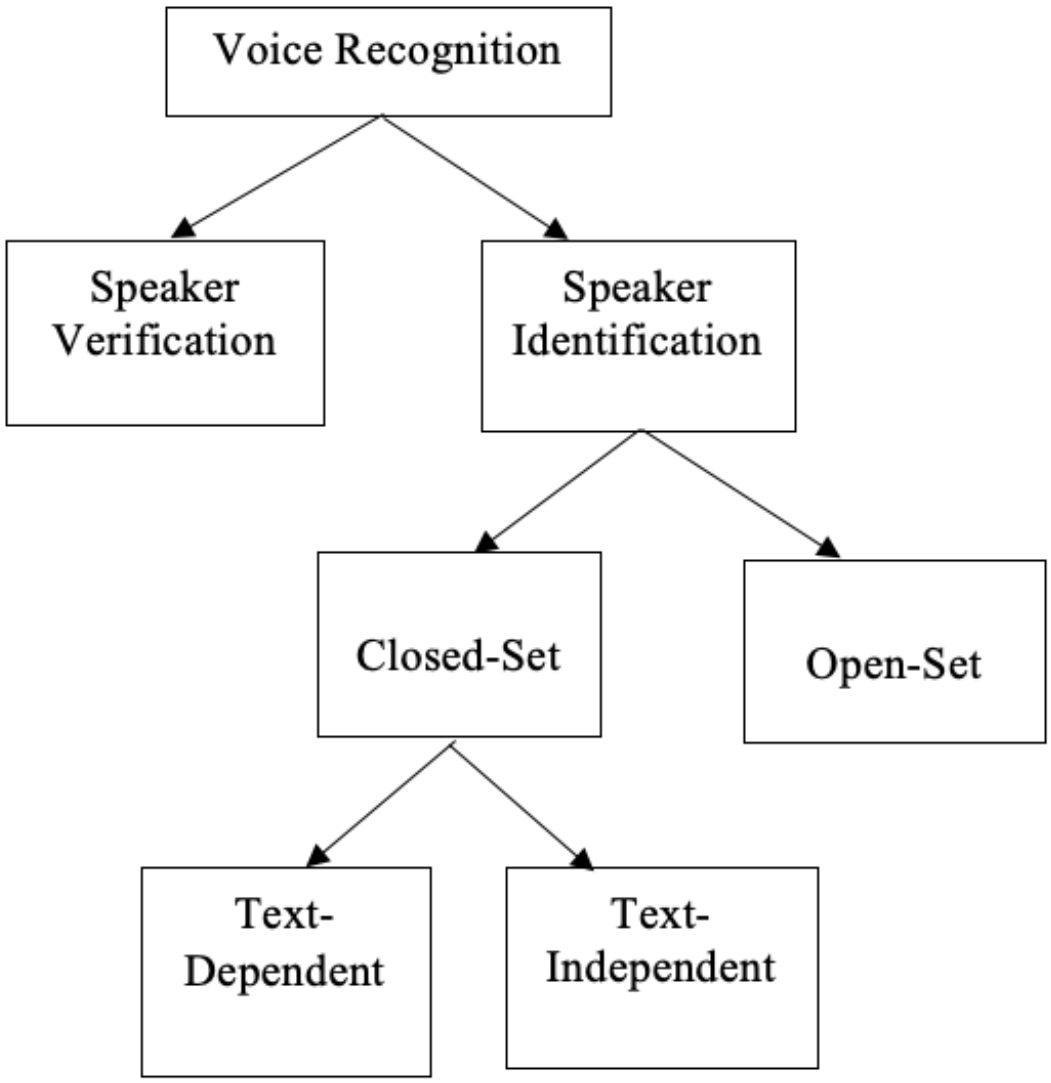

Fig. 7.1 Block Diagram of a Different Approach to Speaker Recognition Systems

7.3.1 Security Application

The most commonly cited example of speaker identification is its use for identifying a particular person for security reasons, such as for telephone banking. Typically a person is enrolled for this security service and is asked to utter a particular phrase. The person's utterance is then compared with a stored reference, which is an example of closed-set text-dependent recognition. If the match is within a predetermined threshold, the caller is permitted to access the banking services. If the match is not made, the caller is passed to an alternative authentication service, as shown in Figure 7.2.

Once it has been demonstrated that the ability to use speaker recognition meets the necessary security requirements in terms of recognizing an individual uniquely, it would mean that someone who finds it difficult to remember PINs and passcodes may access such services more easily [117].

Beyond the obvious banking applications, it is straightforward to think of situations where someone who is older or has severe functional impairments may wish to benefit from such technology, such as additional security at home. For instance, a door locking can be used to increase the physical security of a potentially vulnerable person, where both a key and a stored passphrase are used by a caregiver or cleaner to enter the house.

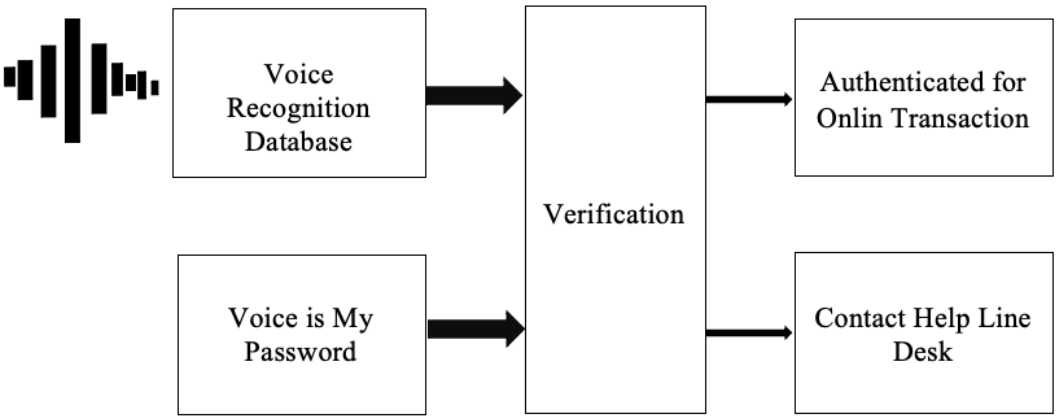

Fig. 7.2 Speaker Authentication for Telephone Banking Services

Similarly, there have been numerous newspaper reports over the years of scammers targeting potential victims by posing as representatives of financial institutions, such as banks, to obtain log-on information to banking services[118]. A layer of speaker recognition technology would render such an approach useless for telephone banking services.

It is possible to go even further with the technology and picture a system where all representatives of agencies that supported a potentially vulnerable person had their voices stored in a centralized database. That person could phone the database service and ask the agency representative to speak on the telephone. The database service could then authenticate whether the person who spoke was a genuine representative.

A simple example would be an older woman living on her own has the front doorbell ring. She opens the door and finds a large man is claiming to be from the gas company. He says that there has been a report of the smell of gas from a neighbor and he needs to come into the house to check to make sure that there is no leak in the house. Before allowing him into the house, she could phone the gas company's voice checking service and ask him to

speak on the telephone. If the database services tell her that the voice is recognized and gives the same name as the ID tag on his suit, she can be reassured that he is genuine. Should the voice check fail, though, she would know not to let him into her house and to call the police instead.

7.3.2 Forensic Speaker Recognition

Unfortunately, criminals will find a way past even the best security. Even then, though, speaker recognition technology can help through the identification of an unknown speaker(s) [119].

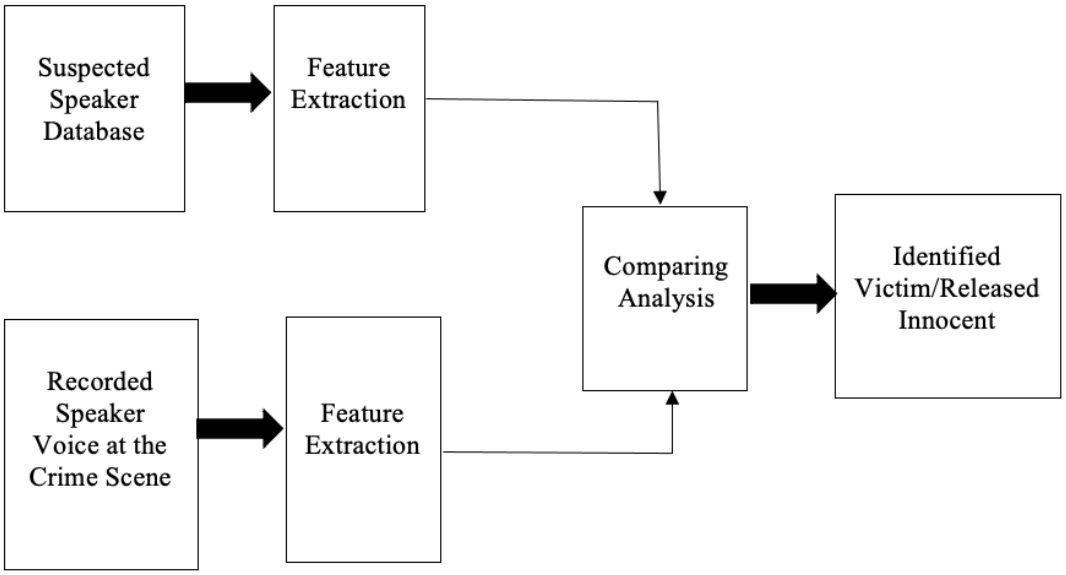

Fig. 7.3 Identifying a Speaker from a Crime Record

However, there is still no source providing a 100 percent confirmation of speaker recognition at the moment due to the differences in speech sample environments. However many criminals end up thinking that committing crimes using voice-related tricks such as ransom calls, harassment calls, or blackmail threats will mask their crimes. With the help of speaker identification, this is no longer true.

Speaker recognition for surveillance could stop such crimes from taking place by monitoring speech via mobiles, telephones, computers, etc., which are constantly listening for certain words related to terrorism or criminal activity [120].

Voice samples of suspects could be compared with the unknown (criminal) speaker. Speaker identification requires comparison of one unknown speaker to a database of stored

samples/suspects, from which, a set of suspects are separated whose voice is within the range of the criminal voice as shown in Figure 7.3. Speaker identification comes down to analyzing the acoustic parameters of voice closely related to voice characteristics.

7.3.3 Identifying a Speaker from Multiple Speakers

Moving to an example where someone is situationally impaired, consider the case where multiple people are participating in a telephone conference call and the participants are not very familiar with each other. It may be useful to help each participant to identify or recognize who is talking at any particular point in time on the conference call.

Participants could register their voices as part of the process of logging on to the conference call. They could be asked to repeat several phrases to allow the system to build a model of their speech. Once the speech models have been logged, the system can analyze who is talking at any particular point and display that information to the other listeners. As with the earlier examples for assisting people with memory loss, additional information about the person who is speaking can be displayed at the same time to assist the other listeners.

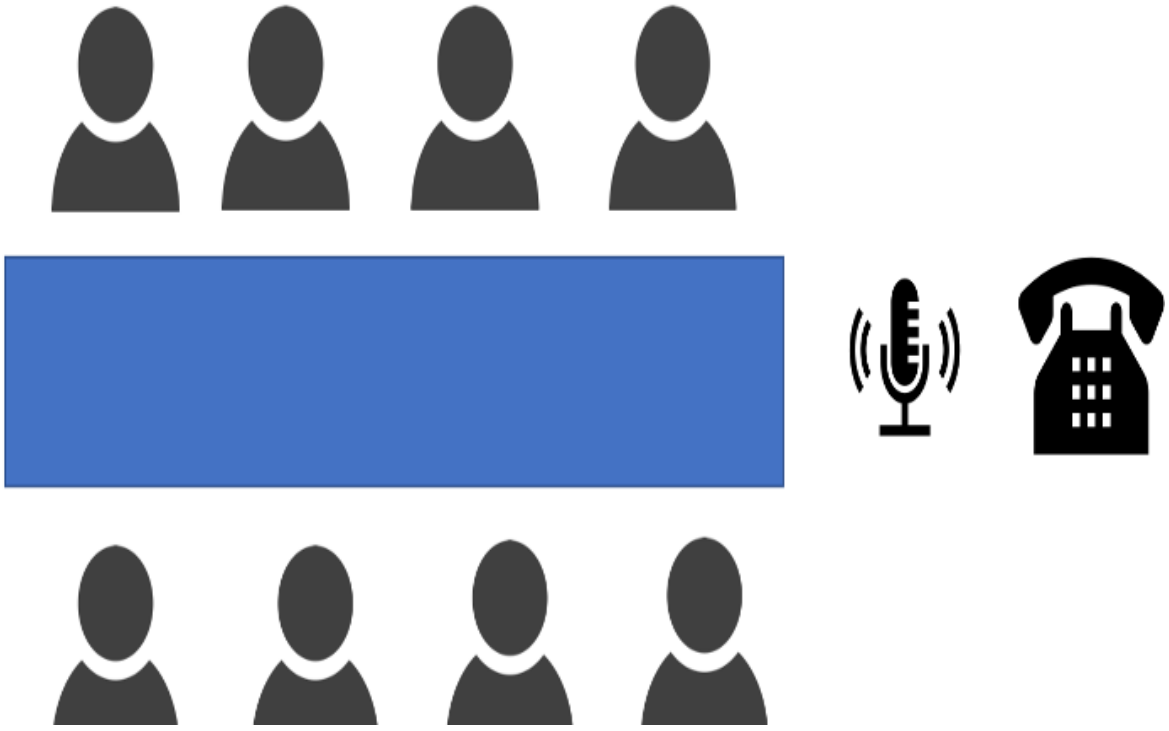

Fig. 7.4 Simultaneous Multiple Speaker Detection and Recognition

The same approach could be used to assist the transcription of meetings, as shown in Figure 7.4. Speaker recognition is being widely and progressively being used to transcribe

data. It reduces time consumption and manual work and produces higher results with greater accuracy. Such programs can be created with in-depth knowledge, and terminology in a specific field, producing error-free documents.

Speaker recognition transcription can be used in numerous places such as in interviews, where the interviewer is constantly taking down the answers provided by the interviewee. Here, a speech recognition device can be used, which can differentiate between the interviewer and interviewee and take down all information instead of manually writing, and breaking the flow of the interview.

Another example is legal transcription, where attending depositions, or hearings can lead to the accumulation of important data left to be transcribed. Instead, they use of a speaker recognition program in such cases, designed to differentiate between the jury, the judge, the witnesses, and the lawyers and transcribe multiple-voice recordings, can save huge amounts of time.

Medical transcription is one of the most important areas where the use of a speaker identification device can reduce the time taken in the treatment or diagnosis of a patient. For example, in the emergency ward, where trauma is incoming 24/7, and patient treatment is the number one priority, the use of a speech recognition device to transcribe emergency room reports can save time for authorized professionals; that is, time is taken to write down the reports. Instead, they can use that time to treat, diagnose, operate, or discharge a patient as soon as possible.

7.3.4 Military Activities and Air Force

Speech recognition in these areas needs to be specialized and requires high performance to eliminate poor signals in remote areas, communicate through long distances and limited bandwidth channels, cancel background noise, and assist in maximizing hands-and-eyes operations so that the entire focus can be cast on the task on hand.

The key in these operations is to activate voice-operated tasks, only by officials. In terms of fighter cockpit applications, speaker recognition programs are already installed into fighter air-crafts with the voices of certain commanders, authorized to fly the specialised air-crafts. Before a flight, this program identifies the commander and activates the aircraft, ready to fly and accept future commands during the flight. During the flight, speaker recognition will allow the pilot to give speech commands such as the release of weapons parameters, setting radio frequencies, and commanding the auto-pilot system.

7.3.5 Personal Digital Assistant

The most common, modern, and up-to-date applications of speech recognition are digital assistants such as Siri (Apple), and Alexa (Amazon). A combination of speaker identification, and verification, can allow secure and personal use of these digital assistants, not only at home but at work, in the car, and even during outdoor activities, [120].

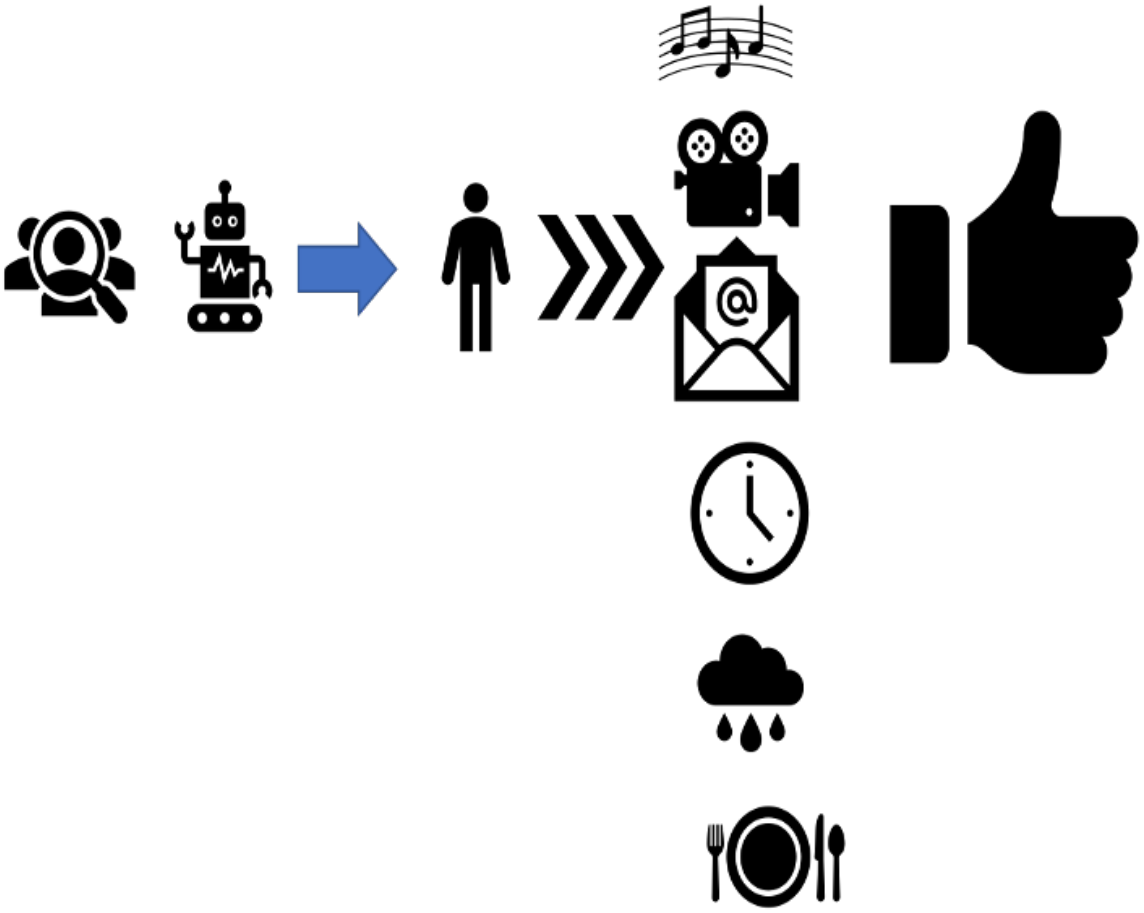

Fig. 7.5 User Interfaces for a Personal Digital Assistant

Considering the example of a smartphone, if the user's hands are occupied but he/she needs to access the phone, speaking to the phone wakes up the voice recognition program and moves on to identifying the user. Once, verified, the system allows the user to make further commands such as set reminders or alarms, call or message contacts, browse the Internet, and so on, using voice identification. This is also beneficial to save time and multitask, thereby reducing screen time and increasing productivity.

The in-car system provides access to information without physically distracting the driver. For instance, if the driver forgets the way to their destination, they can activate the speech recognition system with the push of a button and ask for directions via speech. The sole purpose of introducing speech identification, in this case, would be that if the system realizes that an unknown person is asking for information like directions to home, the system can activate an alarm indicating an intruder while simultaneously sending the location of the vehicle to the police and/or car-owner(s).

7.3.6 Helping Patients in Hospital

The process of recognizing the speaker involves the collection and analysis of many facets of the speech. This includes general features, such as frequency, phase, amplitude, and rate of speech, as well as analysis of more specific contents of the speech, like individual phonemes.

It is possible to combine the speaker recognition system with other technologies, such as a voice stress analyzer to give a snapshot of someone's emotional state. Insurance companies sometimes use such technology to determine whether a caller to their call center is likely to be telling the truth or not, with regards to their claim.

Combining such technologies in a healthcare setting offers new opportunities for monitoring the wellbeing of patients. For example, in a hospital environment, it is not easy to monitor older adults 24 hours a day, 7 days a week. Technological aids, such as using a camera can assist with monitoring activity but are typically not of much use at keeping track of daily activities and health-related issues. An observer is required to keep an eye on the person under observation at all times. Setting up cameras in rooms is also costly and causes obvious privacy issues.

An alternative would be collecting voice samples could be collected from the microphone and using them to identify the speaker while running a basic analysis of the voice for signs of distress, such as a change in frequencies, rate of speech, or other indicators.

Most of the time, saving every tiny bit of information is not required to decide on a patient's health. One can collect voice samples from a patient through a microphone and store them on a database. By listening to a person's voice, one can predict their needs such as: hungry, sick in terms of cold or tired, emotional state such as whether they are in pain, sad, angry, etc.

7.4 Summary

The human brain consists of several systems and subsystems that have proven to be particularly challenging to discover and understand. Everything in the world around us is based on patterns. For example, language is a combination of patterns at various levels, such as letters combine to form syllables; syllables combine to form words; words combine to form clauses and sentences, and then these are all stored in sequential order. So, the question that emerges is, how is this information presented as patterns and how those patterns are stored possibly including sequential/timing information. For example, if someone wanted to explain their house to a friend then normally the explanation starts from the entrance moving on to e.g. the kitchen or living area, and then to the playroom. Even though someone's house is well known to them, still there is a priority to the order in which they tell their friend. This happens all information patterns are associated with one another through a massive network.

As humans use a combination of parameters to be able to identify a speaker. Parameters could be varied based on applications and people as well. For example, humans can identify local and non-local people based on their accent and pronunciation, etc.

This chapter has explored several examples where the use of such technology can be of benefit in achieving the goals of universal access. The examples have included applications where the user either has an innate impairment or a situationally induced one. The technology offers several clear benefits for a wide range of users in many contexts.

Chapter 8

Conclusion and Future Work

Identification of a speaker is the task of identifying or recognising a speaker from a group of people or a database. Voice can be used as a biometric measure to recognise a person. The Human voice is unique to an individual and there will be a difference within the speaker too. However, those differences can be audible, measurable, and differentiate among others. The core aim of the research is to understand, identify and analyse the voice characteristics to increase the chances of identifying a speaker, by narrowing down the population of potential candidates. The voice characteristics results helped to deduce the following conclusions.

Participants successfully identified 90% of the familiar voices to the correct movie artist in less than 60 seconds. Participants tended to do better at identifying voices belonging to voices with who they did express familiarity at the start of the test. Participants who generally have a greater familiarity with famous movie artists tended to perform better. The analysis showed that 10% of participants had taken more than 60 seconds and less than 120 seconds, to identify a speaker. The results of this kind are indicative of a sub-population of people who have an unusually pronounced difficulty with speaker identification. Participants need more training data as they prefer to listen to the audio files multiple times before they provide an answer.

The second part of the experiment demonstrates that people do not need to have prior knowledge of an individual's language to identify, but it is slower to do so. The study shows that people could identify the voices of speakers with 95 % accuracy after hearing a movie artist whose language is familiar with, in less than 60 seconds, but whereas with unfamiliar languages, in 71 seconds.

It was observed from the final experiment that the further distance a participant goes away from the receiver, the harder it becomes to identify the speaker. Since the receiver cannot hear a participant's voice clear, enough to be able to identify them as shown in Figure 3.4. For example, if a person is talking next to you or within a limited distance, you can hear

them clearly and that helps to identify them. However, imagine a person is talking far away from you, for example, 200 meters away, then the listener cannot hear them properly to be able to identify them. In this experiment, energy played an important role and it was hard for participants to maintain the same volume they produce, independent of distance.

A representation of the voice in a spectral analysis helps to identify a voice pattern by measuring fundamental frequency and amplitude. Voice features would be an indirect representation of the human's vocal system including nose, mouth, and throat cavities. Based on the observations, results have shown that fundamental frequency can be used to assist in identifying a person. The analysis of voice characteristics has been observed and results were published.

In the candidate list, initially, the frequency range can be used to delete people who cannot be the speaker, then the next parameter we could use would be speech rate that is participants can be eliminated based on the number of words spoken per minute and both minimum and maximum words per minute can be compared. Those who do not meet these parameters will be eliminated and the list of people remaining to be compared will be narrower. Next, the accent, pronunciation, and repetitively used words (such as some people tend to use some words very often) will be compared.

It was concluded that participants and members have higher SR values and faster AR values in their native language than the learned language. Volume and frequency are dependent on the language. But, differ in pauses to speech ratio and the number of words per pause, possibly reflecting the words/syllables difference across English and his/her native language. Native language (participant's first language/mother-tongue) had a significantly higher speaker rate since participants are very familiar with their first language when compared to the English language. Speech rate depends on the speaker's topic of the conversation and the situation, whereas AR depends on the phones and pauses between the words. The results showed, that there is no significant relationship between SR and AR values. The results proved that language is not a barrier to identify a person. The experiments helped to understand the language barrier in speech production and perception.

The fundamental frequency does not change when the speaker talks in different languages. In 5.3.1, frequency values do not change significantly, because participants have the script before the recording starts. Participants were well prepared to read the script both in English and their native languages. On the other hand in section 5.3.2, there was a minor change in the frequency, where the audience talked on the topic on the spot, that is, without any preparation.

According to the results of the experiments, the fastest speaker in speech rate (SR) has the fastest articulation rate (AR) as well, but the slowest speaker in speech rate SR does not

have the slowest AR. This demonstrates that there is no direct relationship between SR and AR. Speech rate does have a direct influence on speakers, where AR depends on phonemes and movements of the mouth. AR shows the differences within the speaker rather than the speaker's comparison with others. However, not all speakers can vary from others at the same time, so AR would still be considered as one of the parameters which can be used to help recognise a speaker.

The number of pauses and duration of those pauses did not show significant results. The values change based on the languages and duration of the words. The results showed that the number of pauses has some impact on the change of speech rate. People tend to speak faster in their native/first language when compared to a learned language. The results concluded that fundamental frequency, number of pauses, and volume of speech have an impact on differentiating, either the fastest or the slowest speaker.

The volume of a speaker is not consistent enough to be used for identification. However, the volume factor would be helpful to predict the speech rate of the speaker. The volume of the speech is a speaker's controlled variable. For example, the speaker would choose to be loud or soft according to the situation or their mood.

Phonemes are good enough to identify their origin, but inconsistent to identify a person. Even in linguistics, the aim of the listener is not to concentrate on individual phoneme, but to understand the meaning of the words/sentences. It is difficult to extract phonemes from a voice signal manually.

Nowadays, many people tend to go abroad to pursue their higher studies or for their dream job. One tends to learn or adopt a foreign language in terms of accent and pronunciation. However, some individuals pronounce certain words in a unique style, which helps identify their origin. For instance, the emphasis on a certain letter of a word is different in different accents like 'water' in some British English accents, has the 't' silent when pronounced, whereas, in an Indian accent the "ter" in 'water' is pronounced as turr, with an emphasis on the "r". Production of sounds in the vocal tract during speech describes and characterizes the sounds. There are two types of sounds: voiced and unvoiced/voiceless. A voiced sound will produce vibrations in the vocal cord as compared to unvoiced sounds. Unvoiced sounds produce no vibrations in the vocal cord but still generate sounds through the mouth and lips.

Several questions arise about the recording of the above data. These questions include: why are phonemes are changing even though the same person is speaking? Why are some phonemes are distinctive to a participant, while some of them are very similar between different participants? There are several limitations of using phonemes as a fundamental factor that affects voice recognition: Phonemes produce different sounds because of the exaltation of air from our mouths. It is hard to keep track of how these different sounds are

produced, as it is dependent on many factors such as how much air is exhaled whilst speaking, the opening size of the vocal cord is open, the shape of lips, placement of the tongue, etc.

Participants can adjust the boundaries of a phonemes frequency based on the context. For example, the participant will learn how to say words in different ways. There are numerous papers focused on how phonemes are used to identify a person, but it is only available in a few languages. This is mainly because they only know phonemes that are used very often in their language. Phonemes mainly arise from a language perceptive. Humans do not listen to phonemes on their own, however, humans does listen to complete phonemes to understand the language, but not to identify a speaker. Language carries information from human speech, by using words.

Changes in the position of a phoneme create a lot of difference that would reflect a different pattern of human speech, making it more difficult to identify a speaker. Moreover, English is not a phonetic language. In the English language, one phoneme can be represented by using different letters. For example the phoneme \k\ at the start of, Cat, kite, KitKat is represented by the letter 'c'.

Overall, it is clear from the results that speaker identification does not depend on only one parameter or one method. We need different parameters and several methods should be used to eliminate the person whom we are not looking for. SR is the most efficient factor to differentiate between two speakers and AR would be helpful to identify the variations within the speaker. Speakers have fewer pauses in their native language when compared to a learned language. Overall, the number of pauses, percentage of pauses, and duration of time would change and help to analyse the variations within the speaker.

Humans can access several types of information to be able to identify a speaker. For example, when someone calls you, humans do not tend to ask them " Who are you?", humans initially process and question ourself", "is this X, cause her/his accent seems Y country" or "Is this A, cause her/his using a "word" more often" etc. So, humans will ask several types of questions before her/him recognise the person. Whereas, machines do not have access to all the information and their access to information is limited. Machines are not as fundamentally intelligent as humans to recognise. It would be useful, if machines could access more information like humans so that, in the future, machines would be good at identifying a speaker. In this way, machines also can have filters to eliminate people whom they are not looking for.

The principal contributions of this research has been an exploration of how a system can be designed to provide language-independent speaker identification based on characteristics of human voices. The aim was to investigate how a working system could be designed that required a minimum of training and computer processing. The aim was not to develop a

system that would recognise a single individual from a potential pool of millions, but rather how clearly incorrect speakers could be removed from that pool. As such, this would be an additional layer of security over other forms of security and identity checking.

The research thus focused on features that are easy to measure and identify, such as dominant frequencies, pauses and other such attributes. These do not identify an individual uniquely, but do allow the system to identify people who are definitely not correct. By repeating the use of different features to compare to the speaker, the pool of potential candidates can be trimmed down in size quite notably and all very quickly, with minimal data and comparative little computer processing.

By focusing on fundamental features of how someone speaks, the resultant approach is also language and text independent. The data collected showed that a number of the features investigated were independent of what was being said and what language was being used.

The data results collected showed that for the 100 participants involved in the experiments it was possible to eliminate 70 participants based on the dominant frequencies alone. Then, using the speech rate, it was possible to eliminate another 20 people from the population. Phonemes helped to remove another 5, then there were only 5 people left. These simple speaking features thus eliminated 95 % of the candidate population.

Other approaches could generate more specific results, but typically involve substantially more training data, more computational power or restricting participants to having to use defined phrases in a specified language. The approach taken in this research in this thesis has none of these limitations or requirements.

There is further fundamental research to be undertaken to build on this approach such as: how humans identify a person, how much data do humans need to identify a person, how much time can humans take to identify familiar and unfamiliar voices and languages, which aspects of speech do humans use to make the identification process faster, etc., which still needs to be carried out. Identifying voice characteristics independent of the language used and environmental factors would be interesting research to follow. It would be fascinating to learn and understand the individual speaker-specific features to identify a speaker.

8.1 Future Work

Although there have been advances in voice recognition systems to identify a speaker, still there is a downside of the technology that stress, health issues can impact the results. A speaker's voice is subject to change based on her/his health and emotional state.

A voice pattern requires the speaker to speak in a normal voice that was recorded for training. If the speaker suffers from any health issues such as cold, then the voice pattern will

not match with the pattern that has been stored during the enrolment. Human speech can be processed to recognise emotion. To detect emotion from human speech, peak to peak distance was calculated from the graphical representation of FFT. However, it was concluded that to get better accuracy, a voice should be recorded from one person with different emotions and compare with the same person rather than a group of people. Detecting emotion from the human speech is how a person speaks rather than what the speaker says.

Currently, very few machine learning algorithms care about the information coming from the environment. It would therefore be interesting to identify suitable ways to use multilevel alignment for information to create patterns similar to those in the brain. These machines should then also be able to retrieve data automatically and even predict the output results based on incomplete patterns being presented. While overall, the data would be stored efficiently due to the inherent compression and the system would easily be able to learn and continuously improve its patterns. The research will propose such a multi-level model, build a prototype simulation and evaluate its performance against baseline comparisons.

Input to the human brain travels through the cerebellum and striatum, before reaching the cortex. Within the cortex, it is generally believed that there would be one unique algorithm that processes the information which is received from sensory organs. The cortex also stores information as patterns in a hierarchical structure. However, it appears that the brain does not know the difference between information received from sensory organs and virtual creations of the brain itself. Consequently, being able to create a multilevel hierarchy to store different patterns will help a machine learn like humans. This should allow the machine to deliver similar functionality to that of the human brain.

One more problem with machines is their memory. Humans have typically two types of memories: short-term and long-term memory. Daily activities or reminders for a particular time are stored in short-term memory, only for as long as needed. In long-term memory, several types of information would be there and a combination of all the information will be going in and out. A hard disk is used as long-term memory where the machine can retrieve different types of information for identification.

References

[1] Beukelman, D.R. and P. Mirenda, 1998. “Augmentative and alternative communication”, Baltimore: Paul H. Brookes.

[2] Hawley M. S., Cunningham, S. P., Green, P. D., Enderby, P., Palmer, R., Sehgal, S. and P. O’Neill, “A Voice-Input Voice-Output Communication Aid for People With Severe Speech Impairment”, in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 21, no. 1, pp. 23-31, Jan. 2013, doi: 10.1109/TNSRE.2012.2209678.

[3] Matsui, T. and Furui, S., “Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMM’s”, in IEEE Transactions on Speech and Audio Processing, vol. 2, no. 3, pp. 456-459, July 1994, doi: 10.1109/89.294363.

[4] Shirali-Shahreza, S., Abolhassani, H. and Shirali-shahreza, M. H., “Fast and Scalable System for Automatic Artist Identification”, in IEEE Transactions on Consumer Electronics, vol. 55, no. 3, pp. 1731-1737, August 2009, doi: 10.1109/TCE.2009.5278049.

[5] Ramli, I. and Ortega-Sanchez, C., “Pattern recognition using hierarchical concatenation”, 2015 International Conference on Computer, Control, Informatics and its Applications (IC3INA), Bandung, 2015, pp. 109-113, doi: 10.1109/IC3INA.2015.7377756.

[6] Prachi, K. and Bhope, V.P., “A Review on Speech to Text Conversion Methods”, in International Journal of Advanced Research in Computer Engineering and Technology (IJARCET), Vol. 4, Issue 7, July 2015.

[7] Ignatenko, T. and Willems, F. M. J., “Biometric Systems: Privacy and Secrecy Aspects”, IEEE Transactions on Information Forensics and Security, vol. 4, no. 4, pp. 956–973, 2009.

[8] Gerasimos, P., Chalapathy, N., Juergen L. and Iain,M., “Audio-Visual Automatic Speech Recognition: An Overview” International Journal of Issues in audio-visual speech processing, MIT Press, 2004.

[9] Al-haddad, S. A. R., Samad, S. A., Hussain, A., Ishak, K. A., Noor, A. O. A., "Robust Speech Recognition Using Fusion Techniques and Adaptive Filtering", in American Journal of Applied Sciences, Vol. 6(2), pp. 290-295, 2009, doi: 10.3844/ajassp.2009.290.295

[10] http://www.biometrics.gov/Documents/SpeakerRec.pdf.

[11] Divya, R. and Vijayalakshmi, V., "Analysis of Multimodal Biometric Fusion Based Authentication Techniques for Network Security", International Journal of Security and Its Applications. Vol. 9, No. 4 (2015), pp. 239-246, doi: org/10.14257/ijsia.2015.9.4.22.

[12] Panda, S. P., "Automated speech recognition system in advancement of human-computer interaction", 2017 International Conference on Computing Methodologies and Communication (ICCMC), Erode, pp. 302-306, 2017, doi: 10.1109/ICCMC.2017.8282696.

[13] Rosdi, F. and R. N. Ainon, "Isolated malay speech recognition using Hidden Markov Models", in International Conference on Computer and Communication Engineering, Kuala Lumpur, 2008, pp. 721-725, doi: 10.1109/ICCCE.2008.4580699.

[14] Abberton, E. and Fourcin, A.J., "Intonation and speaker identification. Language and Speech", International Journal Vol: 21 issue: 4, pp: 305-318., 1978, doi.org/10.1177/002383097802100405.

[15] George, D. and Hawkins, J., "A hierarchical Bayesian model of invariant pattern recognition in the visual cortex", in IEEE International Joint Conference on Neural Networks, Montreal, Que., Vol. 3, pp. 1812-1817, 2005, doi: 10.1109/IJCNN.2005.1556155.

[16] Dudley, H. "The Vocoder", Bell Labs Record, Vol.17, pp. 122-126, 1939, doi: 10.1038/145157a0.

[17] Dudley, H., Riesz, R. and S. A. Watkins, "A Synthetic Speaker", Journal of the Franklin Institute, vol. 227, pp. 739-764, 1939, doi.org/10.1016/S0016-0032(39)90816-1.

[18] Weintraub, M., Murveit, H., Cohen, M., Price, P., Bernstein, J., Baldwin, G. and D. Bell, "Linguistic constraints in Hidden Markov Model-based speech recognition", International Conference on Acoustics, Speech, and Signal Processing, Glasgow, UK, pp. 699-702 vol.2, 1989, doi: 10.1109/ICASSP.1989.266523.

[19] Manfred R. Schroeder, "A brief history of synthetic speech", in Journal of Speech Communication, Volume 13, Issues 1–2,1993, Pages 231-237, doi.org/10.1016/0167-6393(93)90074-U.

[20] Stevens S. S., Volkman J., Newman E., “A scale for the measurement of the psychological magnitude pitch”, in Journal of Acoustical Society of America, vol. 8, pp.185–190, 1937.

[21] Cardin, R., Normandin, Y. and E. Millien, “Inter-word coarticulation modelling and MMIE training for improved connected digit recognition”, in IEEE International Conference on Acoustics, Speech, and Signal Processing, Minneapolis, MN, USA, pp. 243-246 vol.2, 1993, doi: 10.1109/ICASSP.1993.319280.

[22] Normandin, Y., Cardin, R. and R. De Mori, “High-performance connected digit recognition using maximum mutual information estimation”, in IEEE Transactions on Speech and Audio Processing, vol. 2, no. 2, pp. 299-311, April 1994, doi: 10.1109/89.279279.

[23] Cardin, R., Normandin, Y. and De Mori, R. “High performance connected digit recognition using codebook exponents”, [Proceedings] ICASSP-92: in IEEE International Conference on Acoustics, Speech, and Signal Processing, San Francisco, CA, USA, pp. 505-508 vol.1, 1992, doi: 10.1109/ICASSP.1992.225861.

[24] Kratzenstein, C. G., “Sur la raissance de la formation des voyelles. (eng.: On the origin and the formation of vowels)” In: J. de. Physique., Vol 21, pp. 358-380, 1782

[25] Davis, K. H., Biddulph, R. and S. Balashek, “Automatic Recognition of Spoken Digits”, The Journal of the Acoustical Society of America, Vol 24, No. 6, pp. 627-642, 1952.

[26] Itakura, F., “Minimum Prediction Residual Principle Applied to Speech Recognition”, IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 23, no. 1, pp. 67-72, February 1975. Reprinted in Waibel and Lee (1990).

[27] Richard Peacocke, D. and Daryl Graf, H, “An Introduction to Speech and Speaker Recognition in News Briefs” in Computer, vol. 23, pp. 26-28, 2000, doi: 10.1109/MC.2000.10068.

[28] Hillenbrand J., Getty L. A., Clark M. J., Wheeler K., “Acoustic characteristics of American English vowels “, in Journal of Acoustical Society of America, vol. 97, pp.3099–3111, 1995.

[29] Joe, T. “Speech Recognition using Neural Networks”, CMU-CS-95-142, School of Computer Science, May 1995, Carnegie Mellon University Pittsburgh, Pennsylvania 15213-3890.

[30] Olson, H. F. and Belar, H., “Phonetic Typewriter”, The Journal of the Acoustical Society of America, Vol. 28, No. 6, pp. 1072-1081, 1956.

[31] Rabiner, L. R., Wilpon, J. G. and Soong, F. K. “High performance connected digit recognition using hidden Markov models”, in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 8, pp. 1214-1225, Aug. 1989, doi: 10.1109/29.31269.

[32] Forgie, J. W., and Forgie, C. D., “Results Obtained from a Vowel Recognition Computer Program”, The Journal of the Acoustical Society of America., Vol. 31, No. 11, pp. 1480-1489, 1959.

[33] Rosenhouse, G., “Biomimetics of sound production, synthesis and recognition”. WIT Transactions on Ecology and the Environment, vol. 138., pp. 273-287, 2010, doi: 10.2495/DN100241.

[34] Furui, S., “Cepstral analysis technique for automatic speaker verification “. IEEE Transactions on Acoustics, Speech and Signal Processing, vol.29, pp. 254–272, 1981.

[35] Rabiner, L. and Juang, B.H., “An Introduction to Hidden Markov Models”, IEEE ASSP Magazine, Vol. 3, No.1, Part 1, pp. 4-16., 1986.

[36] Rabiner, L., “A Tutorial on Hidden Markov Models and selected Application in Speech Recognition”, in proceedings of IEEE, Vol. 77, No. 2, pp. 257-286., 1989.

[37] Picone, J., “Continues Speech Recognition using Hidden Markov Models”, IEEE ASSP Magazine, Vol.7, Issue 3, pp. 26-41., 1990.

[38] Flahert, M. J. and Sidney, T., “Real Time Implementation of HMM speech recognition for telecommunication applications”, in proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP), Vol. 6, pp. 145-148., 1994.

[39] Manning, C. and Schütze, H., “Foundations of Statistical Natural Language Processing”, Cambridge: MIT Press, 1999.

[40] Jin, W., Liu, X., Scordilis, M. S. and Han, L., “Speech Enhancement Using Harmonic Emphasis and Adaptive Comb Filtering”, in IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 2, pp. 356-368, Feb. 2010, doi: 10.1109/TASL.2009.2028916.

[41] Davis, S., and Mermelstein, P., “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences “, in IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 28, pp. 357–366, 1980.

[42] Titze, I. and M. Daniel, "Principles of Voice Production", Book, published by Prentice Hall, Englewood Cliffs (1994).

[43] Yegnanarayana, B., Prasanna, S. R. M., Zachariah, J. M. and C. S. Gupta, "Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system", in IEEE Transactions on Speech and Audio Processing, vol. 13, no. 4, pp. 575-582, July 2005, doi: 10.1109/TSA.2005.848892.

[44] Atal, B., "Automatic Speaker Recognition Based on Pitch Contours", Journal of the Acoustical Society of America, Vol. 52, pp. 1687–1697 (1972), doi: org/10.1121/1.1913303.

[45] S. Furui, "An Overview of Speaker Recognition Technology", In: Lee CH., Soong F.K., Paliwal K.K. (eds) Automatic Speech and Speaker Recognition. The Kluwer International Series in Engineering and Computer Science (VLSI, Computer Architecture and Digital Signal Processing), vol 355. 1996, Springer, Boston, MA.

[46] Furui, S., "Cepstral analysis technique for automatic speaker verification", IEEE Transactions on Acoustics, Speech and Signal Processing, vol.29, pp. 254–272, 1981.

[47] Kishore, P., Sudhakar, V., Ranganatham, V., Bharat, G., Krishna, M. and Debashish Roy, S., "Significance of Formants from Difference Spectrum for Speaker Identification", in conference proceedings INTERSPEECH-2006, ICSLP, Ninth International Conference on Spoken Language Processing, 2006.

[48] Chakroborty, S. and Goutam, S., "Improved Text-Independent Speaker Identification using Fused MFCC and IMFCC Feature Sets based on Gaussian Filter" in International Journal of Signal Processing, Vol. 35, 2009.

[49] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P. and Ouellet, P., "Front-end factor analysis for speaker verification", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 4, pp.788-798, 2010.

[50] Atal, B., "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", in Journal of the Acoustic Society of America, vol. 55, pp.1304–1312, 1974.

[51] Charlet, D., and Jouvet, D., "Optimizing feature set for speaker verification", in Pattern Recognition Letters, vol. 18, pp. 873–879, 1997.

[52] Furui, S., “Recent advances in speaker recognition”, Pattern Recognition Letters Vol.18, pp.859–872, 1997.

[53] Revathi1, A., Ganapathy, R. and Venkataramani, Y., “Text Independent Speaker Recognition and Speaker Independent Speech Recognition Using Iterative Clustering Approach”, in International Journal of Computer science and Information Technology (IJCSIT), Vol 1, No 2, November 2009.

[54] Chaudhari, U., Navratil, J., and Maes, S., ´ “Multigrained modeling with pattern specific maximum likelihood transformations for text-independent speaker recognition”, in IEEE Trans. on Speech and Audio Processing, vol.1, 61–69, 2003.

[55] Todor, G., Nikos, F., George, K., “Comparative evaluation of various MFCC implementations on the speaker verification task”, in Proceedings of the SPECOM, Vol. 1, 2005.

[56] Chen, K., Wang, L., and Chi, H., “Methods of combining multiple classifiers with different features and their applications to text-independent speaker recognition “, in International Journal of Pattern Recognition and Artificial Intelligence, Vol. 11, pp. 417–445, 1997.

[57] Shirali-Shahreza, S., Abolhassani, H. and Shirali-shahreza, M. H., “Fast and Scalable System for Automatic Artist Identification”, in IEEE Transactions on Consumer Electronics, vol. 55, no. 3, pp. 1731-1737, August 2009, doi: 10.1109/TCE.2009.5278049.

[58] Pruzansky, S., “Pattern-matching procedure for automatic talker recognition”, The Journal of the Acoustical Society of America, vol. 35, pp. 354-358, 1963.

[59] Duda, R., Hart, P., and Stork, D., “Pattern Classification”, second ed. Wiley Interscience, New York, 2000.

[60] Bharti, R. and Bansal, P., “Real time speaker recognition system using MFCC and vector quantization technique”, International Journal of Computer Applications, Vol. 117, pp.25-31, 2015, doi: 10.5120/20520-2361.

[61] Gersho, A., and Gray, R., “Vector Quantization and Signal Compression”, Kluwer Academic Publishers, Boston, 1991.

[62] Pahini Trivedi, A., “Introduction to Various Algorithms of Speech Recognition: Hidden Markov Model, Dynamic Time Warping and Artificial Neural Network”, International

Journal of Engineering Development and Research (IJEDR), ISSN:2321-9939, Vol.2, Issue 4, pp.3590-3596, December 2014, doi=10.1.1.677.7302.

[63] Sharma, S., "Speech Recognition with Hidden Markov Model: A Review", International Journal of Scientific and Engineering Research, Volume 6, Issue 11, November-2015.

[64] Charlet, D., Jouvet, D., and Collin, O., "An alternative normalization scheme in HMM-based text-dependent speaker verification", in Speech Communication, vol. 32, pp. 113–120, 2000.

[65] Fukunaga, K., "Introduction to Statistical Pattern Recognition", second ed. Academic Press, London, 1990.

[66] Mark, G., and Y. Steve, "The Application of Hidden Markov Models in Speech Recognition", Foundations and Trends in Signal Processing, Vol. 1, No. 3 (2007), pp. 195–304, doi: 10.1561/2000000004.

[67] Alexander, A., Botti, F., Dessimoz, D., and Drygajlo, A., "The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications" in Forensic Science International 146S, pp. 95-99, 2004.

[68] Wolff, J. G., "The SP theory of intelligence and the representation and processing of knowledge in the brain", 2016.

[69] Hawkins, J. and Blakeslee, S., "On intelligence: How a new understanding of the brain will lead to the creation of truly intelligent machines". Macmillan, 2007.

[70] Xue, M. and Zhu, C., "A study and application on machine learning of artificial intelligence". In 2009 International Joint Conference on Artificial Intelligence, pp. 272-274, 2009, IEEE.

[71] Kosuge, K. and Hirata, Y., "Human-robot interaction". In 2004 IEEE International Conference on Robotics and Biomimetics, pp. 8-11, 2004, IEEE.

[72] Van Lancker, D., Kreiman,J. and Emmorey, K. , "Familiar voice recognition: patterns and parameters Part I: Recognition of backward voices", Journal of Phonetics,Volume 13, Issue 1, 1985,Pages 19-38, doi.org/10.1016/S0095-4470(19)30723-5.

[73] Short, E., Feil-Seifer, D. and Matarić, M., "A comparison of machine learning techniques for modeling human-robot interaction with children with autism". In 2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 251-252, 2011, IEEE.

[74] Kazuaki, T., Motoyuki, O. and Natsuki, O., "The hesitation of a robot: A delay in its motion increases learning efficiency and impresses humans as teachable". In 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 189-190, 2010, IEEE.

[75] Gao, J., Li, C.F., Liu, Z.G. and Liu, L.Z., "Elicitation of machine learning to human learning from iterative error correcting". In 2013 International Conference on Machine Learning and Cybernetics, Vol. 1, pp. 229-234, 2013, IEEE.

[76] Amino, K. and Arai, T., "Effects of linguistic contents on perceptual speaker identification: Comparison of familiar and unknown speaker identifications", In journal of Acoustical Science and Technology, Vol. 30, pp. 89-99, 2009.

[77] http://www.mplsvpn.info/2017/11/what-is-neuron-and-artificial-neuron-in.html.

[78] Kinnunen, T.H., "Optimizing spectral feature-based text-independent speaker recognition". 2005, University of Joensuu.

[79] George, D. and Hawkins, J., "A hierarchical Bayesian model of invariant pattern recognition in the visual cortex. Neural Networks", in Proceedings. 2005 IEEE International Joint Conference on, 2005.

[80] Zaw Win, A., "A Robust Speaker Identification System" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-2, Issue-5, August 2018, pp.2057-2064.

[81] Saini, P. and Rao, P., "Multimodal Biometrics Security: A Review", International Journal of Innovative Research in Engineering and Multidisciplinary Physical Sciences, Volume 6, Issue 1, January-February 2018. doi: 10.17605/OSF.IO/4R8NA.

[82] Atal, B., "Automatic speaker recognition based on pitch contours", in Journal of the Acoustic Society of America 52, Vol.6, pp. 1687–1697, 1972.

[83] Monson Brian, B., Hunter Eric, J., Lotto Andrew, J. and Story Brad, H., "The perceptual significance of high-frequency energy in the human voice", in Jounral of Frontiers in Psychology. Vol. 4, pp. 587-597, 2004, doi:10.3389/fpsyg.2014.00587.

[84] Shuren, B.O., "Extraction of Instantaneous Frequency Characteristic Using Time-frequency Ridges", in Journal of Mechanical Engineering, Vol,10, 2008.

[85] Acoustical Terminology, American Standard Association, 1960, N.Y.

[86] Plomp, R., “The Intelligent Ear: On the Nature of Sound Perception”, 1st ed. book, New York: Psychology Press, 2001.

[87] Fletcher, H., “Loudness, pitch and the timbre of musical tones and their relation to the intensity, the frequency and the overtone structure”, in Journal of the Acoustical Society of America, vol. 6, no. 2, pp. 59–69, 1934, doi:org/10.1121/1.1915704

[88] Olson, H. F., “Music, Physics and Engineering", Dover books on music, music history, Dover, 1967.

[89] Sadkhan, S.B., Al-Shukur, B.K. and Mattar, A.K., “Human voice extracted biometric features: What can be used for”, In International Conference on Current Research in Computer Science and Information Technology (ICCIT) pp: 7-12, 2017, IEEE.

[90] Reynolds, D.A., “An overview of automatic speaker recognition technology”, In IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 4, pp. IV-4072, May 2002, IEEE.

[91] Farrell, K., Ramachandran, R., and Mammone, R., “An analysis of data fusion methods for speaker verification “. In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 1998) (Seattle, Washington, USA, 1998), vol. 2, pp. 1129–1132, 1998.

[92] Shaver, C.D. and Acken, J.M., “A brief review of speaker recognition technology”, Electrical and Computer Engineering Faculty Publications and Presentation, pp. 1-7, 2016.

[93] Bartle, A. and Dellwo, V., “Auditory speaker discrimination by forensic phoneticians and naive listeners in voiced and whispered speech”, in International Journal of Speech Language and the Law. vol. 22, pp. 229-248, 2015, doi: 10.1558/ijsll. v22i2.23101.

[94] Cao, H. and Wang, Y. “A Forensic Aspect of Articulation Rate Variation in Chinese”. Hong Kong, Key Laboratory of Evidence Science (China University of Political Science and Law), Ministry of Education, China; Department of Chinese Language and Literature, Peking University, Beijing, China; Centre of Criminal Technology, Public Security Bureau of Guangdong Province, 17-21 August 2011.

[95] Ezzaidi, H., Rouat, J., and O’Shaughnessy, D., “Towards combining pitch and MFCC for speaker identification systems “, In Proc. 7th European Conf. on Speech Communication and Technology, pp. 2825–2828, 2001.

[96] Hawkins, J. and Blakeslee, S., “On intelligence”: How a new understanding of the brain will lead to the creation of truly intelligent machines. Book, 2007 Macmillan.

[97] Andrews, W.D., Kohler, M.A., Campbell, J.P., Godfrey, J. and Hernandez-Cordero, J. J., "Gender- Dependent Phonetic Refraction for Speaker Recognition", In: Proc. ICASSP. Orlando, vol. 1, pp. 149–152 (2002).

[98] Saini, P. and Rao, P., "Multimodal Biometrics Security: A Review", International Journal of Innovative Research in Engineering and Multidisciplinary Physical Sciences, Volume 6, Issue 1, January-February 2018. doi: 10.17605/OSF.IO/4R8NA.

[99] Bazyar, M. and Sudirman, R., "A new speaker change detection method in a speaker identification system for two-speakers segmentation". In 2014 IEEE Symposium on Computer Applications and Industrial Electronics (ISCAIE), pp. 141-145, April 2014, IEEE.

[100] Chowdhury, M.F.R., Selouani, S.A. and O'Shaughnessy, D., "Distributed automatic text-independent speaker identification using GMM-UBM speaker models". In 2009 Canadian Conference on Electrical and Computer Engineering, pp. 372-375, 2009, IEEE.

[101] Al-Hattami, A.A., "A Phonetic and Phonological Study of the Consonants of English and Arabic". Language in India, vol. 10, pp.5 -10, 2010.

[102] Eatock, J., and Mason, J. A., "quantitative assessment of the relative speaker discriminating properties of phonemes", In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 1994) (Adelaide, Australia, 1994), pp. 133–136, 1994.

[103] Zhao, C., Wang, H., Hyon, S., Wei, J. and Dang, J., "Efficient feature extraction of speaker identification using phoneme mean Fratio for Chinese". In 2012 8th International Symposium on Chinese Spoken Language Processing, pp. 345-348, 2012, IEEE.

[104] Bacha, S., Ghozi, R., Jaidane, M. and Gouider-Khouja, N., "Arabic adaptation of phonology and memory test using entropy-based analysis of word complexity". In 2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA), pp. 672-677, 2012, IEEE.

[105] Akhila, K.S. and Kumaraswamy, R., "Comparative analysis of Kannada phoneme recognition using different classifiers". In 2015 International Conference on Trends in Automation, Communications and Computing Technology (I-TACT-15), pp. 1-6, 2015, IEEE.

[106] Andrews, W., Kohler, M., Campbell, J., and Godfrey, J., "Phonetic, idiolectal, and acoustic speaker recognition", In Proc. Speaker Odyssey: The Speaker Recognition Workshop (Odyssey), pp. 55–63, 2001.

[107] Ngo, G.H., Nguyen, M. and Chen, N.F., "Phonology-augmented statistical framework for machine transliteration using limited linguistic resources". IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 27, pp.199-211, 2018.

[108] Shih, S.S. and Inkelas, S., "Auto segmental aims in surface-optimizing phonology". Linguistic Inquiry, Vol. 50, pp.137-196, 2018.

[109] Bartle, A. and Dellwo, V., "Auditory speaker discrimination by forensic phoneticians and naive listeners in voiced and whispered speech", in International Journal of Speech Language and the Law. vol. 22, pp. 229-248, 2015, doi: 10.1558/ijsll. v22i2.23101.

[110] Upmanyu, M., Narnboodiri, A. M., Srinathan, K.and Jawahar, C. V., "Blind Authentication: A Secure Crypto-Biometric Verification Protocol", IEEE Transactions on Information Forensics and Security, vol.5, no. 2, pp. 255-265, 2010.

[111] Nagaraja, B.G. and Jayanna, H.S., December. "Efficient window for monolingual and crosslingual speaker identification using MFCC". In 2013 International Conference on Advanced Computing and Communication Systems, pp. 1-4, 2013, IEEE.

[112] Andrews, W.D., Kohler, M.A., Campbell, J.P., Godfrey, J. and Hernandez-Cordero, J. J., "Gender- Dependent Phonetic Refraction for Speaker Recognition", In: Proc. ICASSP. Orlando, vol. 1, pp. 149–152 (2002).

[113] Faisal, B., Saira, B. and Khan, M. F. "Controlling Home Appliances Remotely through Voice Command", in International Journal of Computer Applications, vol. 48, June 2012, doi: 10.5120/7437-0133.

[114] Mark, H., Stuart, C., Phil, G., Pamela, E., Rebecca, P., Siddharth, S. and Peter, O., "A Voice-Input Voice-Output Communication Aid for People with Severe Speech Impairment", IEEE transactions on neural systems and rehabilitation engineering: a publication of the IEEE Engineering in Medicine and Biology Society. vol. 21, no. 1, pp. 23-31, Jan 2013. doi: 10.1109/TNSRE.2012.2209678.

[115] Baig, F., Saira, B. and Khan, M.F., "Zigbee based home appliances controlling through spoken commands using handheld devices", International Journal of Smart Home Vol. 7, No. 1, pp. 19-26, Jan 2013.

[116] Arthi, J.E. and Jagadeeswari, M., "Control of Electrical Appliances through Voice Commands", IOSR Journal of Electrical and Electronics Engineering, vol.9, pp. 13-18, February 2014.

[117] Divya, R. and Vijayalakshmi, V., "Analysis of Multimodal Biometric Fusion Based Authentication Techniques for Network Security", International Journal of Security and Its Applications. Vol. 9, No. 4 (2015), pp. 239-246, doi: org/10.14257/ijsia.2015.9.4.22.

[118] Apsingekar, V. R. and De Leon, P. L. "Speaker Model Clustering for Efficient Speaker Identification in Large Population Applications", in IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 4, pp. 848-853, May 2009, doi: 10.1109/TASL.2008.2010882.

[119] Upmanyu, M., Narnboodiri, A. M., Srinathan, K.and Jawahar, C. V., "Blind Authentication: A Secure Crypto-Biometric Verification Protocol", IEEE Transactions on Information Forensics and Security, vol.5, no. 2, pp. 255-265, 2010.

[120] Ortega-Garcia, J., Gonzalez-Rodriguez, J. and Cruz-Llanas, S., "Speech variability in automatic speaker recognition systems for commercial and forensic purposes", in IEEE Aerospace and Electronic Systems Magazine, vol. 15, no. 11, pp. 27-32, Nov. 2000, doi: 10.1109/62.888324.

Appendix A

Tables of Chapter 3

Table A.1 The Time Taken by Participants to Recognise a Familiar Voice from a Second Audio Clip

Participant	Recognition of a Familiar Voice Audio clip of 60 seconds	
	Measure Time in seconds	
	Female	Male
51	10	10
52	20	10
53	10	20
54	20	20
55	20	10
56	10	10
57	20	10
58	20	30
59	20	10
60	20	10
61	10	10
62	20	10
63	10	20
64	20	20
65	20	10
66	10	10
67	20	10

Table A.1 continued from previous page

68	20	30
69	20	10
70	20	10
71	10	10
72	20	10
73	10	20
74	20	20
75	20	10
76	10	10
77	20	10
78	20	30
79	20	10
80	20	10
81	10	10
82	20	10
83	10	20
84	20	20
85	20	10
86	10	10
87	20	10
88	20	30
89	20	10
90	20	10
91	10	10
92	20	10
93	10	20
94	20	20
95	20	10
96	10	10
97	20	10
98	20	30
99	20	10
100	10	10

A.1 Recognition of an Unfamiliar Voice

Table A.2 The Time Taken by Participants to Recognise Unfamiliar Voice from a Second Audio Clip

Participant	Recognition of Unfamiliar Voice Audio clip of 60 seconds	
	Measure Time in seconds	
	Female	Male
51	30	40
52	40	40
53	20	20
54	20	40
55	40	40
56	20	30
57	20	30
58	20	30
59	20	10
60	40	40
61	20	30
62	40	20
63	20	60
64	50	10
65	40	30
66	20	40
67	40	30
68	50	30
69	20	10
70	30	40
71	40	50
72	30	20
73	50	20
74	50	10
75	50	40
76	20	10
77	50	30

Table A.2 continued from previous page

78	30	40
79	40	40
80	10	50
81	30	10
82	40	20
83	20	20
84	60	50
85	10	60
86	20	60
87	30	30
88	80	40
89	40	10
90	10	40
91	50	30
92	40	60
93	30	30
94	30	30
95	20	20
96	30	30
97	50	20
98	10	40
99	40	40
100	30	30

A.2 Time Taken to Identify a Speaker Whose Language is Familiar

Table A.3 Time taken to identify a speaker whose language is familiar

Participant	Familiar Language Audio clip of 60 seconds	
	Time taken to recognise a speaker in a familiar language	
	Female	Male
51	70	40
52	60	50
53	20	30
54	40	20
55	60	40
56	40	10
57	30	30
58	40	40
59	40	10
60	50	60
61	30	20
62	10	40
63	30	10
64	40	50
65	50	70
66	20	10
67	20	30
68	30	20
69	10	20
70	10	10
71	10	40
72	60	20
73	20	10
74	50	10
75	30	10
76	40	10
77	10	30
78	30	20
79	40	50

Table A.3 continued from previous page

80	50	30
81	60	40
82	30	30
83	20	10
84	30	20
85	50	30
86	80	60
87	30	40
88	20	20
89	30	30
90	40	10
91	50	20
92	10	30
93	10	10
94	10	40
95	30	20
96	40	10
97	40	10
98	70	40
99	10	20
100	10	10

A.3 Time Taken to Identify a Speaker Whose Language is an Unfamiliar

Table A.4 Time taken to identify a speaker whose language is an unfamiliar

Participant	**Familiar Language Audio clip of 60 seconds**	
	Time taken to recognise a speaker in an unfamiliar language	
	Female	**Male**
51	50	60
52	20	50

Table A.4 continued from previous page

53	10	40
54	10	40
55	60	30
56	70	100
57	90	120
58	100	80
59	170	120
60	60	50
61	50	30
62	40	50
63	70	80
64	80	70
65	150	100
66	130	80
67	60	70
68	80	100
69	90	50
70	20	60
71	100	130
72	60	40
73	50	50
74	40	60
75	100	70
76	90	80
77	80	70
78	100	120
79	90	70
80	80	50
81	130	80
82	50	40
83	120	100
84	110	70
85	150	80
86	170	90

Table A.4 continued from previous page

87	170	100
88	110	70
89	30	60
90	30	70
91	50	30
92	60	40
93	20	10
94	20	20
95	20	40
96	30	20
97	100	70
98	80	50
99	90	100
100	100	80

Appendix B

Tables of chapter 4

Table B.1 Analysis of Fundamental Frequency of People's Voices

Participant	**Mean. Freq (Hz)**	**Median. Freq (Hz)**	**Min Freq (Hz)**	**Max Freq (Hz)**
51	129	190	145.4	135
52	167	243	188.9	172.5
53	97	132	111.6	105
54	174	210	186.7	187.5
55	180	255	238.9	246.5
56	115	126	118.2	116.5
57	178	255	201.8	190
58	108	193	136.2	128.5
59	132	213	164.6	147.5
60	104	204	154.3	157.5
61	110	190	154.7	152.5
62	105	198	143.8	136.5
63	111	178	133.7	118.5
64	113	198	137.9	128.5
65	156	255	212.7	227.5
66	102	223	154.9	150
67	120	199	163.8	171
68	185	239	219.2	223
69	149	196	163	158.5
70	176	195	185.4	187

71	168	200	179.9	174
72	245	250	247.9	248
73	138	179	150	142.5
74	198	250	212.8	199.5
75	229	245	235.4	234
76	128	251	199	192.5
77	125	250	188.1	188.5
78	181	255	216.7	220.5
79	110	135	123.3	121.5
80	162	210	181.2	178.5
81	83	99	89.1	86.5
82	180	250	207.8	200.5
83	237	255	245.9	248.5
84	85	160	126.2	131.5
85	85	255	161.3	159.5
86	117	243	168.2	174
87	90	255	160.4	148.5
88	100	255	171.3	155.5
89	85	121	96.4	95
90	115	255	173.1	160
91	100	120	110.3	112
92	100	255	192.9	215.5
93	100	231	155.8	146.5
94	87	210	145.7	130.5
95	231	255	246.4	249.5
96	130	231	204.3	205.5
97	87	111	95.5	95
98	115	232	177.4	183.5
99	100	234	158.6	149
100	91	231	163.5	187

B.1 Participants Speech Rate Was Observed

Table B.2 Participants Speech Rate was Observed

Participant	Min SR (WPM)	Max SR (WPM)	Mean SR (WPM)	Median SR (WPM)
51	120	145	133.3333333	135
52	124	134	128.8333333	129
53	120	135	128.1666667	129.5
54	127	135	130	129.5
55	135	140	138	139
56	132	140	136.6666667	138
57	120	135	126.1666667	126
58	130	145	137.1666667	137.5
59	120	140	130.3333333	129
60	90	98	94.66666667	95
61	140	150	144.6666667	145
62	135	140	137.5	137.5
63	123	135	129.8333333	129.5
64	90	96	93	93.5
65	138	145	140.8333333	140
66	130	137	133.6666667	135
67	125	129	126.6666667	126.5
68	100	105	101.1666667	100
69	120	129	125.3333333	126.5
70	123	130	126.6666667	126.5
71	149	150	149.6666667	150
72	100	120	107.3333333	107
73	124	140	132.5	132.5
74	120	125	122	121.5
75	110	134	122.6666667	124
76	125	130	127.8333333	128.5
77	125	130	128.1666667	129.5
78	120	130	126.3333333	126.5
79	115	125	120	120
80	120	135	126.1666667	126
81	100	123	107.1666667	105
82	120	130	125.8333333	127.5

Table B.2 continued from previous page

83	120	125	122.5	122.5
84	90	95	91.5	91
85	100	130	113.3333333	115
86	120	150	140.3333333	143.5
87	140	146	143.5	145
88	132	138	135.1666667	136
89	115	128	121.6666667	121
90	128	136	131.8333333	131
91	110	130	122.3333333	122.5
92	125	140	131.3333333	130
93	140	148	142.6666667	141.5
94	120	130	126.3333333	126.5
95	115	130	122.1666667	122.5
96	100	120	114.1666667	117.5
97	130	140	134.6666667	134.5
98	140	150	147.6666667	149
99	110	126	119	119
100	120	135	127.1666667	126.5

Appendix C

Tables of chapter 5

C.1 Experiment 1: Voice Characteristics for Scripted Speech

Table C.1 Observation of pauses and with their types for Participant 2 (English Language)

Pauses	Type of pause	Duration of pause (msec)
Pause 1	Silent	230
Pause 2	Silent	280
Pause 3 [Filled]	Insertion	290
Pause 4	Silent	420
Pause 5	Silent	300
Pause 6 [Filled]	Prolongation	280

Table C.2 Observation of pauses and with their types for Participant 2 (Native Language)

Pauses	Type of Pause	Duration of pause (msec)
Pause 1	Insertion	200
Pause 2 [Filled]	Silent	300
Pause 3	Silent	410
Pause 4 [Filled]	Insertion	290

Table C.3 Observation of pauses and with their types for Participant 3 (English Language)

Pauses	**Type of Pause**	**Duration of pause (msec)**
Pause 1 [Filled]	Prolongation	400
Pause 2	Silent	310
Pause 3 [Filled]	Repetition	520
Pause 4 [Filled]	Insertion	180
Pause 5	Silent	220

Table C.4 Observation of pauses and with their types for Participant 3 (Native Language)

Pauses	**Type of Pause**	**Duration of pause (msec)**
Pause 1	Silent	300
Pause 2 [Filled]	Insertion	450
Pause 3 [Filled]	Repetition	450

Table C.5 Observation of pauses and with their types for Participant 4 (English Language)

Pauses	**Type of Pause**	**Duration of pause (msec)**
Pause 1	Silent	470
Pause 2	Silent	380
Pause 3 [Filled]	Repetition	590
Pause 4 [Filled]	Insertion	410
Pause 5	Silent	225
Pause 6 [Filled]	Prolongation	325

Table C.6 Observation of pauses and with their types for Participant 4 (Native Language)

Pauses	**Type of Pause**	**Duration of pause (msec)**
Pause 1	Silent	420
Pause 2 [Filled]	Insertion	500
Pause 3 [Filled]	Repetition	380
Pause 4	Silent	500

Table C.7 Observation of pauses and with their types for Participant 5 (English Language)

Pauses	**Type of Pause**	**Duration of pause (msec)**
Pause 1	Silent	300
Pause 2	Silent	410
Pause 3	Silent	370
Pause 4 [Filled]	Insertion	320

Table C.8 Observation of pauses and with their types for Participant 5 (Native Language)

Pauses	**Type of Pause**	**Duration of pause (msec)**
Pause 1	Silent	471
Pause 2	Silent	378
Pause 3 [Filled]	Insertion	251

Table C.9 Observation of pauses and with their types for Participant 6 (English Language)

Pauses	**Type of Pause**	**Duration of pause (msec)**
Pause 1 [Filled]	Insertion	350
Pause 2	Silent	270
Pause 3 [Filled]	Repetition	430
Pause 4	Silent	510
Pause 5	Silent	340

Table C.10 Observation of pauses and with their types for Participant 6 (Native Language)

Pauses	**Type of Pause**	**Duration of pause (msec)**
Pause 1	Silent	400
Pause 2 [Filled]	Insertion	380
Pause 3	Silent	420

Table C.11 Observation of pauses and with their types for Participant 7 (English Language)

Pauses	**Type of Pause**	**Duration of pause (msec)**
Pause 1 [Filled]	Insertion	400
Pause 2 [Filled]	Repetition	310
Pause 3	Silent	280
Pause 4	Silent	430
Pause 5 [Filled]	Insertion	350
Pause 6 [Filled]	Prolongation	330

Table C.12 Observation of pauses and with their types for Participant 7 (Native Language)

Pauses	**Type of Pause**	**Duration of pause (msec)**
Pause 1 [Filled]	Insertion	500
Pause 2	Silent	600
Pause 3	Silent	600

Table C.13 Observation of pauses and with their types for Participant 8 (English Language)

Pauses	**Type of Pause**	**Duration of pause (msec)**
Pause 1 [Filled]	Insertion	500
Pause 2	Silent	450

Pause 3 [Filled]	Repetition	550

Table C.14 Observation of pauses and with their types for Participant 8 (Native Language)

Pauses	Type of Pause	Duration of pause (msec)
Pause 1 [Filled]	Insertion	450
Pause 2	Silent	550

Table C.15 Observation of pauses and with their types for Participant 9 (English Language)

Pauses	Type of Pause	Duration of pause (msec)
Pause 1	Silent	550
Pause 2 [Filled]	Insertion	450
Pause 3 [Filled]	Prolongation	350
Pause 4 [Filled]	Repetition	650

Table C.16 Observation of pauses and with their types for Participant 9 (Native Language)

Pauses	Type of Pause	Duration of pause (msec)
Pause 1	Silent	874
Pause 2 [Filled]	Insertion	726

Table C.17 Observation of pauses and with their types for Participant 10 (English Language)

Pauses	Type of Pause	Duration of pause (msec)
Pause 1 [Filled]	Insertion	500
Pause 2 [Filled]	Repetition	700
Pause 3	Silent	600

Pause 4 [Filled]	Prolongation	274
Pause 5 [Filled]	Insertion	126

Table C.18 Observation of pauses and with their types for Participant 10 (Native Language)

Pauses	Type of Pause	Duration of pause (msec)
Pause 1 [Filled]	Insertion	435
Pause 2	Silent	510
Pause 3 [Filled]	Insertion	355

C.2 Experiment 2: Voice Characteristics for Unscripted Speech

Table C.19 Observation of pauses and with their types for audience member 2 (English Language)

Pauses	Type of Pause	Duration of a pause (msec)
Pause 1 [Filled]	Insertion	200
Pause 2	Silent	260
Pause 3	Silent	360
Pause 4 [Filled]	Prolongation	400
Pause 5	Silent	250
Pause 6	Silent	280
Pause 7	Silent	350
Pause 8	Silent	320
Pause 9 [Filled]	Insertion	240
Pause 10 [Filled]	Repetition	450

Table C.19 continued from previous page

Pause 11	Silent	310
Pause 12	Silent	187
Pause 13 [Filled]	Prolongation	195
Pause 14	Silent	178
Pause 15	Silent	180
Pause 16	Silent	180
Pause 17 [Filled]	Prolongation	370
Pause 18 [Filled]	Repetition	300

Table C.20 Observation of pauses and with their types for audience member 2 (Tamil Language)

Pauses	**Type of Pause**	**Duration of pause (msec)**
Pause 1	Silent	300
Pause 2	Silent	260
Pause 3 [Filled]	Prolongation	277
Pause 4 [Filled]	Insertion	396
Pause 5	Silent	300
Pause 6	Silent	279
Pause 7	Silent	310
Pause 8	Silent	350
Pause 9 [Filled]	Insertion	400
Pause 10	Silent	360

Table C.21 Observation of pauses and with their types for audience member 3 (English Language)

Pauses	Type of Pause	Duration of pause (msec)
Pause 1 [Filled]	Insertion	345
Pause 2 [Filled]	Repetition	278
Pause 3	Silent	498
Pause 4	Silent	425
Pause 5 [Filled]	Prolongation	520
Pause 6 [Filled]	Prolongation	480
Pause 7	Silent	460
Pause 8	Silent	482
Pause 9	Silent	322
Pause 10 [Filled]	Repetition	370
Pause 11 [Filled]	Insertion	680
Pause 12 [Filled]	Insertion	410
Pause 13	Silent	470
Pause 14 [Filled]	Repetition	410
Pause 15 [Filled]	Insertion	600

Table C.22 Observation of pauses and with their types for audience member 3 (Tamil Language)

Pauses	Type of Pause	Duration of pause (msec)
Pause 1	Silent	670
Pause 2	Silent	510
Pause 3	Silent	438

Table C.22 continued from previous page

Pause 4 [Filled]	Insertion	537
Pause 5	Silent	515
Pause 6	Silent	480
Pause 7 [Filled]	Prolongation	390
Pause 8	Silent	670

Table C.23 Observation of pauses and with their types for audience member 4 (English Language)

Pauses	**Type of Pause**	**Duration of pause (msec)**
Pause 1 [Filled]	Prolongation	115
Pause 2	Silent	215
Pause 3 [Filled]	Repetition	180
Pause 4	Silent	113
Pause 5	Silent	193
Pause 6 [Filled]	Insertion	147
Pause 7 [Filled]	Insertion	210
Pause 8	Silent	330
Pause 9 [Filled]	Prolongation	110
Pause 10 [Filled]	Repetition	247
Pause 11 [Filled]	Insertion	180
Pause 12 [Filled]	Insertion	132
Pause 13	Silent	315
Pause 14	Silent	270

Table C.23 continued from previous page

Pause 15 [Filled]	Insertion	178
Pause 16 [Filled]	Insertion	150

Table C.24 Observation of pauses and with their types for audience member 4 (Tamil Language)

Pauses	**Type of Pause**	**Duration of pause (msec)**
Pause 1	Silent	136
Pause 2	Silent	176
Pause 3 [Filled]	Insertion	110
Pause 4 [Filled]	Prolongation	150
Pause 5	Silent	183
Pause 6 [Filled]	Repetition	210
Pause 7 [Filled]	Insertion	221
Pause 8	Silent	174
Pause 9 [Filled]	Insertion	200

Table C.25 Observation of pauses with their types for audience member 5 (English Language)

Pauses	**Type of Pause**	**Duration of pause (msec)**
Pause 1	Silent	300
Pause 2	Silent	250
Pause 3 [Filled]	Insertion	410
Pause 4 [Filled]	Insertion	330
Pause 5 [Filled]	Prolongation	280

Table C.25 continued from previous page

Pause 6	Silent	260
Pause 7	Silent	330
Pause 8 [Filled]	Repetition	175
Pause 9	Silent	390
Pause 10 [Filled]	Insertion	430
Pause 11 [Filled]	Insertion	350
Pause 12	Silent	250
Pause 13	Silent	410
Pause 14 [Filled]	Prolongation	200
Pause 15 [Filled]	Prolongation	430
Pause 16	Silent	185
Pause 17 [Filled]	Insertion	250

Table C.26 Observation of pause and with their types for audience member 5 (Tamil Language)

Pauses	**Type of Pause**	**Duration of pause (msec)**
Pause 1	Silent	520
Pause 2	Silent	400
Pause 3 [Filled]	Prolongation	320
Pause 4	Silent	260
Pause 5 [Filled]	Prolongation	280
Pause 6	Silent	470
Pause 7	Silent	400
Pause 8 [Filled]	Insertion	620

Table C.26 continued from previous page

Pause 9 [Filled]	Insertion	400

Table C.27 Observation of pauses and with their types for audience member 6 (English Language

Pauses	**Type of Pause**	**Duration of pause (msec)**
Pause 1	Silent	200
Pause 2	Silent	150
Pause 3	Silent	170
Pause 4 [Filled]	Insertion	130
Pause 5 [Filled]	Prolongation	190
Pause 6	Silent	110
Pause 7	Silent	170
Pause 8	Silent	100
Pause 9 [Filled]	Insertion	125
Pause 10 [Filled]	Repetition	115
Pause 11	Silent	165
Pause 12 [Filled]	Insertion	135
Pause 13	Silent	120
Pause 14	Silent	140
Pause 15	Silent	170
Pause 16	Silent	130
Pause 17 [Filled]	Insertion	180
Pause 18	Silent	100
Pause 19	Silent	250
Pause 20	Silent	320
Pause 21	Silent	200
Pause 22	Silent	140

Table C.28 Observation of pauses and with their types for audience member 6 (Tamil Language)

Pauses	**Type of Pause**	**Duration of Pause**
Pause 1	Silent	150
Pause 2	Silent	250
Pause 3	Silent	400
Pause 4 [Filled]	Insertion	150
Pause 5	Silent	170
Pause 6	Silent	200
Pause 7	Silent	180
Pause 8	Silent	160
Pause 9 [Filled]	Repetition	100
Pause 10	Silent	150
Pause 11	Silent	180
Pause 12	Silent	230

Table C.29 Observation of pauses and with their types for audience member 7 (English Language)

Pauses	**Type of Pause**	**Duration of pause (msec)**
Pause 1	Silent	420
Pause 2	Silent	280
Pause 3	Silent	345
Pause 4	Silent	287
Pause 5 [Filled]	Insertion	414
Pause 6 [Filled]	Prolongation	378
Pause 7	Silent	230
Pause 8	Silent	170
Pause 9 [Filled]	Repetition	210
Pause 10 [Filled]	Insertion	170

Table C.29 continued from previous page

Pause 11	Silent	480
Pause 12 [Filled]	Insertion	620
Pause 13 [Filled]	Insertion	183
Pause 14 [Filled]	Repetition	247
Pause 15	Silent	530
Pause 16	Silent	283
Pause 17	Silent	430
Pause 18 [Filled]	Insertion	118
Pause 19	Silent	215
Pause 20	Silent	220

Table C.30 Observation of pauses and with their types for audience member 7 (Tamil Language)

Pauses	**Type of Pause**	**Duration of pause (msec)**
Pause 1	Silent	428
Pause 2	Silent	335
Pause 3	Silent	170
Pause 4 [Filled]	Insertion	430
Pause 5	Silent	185
Pause 6 [Filled]	Repetition	226
Pause 7	Silent	184
Pause 8	Silent	300
Pause 9 [Filled]	Insertion	350
Pause 10 [Filled]	Insertion	400
Pause 11	Silent	525
Pause 12	Silent	343

Table C.30 continued from previous page

Pause 13	Silent	210
Pause 14 [Filled]	Insertion	485

Table C.31 Observation of pauses and with their types for audience member 8 (English Language)

Pauses	**Type of Pause**	**Duration of pause (msec)**
Pause 1 [Filled]	Insertion	526
Pause 2 [Filled]	Insertion	480
Pause 3	Silent	435
Pause 4	Silent	410
Pause 5	Silent	550
Pause 6	Silent	383
Pause 7 [Filled]	Insertion	447
Pause 8	Silent	286
Pause 9 [Filled]	Repetition	237
Pause 10	Silent	347
Pause 11	Silent	570
Pause 12	Silent	497
Pause 13 [Filled]	Insertion	380
Pause 14	Silent	414

Table C.32 Observation of pauses and with their types for audience member 8 (Tamil Language)

Pauses	**Type of Pause**	**Duration of pause (msec)**
Pause 1	Silent	380
Pause 2 [Filled]	Prolongation	428

Table C.32 continued from previous page

Pause 3	Silent	533
Pause 4 [Filled]	Insertion	350
Pause 5 [Filled]	Repetition	430
Pause 6	Silent	290
Pause 7	Silent	479

Table C.33 Observation of pauses and with their types for audience member 9 (English Language)

Pauses	**Type of Pause**	**Duration of pause**
Pause 1	Silent	428
Pause 2	Silent	238
Pause 3 [Filled]	Repetition	515
Pause 4 [Filled]	Insertion	230
Pause 5 [Filled]	Insertion	170
Pause 6	Silent	210
Pause 7 [Filled]	Prolongation	140
Pause 8 [Filled]	Insertion	280
Pause 9 [Filled]	Insertion	407
Pause 10	Silent	693
Pause 11 [Filled]	Repetition	390
Pause 12	Silent	510
Pause 13 [Filled]	Insertion	487
Pause 14 [Filled]	Insertion	297

Table C.33 continued from previous page

Pause 15	Silent	580
Pause 16	Silent	250
Pause 17 [Filled]	Prolongation	300
Pause 18	Silent	275

Table C.34 Observation of pauses and with their types for audience member 9 (Tamil Language)

Pauses	**Type of Pause**	**Duration of pause (msec)**
Pause 1	Silent	470
Pause 2 [Filled]	Insertion	525
Pause 3 [Filled]	Insertion	780
Pause 4 [Filled]	Prolongation	890
Pause 5	Silent	285
Pause 6	Silent	520
Pause 7	Silent	370
Pause 8 [Filled]	Insertion	410

Table C.35 Observation of pauses and with their types for audience member 10 (English Language)

Pauses	**Type of Pause**	**Duration of pause (msec)**
Pause 1	Silent	180
Pause 2	Silent	147
Pause 3 [Filled]	Insertion	171
Pause 4 [Filled]	Prolongation	197
Pause 5 [Filled]	Prolongation	185

Table C.35 continued from previous page

Pause 6 [Filled]	Insertion	230
Pause 7	Silent	310
Pause 8	Silent	170
Pause 9	Silent	210
Pause 10	Silent	114
Pause 11 [Filled]	Repetition	150
Pause 12	Silent	174
Pause 13	Silent	117
Pause 14 [Filled]	Insertion	250
Pause 15 [Filled]	Prolongation	380
Pause 16	Silent	135
Pause 17	Silent	280
Pause 18	Silent	420
Pause 19	Silent	170
Pause 20 [Filled]	Insertion	287
Pause 21 [Filled]	Prolongation	310
Pause 22 [Filled]	Prolongation	173

Table C.36 Observation of pauses and with their types for audience member 10 (Tamil Language)

Pauses	**Type of Pause**	**Duration of pause (msec)**
Pause 1 [Filled]	Prolongation	170
Pause 2	Silent	183
Pause 3	Silent	210
Pause 4 [Filled]	Insertion	140

Table C.36 continued from previous page

Pause 5	Silent	370
Pause 6	Silent	190
Pause 7 [Filled]	Repetition	370
Pause 8 [Filled]	Prolongation	410
Pause 9	Silent	225
Pause 10	Silent	252