UNIVERSITÉ DE LORRAINE

INSTITUT DES SCIENCES DU DIGITAL MANAGEMENT AND COGNITION

MASTER THESIS

Investigating the speaker information carried on by the fundamental frequency

Author:
Shalini Priya

Supervisors:
Pierre CHAMPION
Denis JOUVET
Hubert NOURTEL

A thesis submitted in fulfillment of the requirements for the degree of Master of Science in Natural Language Processing

in the

MULTISPEECH
INRIA, Nancy - Grand Est

August 23, 2022

Declaration of Authorship

I, Shalini Priya, declare that this thesis titled, "Investigating the speaker information carried on by the fundamental frequency" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Shalini Priya

Date: August 23, 2022

“Thanks to my alma mater for sharing a solid foundation, today I understand the importance of words and speech, which I can use to contribute in the field of technology and help build our future.”

Shalini Priya

UNIVERSITÉ DE LORRAINE

Abstract

Institut des sciences du Digital Management and Cognition

Master of Science in Natural Language Processing

Investigating the speaker information carried on by the fundamental frequency

by Shalini Priya

The human voice is made up of different factors, making each voice unique. The speech signal contains many levels of information. Primarily a message is conveyed via the spoken words. Besides the linguistic content itself, a speech signal conveys a lot of information about the speaker, including its identity, gender, age as well as stress, emotions, thus making speech a very personal data. Because of that, there are some studies that investigated how speech data can be anonymized. The ANR project *DEEP- PRIVACY*[1] is one such project. One goal of the project is to anonymize (Anonymization refers to the goal of suppressing personally identifiable attributes of the speech signal, leaving all other attributes intact) speech data to allow for an easier sharing of speech data, which is necessary for large data collection for improving speech technologies.

Currently, the most relevant approach is based on a voice conversion system. A voice conversion system modifies a speech signal to sound as if pronounced by another speaker [13]. In the referred approach, one of the values which is extracted from the speech signal is Fundamental Frequency. The main goal of this thesis was to study Fundamental Frequency (also known as F0), which is determined by the rate of vibration of the vocal folds of the speaker, to investigate the amount of information carried on by it by performing various experiments by anonymizing F0 values. We used SIDEKIT toolkit for carrying out experiments and code was added to extract F0 using the YAAPT package. Experimental results showed that the approach is effective in concealing the speaker's information. It increases the rate of Equal Error Rate while maintaining privacy.

Keywords - Fundamental Frequency, Speaker Recognition, X-vector, Equal Error Rate, Speaker Verification

[1] https://project.inria.fr/deepprivacy/

Acknowledgements

I would like to express my sincere gratitude to the amazing people without whom this work would have not been possible.

I want to start by thanking my supervisors Mr. Denis Jouvet, Pierre Champion and Hubert Nourtel for their mentorship, help and advices. I am extremely honored to have been part of such a great team. Thank you for the time you dedicated to respond to my questions, helped me with the project, report, valuable feedback and insights.

I also want to thank all the employees working at Inria where I met people from different backgrounds and nationality. I got to learn new things and met amazing people.

I also would like to thank Miguel Couceiro and Maxime Amblard. Thank you for being a great teacher and for assisting us throughout the duration of masters.

I would like to take an advantage of this opportunity to thank the entire faculty and staff of INRIA, our dearly beloved school, IDMC for their efforts to ensure a strong program and training that contributed in the success of the internship and played a major part in my professional growth.

I also send my sincere gratitude to the members of the jury Mr. Maxime Amblard, Romain Serizel and Miguel Couceiro and my reviewer for their presence and the time they took from their busy schedule to witness my work.

Finally, I want to thank each one of you who helped me during my internship and motivated me to work hard and deliver a valuable work.

Thank you all from the bottom of my heart. I believe that this work will live up to your expectations.

Contents

List of Figures

List of Tables

List of Abbreviations

ASV	**A**utomatic **S**peaker **V**erification
FAR	**F**alse **A**cceptance **R**ate
FRR	**F**alse **R**ejection **R**ate
EER	**E**qual **E**rror **R**ate

I would like to dedicate this work to:

My Parents

I appreciate your efforts, love, better future planning, hardships and sacrifices in bringing us up to be better individuals. Thank you for nurturing my life and being by my side during my failures as well as successes. I could not have asked you for more. Thank you mom and dad for being an advocate for our education. I can always count on you to help me in my quest to acquire better skills, personal and professional responsibilities. Thanks for setting aside the time for us to experience the realm of life.

My Brother

Thanks for being supportive always and giving me valuable suggestions. I can't wait to see you grow and want to be a part of each of your successes.

My Friends

I admit friends play an important role in your life. I am blessed to have such good heads and souls around me. You have been my guiding light throughout my journey. I learnt from each one of you. I wish I could talk more about each one of you. Thank you Sandeep Mishra, Srikanth Kyathegowda and Sunil Sahu for helping me make good decisions to excel in my career.

My Teachers

To all the amazing teachers who helped me with lessons and learnings throughout this beautiful journey.

Shalini

Well, you made it! You have come a very long way by overcoming obstacles, pain, sleeepless nights. This entire duration was worth. You are about to end a very beautiful journey and there are more to come your way. So buckle up, you have next adventurous journey ahead of you.

Chapter 1

General context of the project

This chapter focuses on the overview of the working environment at Inria, detailed introduction of the project and thesis structure.

1.1 Internship Environment

1.1.1 Inria - France

Inria [1] stands for "Institut National de Recherche en Informatique et en Automatique", is the French national research institute for digital science and technology. It is the public research platform for France's digital power. It's known for research, innovation and entrepreneurial risk. There are around 200 projects consisting of more than 3,900 researchers and engineers who explore new paths, often in an interdisciplinary manner in collaboration with industrial partners and major research universities to meet ambitious challenges. As a technological institute, Inria shows diversity in innovation right from the open source software publishing to the creation of technological startups. An agile project team environment model promotes scientific excellence, technological development and innovation.

FIGURE 1.1: Logo of the Inria research institute
Source - *Inria website* [1]

The 200 project-teams are distributed in nine research centers located throughout France as shown in Figure 1.2. Five partnerships were signed in the second half of 2021 with the Université Côte d'Azur, the University of Bordeaux, the Institut Polytechnique de Paris, the University of Paris-Saclay and the University of Lille, respectively. These new types of agreements herald the creation of Inria centers at the relevant universities, where the institute proposes to operate its research and innovation facilities on behalf of its partners as part of a shared strategy and joint implementation. Inria is also in close proximity with its academic partners (more than 45 universities and other research bodies) and is invested in local innovation

[1] https://www.inria.fr/en/inria-ecosystem

innitiatives and also a partner of several academic institutions in United States, Asia, Latin America, Africa and Middle Est. Inria's scientific projects are integrated into the heart of the major university clusters which stimulates fruitful partnerships with academic players and with companies, and are also the crucible of technological startups.

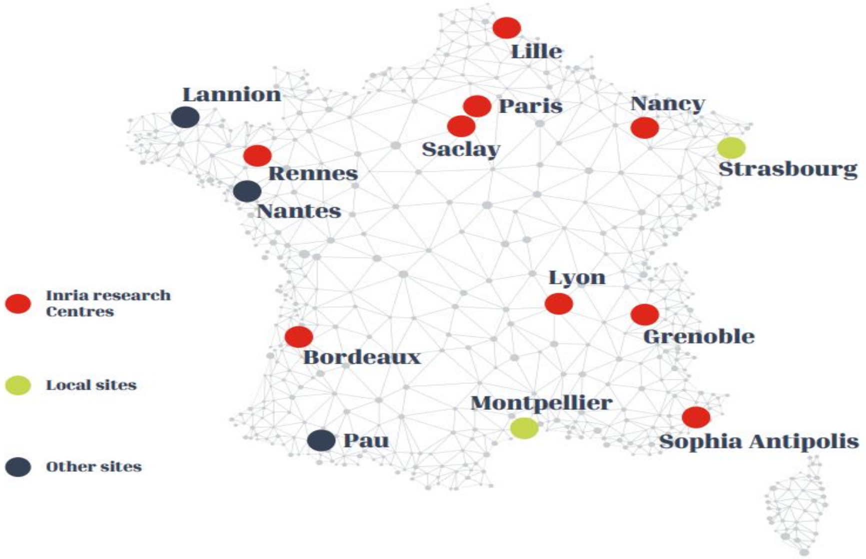

FIGURE 1.2: Inria Research Centers
Source - Inria annual report - 2021

1.1.2 Mission

The main missions of Inria are listed below:

- Maintaining scientific excellence by encouraging scientific themes and adding new "disciplines" for digital technology
- To accelerate innovation by developing strategic relationship with major players in French and European industry
- To contribute in the implementation of public policies such as National Artificial Intelligence Research Program
- A driving force for digital culture education by developing an innovative training program for education and training community
- An exceptional work environment to support collective vitality
- To support the Inria foundation by developing digital sciences in the society

1.1.3 Centre Inria Nancy - Grand Est

Inria Nancy-Grand Est research centre was created in 1986 and is located in Nancy, Strasbourg and Saarbrucken. The centre has 20 project teams, divided between Nancy, Strasbourg and Saarbrücken. It has around 400 people comprising of scientists and research innovation support staff belonging to 45 different nationalities.

It also supports scientists and engineers in the creation of start-ups in various fields such as marketing, 3D simulation and computer security. Few of it's startups are listed below:

- Sonaide - builds intelligent device for preventing loss of autonomy
- Nijta - builds privacy preserving technology to safeguard voice data
- Pulse Audition - provides solution for pulse frames with AI powered speech enhancement

FIGURE 1.3: Centre Inria Nancy - Grand Est

Grand Est region is ambitious to become one of the European leaders by researching on several projects of the Factory of the Future like additive manufacturing, robotics, meshes, speech etc.

In addition to meeting it's technological needs, it also encourages recreational activities like yoga, indoor and outdoor games, music etc as well as self development programs like summer school in various topics.

1.1.4 Team - MultiSpeech

I am a part of *MULTISPEECH*[2] team which is a joint research team between the Université of Lorraine, Inria, and CNRS. It is part of department D4 "Natural language and knowledge processing" of LORIA. Its research focuses on speech processing, with particular emphasis to **multisource** (source separation, robust speech recognition), **multilingual** (computer assisted language learning), and **multimodal** aspects (audiovisual synthesis).

There are **8 Researchers, 4 Faculty Members, 3 Post Doctoral Fellows, 15 PhD students, 4 Co-supervised PhD students, 5 Engineers and 12 Interns currently** in *MultiSpeech Team* [3] led by Denis Jouvet.

[2] https://www.inria.fr/en/multispeech

[3] https://team.inria.fr/multispeech/team-members/

TABLE 1.1: Project Team

Name	Role
Denis JOUVET	Team Leader
Pierre CHAMPION	PhD student
Hubert NOURTEL	Engineer
Shalini Priya	Intern

I worked in collaborations with *Pierre Champion*, a PhD student, working on the ANR DEEP-PRIVACY project, *Denis Jouvet* who is a Team Leader and a Senior Researcher, *Hubert Nourtel* an Engineer, also working on the ANR DEEP-PRIVACY project. The thesis topic was founded by ANR DEEP-PRIVACY project. Table 1.1 shows my project team.

Weekly meeting was conducted to track the progress of my work and also to address action items, blockers and questions for the upcoming weeks. These frequent review meetings helped me to stay on track and complete the project within allocated time.

1.2 Presentation of the Project

1.2.1 Introduction

As proposed in the paper [35], any sound produced by humans to communicate meanings, ideas, opinions, etc., is called a voice. In a specific term, voice is defined as any sound produced by vocal folds vibration, which occurs when air is under pressure from the lungs. Voice is the most natural communication tool used by humans. It conveys the speaker's traits, such as ethnicity, age, gender, and feeling. Every individual's voice is unique due to the differences in the shapes of the vocal tract, larynx sizes, and other parts of human voice production organs. The features of voice are dependent on its pace or speed, volume, pitch level, and quality, while the articulation rate and speech pauses rely on the speaker's speaking style. According to the author in [36], there are three sources of variations among speakers:

- Differences in vocal folds and vocal tract shape
- Differences in speaking style (including accent)
- Differences in how speakers express themselves (words or phrases used) to convey a particular message

The pitch is an integral part of the human voice and is defined as the "rate of vibration of the vocal folds". The sound of the voice changes as the rate of vibrations varies. Researchers in the field of speech science have been showing interest in analyzing the voice parameters, such as pitch which is the subjective perception of Fundamental Frequency and Fundamental Frequency which describes the actual physical phenomenon.

Prosodic characteristics [27] such as rhythm, stress and intonation in speech conveys some important information. Since each speaker has unique physiological characteristics of speech production and speaking style, speaker-specific characteristics are also reflected in prosody Fundamental frequency (F0) is the most frequently investigated prosodic parameter as per the author in [22].

The two main tasks of Speaker Recognition are, Speaker Identification (identifying a person from the group) and Speaker Verification (authenticating a claimed identity). Deep Privacy project aims at sharing anonymized speech signals that are useful for training Speech Recognition models, but do not carry personal information of the speaker. Current approaches are based on voice conversion systems (i.e., modifying a speech signal so that it sounds as if pronounced by another speaker). To assess how our speech signals are anonymized we use Speaker Verification.

This document represents the summary of the work done as part of my Internship project of my Masters program at Inria. The duration of the internship was five months (1st April, 2022 to 31st August, 2022) which was conducted to finish my training at the IDMC in Natural Language Processing. My intervention took place within the MultiSpeech team in the project which focuses on privacy-preserving deep-learning tasks for speech processing. The project's goal was to investigate Fundamental Frequency values using deep learning based speaker recognition or verification experiments. In this context, the mission of my internship was to add pieces of code to extract Fundamental Frequency values from the speech signal as well as to optimize the deep learning architecture. In order to do so, we decided to start by researching and reading about Speaker Recognition and Fundamental Frequency topics, the new state of the art, datasets, evaluation of the model, results already achieved. Throughout the document at hand, the steps we followed and the tools we used in this project, are presented.

1.2.2 Motivation and Context

As stated in the previous section, Fundamental Frequency is the most studied prosodic parameter that characterizes speech. The goal is to analyze the influence of F0 in Speaker Recognition. In particular, we aimed to incorporate F0 as a feature for Speaker Verification. In addition, we tried different F0 anonymization modification schemes for evaluation. The study of this method was done under various conditions to deduce the effectiveness of privacy.

1.2.3 Problem Statement

There is an increased use of internet and digitalization in the last few years because of the rapidly changing technologies which in turn has increased an amount of personal information that are being uploaded to the Internet. Due to the use of popular social media platforms and other picture/video sharing services, people upload (or stream) their self-portraits, voice samples and video clips much more easily than in the past. The general public are unaware that their face photos, videos and voice samples contain biometric traits to identify the Speakers. Sometimes, these biometric identity can be misused to attack Speaker Verification system [51] [52] [49] [14].

1.2.4 Literature Review

Literature reviews are one of the essential means for data collections in research work and it has contributed enormously in many years in the field of research environment. An extensive scientific literature and journals has been reviewed in this thesis report, aiming to discover what has been researched on Speaker Recognition . The main sources of data collection are as follows:

- Scientific Papers and Journal from the internet (www.researchgate.net, www.sciencedirect.com, https://www.ieee.org/)
- Papers published by MultiSpeech team (Denis Jouvet, Pierre Champion)
- Previous students project report from my University and other reports from the same domain to understand the flow, structure and motivation for writing the report
- Book pdf's to understand the topic in detail

  .

1.2.5 Thesis Structure

The remaining chapters of the thesis are organized as follows: In Chapter 2, we present an overview of the speech production process from the articulatory point of view. Chapter 3, covers the vocal organs mainly responsible for Fundamental Frequency. Chapter 4, gives an overview of Speaker Verification model. Chapter 5, describes the proposed approach and it's related aspects. Chapter 6, describes the experimental setup and results of this thesis. Chapter 7, concludes with a summary of the analysis of this work and future work.

Chapter 2

Speech Production Mechanism

Speech is a natural form of communication for human beings with the ability to understand. We use speech every day almost unconsciously, but an understanding of the mechanisms on which it is based will help to clarify how the brain processes information.

2.1 Phsychological aspects of human voice

Speech is produced as a sequence of sounds. The articulators such as jaw, tongue, velum, lips, mouth and their shapes, sizes and positions changes over time to produce sounds. Speech production can be divided into three stages [30]:

1. **Conceptualization** - Speech actually starts from our brain as a thought process. This process is known as conceptualization
2. **Formulation** - Second stage is speech formulation. In formulation stage, our thought is converted into linguistic form. This is again divided into two steps as shown in Figure 2.1

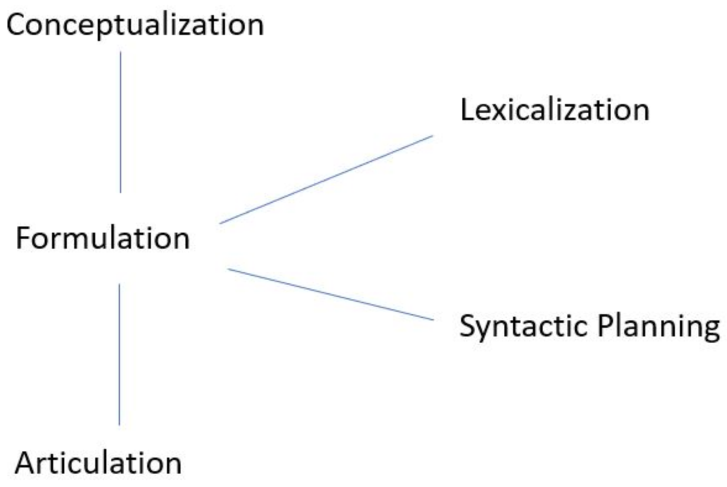

FIGURE 2.1: Physchological aspects of speech production [30]

- **Lexicalization** - Thoughts gets converted into appropriate words

- **Syntactic Planning** - Appropriate words will be arranged in the right way syntactically

3. **Articulation** - is the last stage of speech production. The sound is produced to convey message

2.2 Biological aspects of speech production

The main interest in learning speech production is that how different articulators such as tongue, lips, jaw and other speech organs are involved in making sound because the differences in voice of different speakers lies in the construction of their articulatory organs. Speech can be defined as waves of air pressure created by airflow pressed out of the lungs and going out through the mouth and nasal cavities. The air passes through the vocal folds (cords) via the path from the lungs through the vocal tract, vibrating them at different frequencies. A simplified diagram of human vocal system is shown in the Figure 2.2.

Some of the main articulators and their functions are explained below:

- Pharynx - It is a tube which begins just above the larynx and it is about 7cm in women and about 8cm in men. The primary function of this articulator is to convert a relatively steady flow of air out from the lungs into a puff of air as the glottis opens or closes
- Larynx - This organ of speech contains the vocal folds. It is called as "voice box" which can act as a resonator for any sound produced
- Vocal folds - The vocal folds are responsible for the voicing of speech sounds. Sounds are either voiced (as for example, the vowels) or unvoiced (as for example, /p/, /t/, /k/, /s/, ...)
- Velum - The velum is a thin sheet composed of muscle fibres, tissue, blood vessels, nerves, and glands. It's main function is to separate the nasal cavity from the oral cavity (the mouth). If the velum is raised, it prevents air from going through the nose. If the velum is lowered, air passes through both the nose and mouth
- Palate - Palate forms the entire roof of the mouth that separates the oral cavity from the nasal cavity
- Vocal tract - The air passages above the larynx are known as the vocal tract. The vocal tract can be divided into the oral cavity (the mouth and pharynx), and the nasal cavity (within the nose)

An adult vocal tract is approximately 17 cm long and corresponds to the speech production organs above the vocal folds. The speech production organs includes [39]:

- the laryngeal pharynx (below the epiglottis)
- oral pharynx (behind the tongue, between the epiglottis and velum)
- oral cavity (forward of the velum and bounded by the lips, tongue, and palate)
- nasal pharynx (above the velum, rear end of nasal cavity)

- the nasal cavity (above the palate and extending from the pharynx to the nostrils)
- The larynx comprises of the vocal folds. The area between the vocal folds is called the glottis

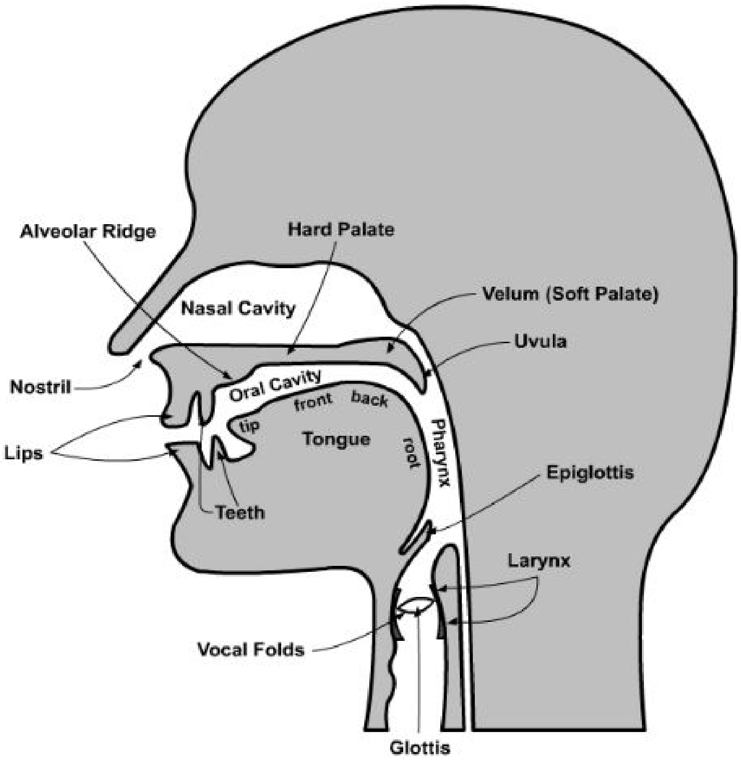

FIGURE 2.2: Human Vocal System [41]

The resonance of the vocal tract alters the spectrum of the acoustic as it passes through the vocal tract. Vocal tract resonances are called formants. Therefore, the vocal tract shape can be estimated from the spectral shape (formant location) of the voice signal. The spectral shape of the speech signal is determined by the shape of the vocal tract (the pipe formed by your throat, tongue, teeth and lips). By changing the shape of the pipe (and in addition opening and closing the air flow through your nose) the spectral shape of the speech signal changes, thus articulating different speech sounds.

Speaker Recognition systems use features generally derived only from the vocal tract. The excitation source of the human vocal also contains speaker specific information. The excitation is generated by the airflow from the lungs, which thereafter passes through the trachea and then through the vocal folds. The excitation is classified as *phonation (or voicing), whispering, frication, compression, vibration or a combination of these* [39].

Phonation (or voicing) excitation is caused when airflow is modulated by the vocal folds. When the vocal folds are closed, pressure builds up underneath them until they blow apart. The folds are drawn back together again by their tension, elasticity and the Bernoulli effect. The oscillation of vocal folds causes pulsed stream excitation of the vocal tract. The frequency of oscillation is called the **Fundamental Frequency**. This is described in detail in chapter 3. The pitch period (T0) is the time interval of each cycle of the vocal fold vibration. The fundamental frequency (F0) or the rate of vibration of the vocalfolds is the inverse of the pitch period (T0). Estimation of F0 is needed in applications such as speech synthesis, voice conversion, gender recognition, speech coding, speaker recognition and speech recognition [3].

Whispered excitation is caused by the flow of air rushing through a small triangular opening between the arytenoids cartilages at the rear of the nearly closed vocal folds. A turbulent airflow results after this, which has a wide band noise characteristic.

Frication excitation is caused due to the constrictions in the vocal tract. The shape of the broadband noise excitation depends upon the place, shape, and degree of constriction. The spectral concentration generally increases in frequency when the constriction moves forward. Sounds that are generated by friction are called fricatives. Frication can occur with or without phonation (or voicing).

Compression excitation is produced from release of a completely closed and pressurized vocal tract. This results in a silence (in the pressure accumulation phase) followed by a short noise burst. If the release is sudden, a stop or plosive is generated. If the release is gradual, an affricate is formed.

Vibration excitation is a result of air being forced through a closure other than the vocal folds, especially at the tongue.

Speech produced by phonated excitation is called voiced, speech produced by phonated excitation plus frication is called mixed voice and speech produced by other types of excitation is called unvoiced.

Chapter 3

Fundamental Frequency

This chapter provides an overview of characteristics of vocal folds and mechanism of vocal folds vibration in producing Fundamental Frequency.

3.1 Dynamics of vocal folds

The human vocal organs depicted in Figure 2.2 gives a basic understanding of how speech is produced. From the point of view of F0 considerations, the most important part of the human vocal system is the larynx, which contains the vocal cords, also known as the vocal folds. Vocal sound is created by the opening and closing of the vocal folds, caused by airflow from the lungs. It is the activity of the vocal folds that determines whether speech is produced as *"breathing","voiced", and "unvoiced"* [17].

The vocal folds are two masses of flesh which stretch between the front and back of the larynx, as illustrated in Figure 3.1. The folds are about 15 mm long in men and 13 mm long in women. The glottis is the slit-like space between the two folds.The folds are fixed at the front of the larynx and attached to the stationary thyroid cartilage. The folds are free to move at the back and sides of the larynx. The size of the glottis is controlled by the arytenoid cartilages and also by the muscles within the folds. Another important property of the vocal folds, in addition to the size of the glottis, is their tension. The tension is controlled primarily by muscle within the folds, as well as the cartilage around the folds.

There are three primary states of the vocal folds: *breathing, voiced, and unvoiced*.

- In breathing state, the glottis is wide open and arytenoid cartilages are streched outward as shown in the right part of the Figure 3.1. So, the air from the lungs flows freely through the glottis with almost no hindrance by the vocal folds
- In the voicing state, the arytenoid cartilages move towards one another. The vocal folds are brought close together as shown in the left part of the Figure 3.1. The air passing through the elastic vocal folds causes them to vibrate, this type of phonation is called as voicing. This vibration is not caused by muscles alone as it vibrates typically at over 100 vibrations per second which is too fast for the muscles to produce it [4]

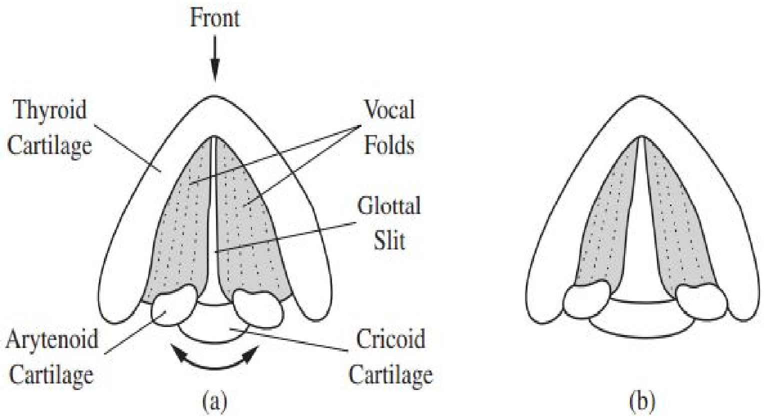

FIGURE 3.1: Downward looking view of human larynx (a) voicing (b) breathing [17]

The paper [19] proposed two basic principles of the myo-elastic-aerodynamic theory of voice production. First, he suggested that the Fundamental Frequency of vocal folds vibration (the rate at which vibration occurs) is determined by a number of interdependent factors including the mass and viscoelasticity of the vocal folds and the subglottal pressure. Secondly, he proposed that during phonation, the vocal folds are driven into vibration by forces that are explained by Bernoulli's principle.

Figure 3.1 shows the simplified sketch of the vocal folds in cross section, seen from the front.

Figure 3.2 describes the Bernoulli effect in vocal folds. The average pressure below the folds supplied from the lungs is greater than the average pressure above the folds. In Figure 3.2 (a), the folds are closed and the pressure below the folds tends to force them upwards and apart, as seen in (b, c). The vocal folds are elastic and under tension so, when stretched, they tend to return to their starting position (d, a).

First, consider the air pressures. During phonation, the pressure inside the glottis (i.e. between the folds) must usually be rather lower than the pressure below the glottis because that pressure difference has accelerated the air to produce a high-speed air jet: the blue arrow in (b). (This is due to Newton's second law and is sometimes called the *Bernoulli effect*[1]. As the jet of air leaves the glottis, it loses much of its kinetic energy which brings them back together.

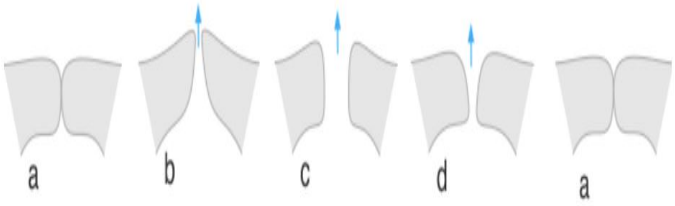

FIGURE 3.2: Vocal folds in cross section, seen from in front [20]

[1] http://hyperphysics.phy-astr.gsu.edu/hbase/pber.htmlbeq/

So, muscles do not directly vibrate the vocal folds. However, muscles in the larynx contribute to the control of vibration, by determining how much the folds are pushed together and how much they are stretched along their length.

- The last state of the vocal folds is unvoicing. In the unvoiced state, the folds are closer together and more tense than in the breathing state, thus air pressure is generated at the folds themselves

3.2 Pitch Period

Typically, with the folds in a closed position, the flow begins slowly, builds up to a maximum, and then quickly decreases to zero when the vocal folds abruptly shuts. The time interval during which the vocal folds are closed, and no flow occurs, is referred to as the glottal closed phase. The time interval over which there is nonzero flow and up to the maximum of the airflow velocity, which results in widest opening point of vocal folds is referred to as the glottal open phase, and the time interval from the airflow maximum to the time of glottal closure is referred to as the return phase. Here, there are two forces that work for reclosing the vocal folds, one, the elastic recoil of the folds and the Bernoulli effect of rushing air below the glottis. The time duration of one glottal cycle is referred to as the *pitch period* and the reciprocal of the pitch period is the corresponding pitch, also referred to as the Fundamental Frequency as shown in Figure 3.3. In conversational speech, during vowel sounds, the number of pitch periods changes with numerous factors such as stress and speaking rate. The rate at which the vocal folds oscillate through a closed, open, and return cycle is influenced by many factors. These include vocal folds muscle tension (as the tension increases, so does the pitch), the vocal fold mass (as the mass increases, the pitch decreases because the folds are more sluggish), and the air pressure behind the glottis in the lungs and trachea, which might increase in a stressed sound or in a more excited state of speaking (as the pressure below the glottis increases, so does the pitch). The specific flow shape can change with the speaker, the speaking style, and the specific speech sound.

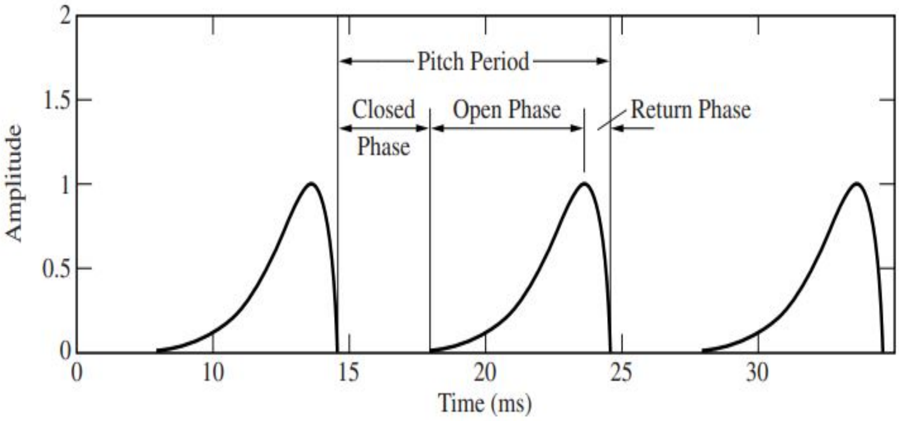

FIGURE 3.3: Illustration of periodic glottal airflow velocity [17]

3.3 What is Fundamental Frequency?

The Fundamental Frequency (F0) is the number of cycles per unit time in the vibration of the vocal folds. It is mutually reciprocal to period $(T0) : F0 = 1/T0$ and $T0 = 1/F0$. One vocal cycle typically consists of closed, open and return phase of the vocal folds. A vibration with 100 cycles per second has a frequency of 100 Hz. This is a typical vocal Fundamental Frequency for adult males. A vibration with 200 cycles per second has a frequency of 200 Hz. This is a typical vocal for adult females.

3.4 What determines Fundamental Frequency?

F0 is determined by multiple interconnected factors. First, it increases with vocal folds length. Second, Fundamental Frequency inversely varies with mass/thickness of the vibrating part of the vocal folds. This implies that F0 is relatively low in thicker vocal folds, because they are supposed to vibrate slower. Thinner vocal folds vibrate faster and therefore produce higher sounding voices. Third, tension in the vibratory part of the vocal folds is important as well because vocal folds are elastic structures with different densities.

Chapter 4

Speaker Recognition

This chapter provides a brief overview of Automatic Speaker Recognition and it's role in two of it's main tasks, Speaker Identification and Speaker Verification. It also lists down the usage of this technology in many applications like forensic, authentication of smart devices etc. We will present some crucial signal processing concepts like feature extraction techniques and various toolkits available to perform Speaker Recognition experiments.

4.1 Automatic Speaker Recognition

Speaker Recognition [37] and *Speech Recognition* [2] are two distinct areas commonly used by researchers in speech processing applications.

Before understanding Speaker Recognition, it is important to understand the difference between Speaker Recognition and Speech Recognition. *Speech Recognition* is concerned with the words being spoken, while the *Speaker Recognition* aims to recognize the speaker rather than the words. Accent, language, speech, emotion, gender, and the speaker's identity are some of the informations contained in the human voice [35] as shown in Figure 4.1.

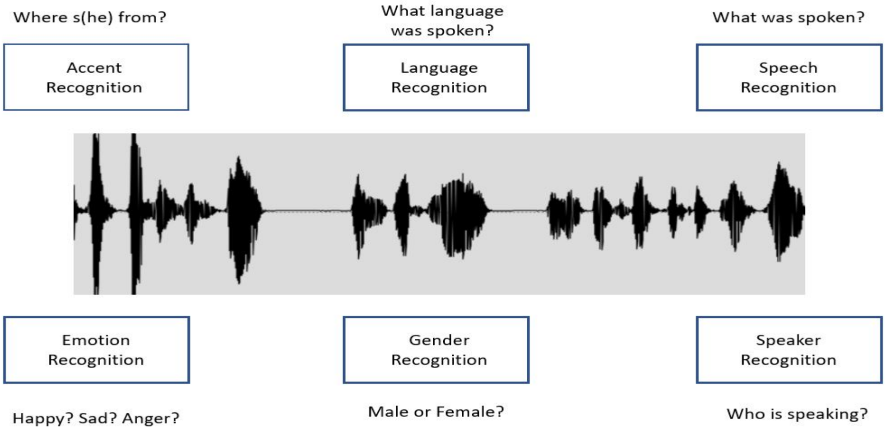

FIGURE 4.1: Some of the information contained in human voice [35]

Automatic Speaker Recognition is the task of recognizing the speaker of a given speech utterance. It can be classified into Automatic Speaker Verification (ASV), authenticating the identity claimed by the speaker, and Automatic Speaker Identification (ASI), determining the identity within a set of known speakers [40].

4.1.1 Speaker Identification

Speaker Identification is the process of determining the identity of an input voice which best matches a group of known voices. In simple words, in Speaker Identification system we speak to the system, and we ask the system to guess the person speaking. The system answers "Which speaker (out of a known group) is this?".

4.1.2 Speaker Verification

Speaker Verification means determining whether an unknown voice matches the speaker whose identity is being claimed. Here, the system takes the decision and addresses "Is the speaker who they claim to be?" .

The general process of Speaker Verification involves two tasks : enrollment and verification. Enrollment is the process of extracting the distinct characteristics of a speaker's voice and is used to create a claimed model to represent the enrolled speaker during verification. In verification, the distinct characteristics of a claimed identity's voice are compared with the previously enrolled claimed speaker model to determine if the claimed identity is correct.

4.1.3 Applications of Speaker Recognition System

Below is an outline of some broad areas where Speaker Recognition technology has been or is currently being used (this is not an exhaustive list) [42]

- Mobile Forensic Applications - Speech is being used in the courts to recognize suspects as guilty or non-guilty
- Call centers or banking sectors where it is required to verify the caller for specific operations
- Healthcare - for database verification for personal accesses
- Authentication of smart devices - Speaker Recognition system is used to verify a person's unique identity for secure access control over their devices
- Personalized user interfaces - Voice-mail has become popular due to the development in speech technologies. The system can recognize the speaker by adapting to his/her needs

4.2 Feature Extraction

At the highest level, each Speaker Recognition system contains a feature extraction module. Feature extraction is the process that extracts a small amount of data from the voice signal that can later be used to represent each speaker. It is a process to extract speaker's personal feature traits. The subsequent sections briefly describes certain features that have been used in the report.

There are a number of feature extraction techniques available. The features can be extracted either directly from the time domain signal or from a transformation domain depending upon the choice of the signal analysis approach [5]. Some of the audio features that have been successfully used for audio classification include

Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), Local Discriminant Bases (LDB), Linear Prediction Cepstral Co- efficients (LPCC) and Perceptual Linear prediction Cepstral Coefficients (PLPC). The most commonly chosen feature extraction technique for the state-of- the-art Speaker Verification system is MFCC [31, 50]. MFCC have shown to provide good performance while staying simple in computation and for it's robustness.

4.2.1 MFCC

MFCCs (Mel Frequency Cepstral Coefficents) are a feature widely used in Speech Recognition and Speaker Recognition.

From chapter 2, we got a brief idea of how speech is produced. Speech produced depends largely on the vocal organs. The shape of the vocal organs determines what kind of sound is made. The shape of the vocal tract is shown in the envelope of the short-term power spectrum of the speech. MFCCs accurately describe this envelope. The clear understanding of the theory behind the generation of MFCC is listed below.

Let's first understand Spectrum [47]. The speech is divided into many frames, each of which corresponds to a spectrum. This represents the energy at each frequency. The amplitudes are mapped to a grayscale representation. The larger the amplitude value, the darker the corresponding area. This way we get spectrogram which changes over time describing the speech signal. Figure 4.2 shows the spectrogram. We represent speech in spectrogram because sounds can be recognized by observing formants and their transitions.

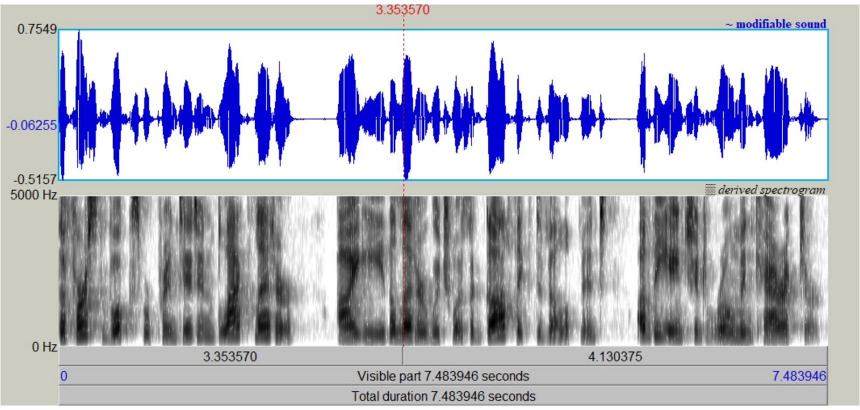

FIGURE 4.2: Spectrogram of a speech signal

4.2.1.1 Mel Frequency Analysis

Experiments on human auditory perception have shown that human auditory perception uses the entire auditory spectrum. However, human better distinguishes frequencies that are rather close together in the low frequencies, than in the high frequencies. The Mel Frequency Analysis is also based on human auditory perception experiments. Experimental observations have found that the filter bank like a human ear, focuses on the entire spectrum and is better in distinguishing frequencies that are rather close together in the low frequencies, than in the high frequencies.

However, these filters are not uniformly distributed on the frequency axis, there are many filters in the low-frequency region, they are more densely distributed, but in the high-frequency region, the number of filters becomes relatively small, and the distribution is very sparse.

4.2.1.2 Mel Frequency Cepstral Coefficients

We pass the spectrum through a set of Mel Filters to get the Mel Spectrum. At this point we perform a reciprocal analysis on Mel Spectrum. The reciprocal coefficient obtained on the Mel Spectrum is called the Mel Frequency Ceptrum Coefficient, referred to as MFCC.
Let's summarize the process of extracting MFCC features.

1. Pre-emphasis, framing and windowing of the speech signal - Pre-emphasis filter is applied to amplify the high frequencies. The signal is then split into short frames. Windowing step is meant to window each individual frame, in order to minimize the signal discontinuities at the beginning and the end of each frame

2. For each short-term analysis window, the corresponding spectrum is obtained through FFT (Fast Fourier Transform, which converts each frame from time domain into frequency domain)

3. The above spectrum is passed to Mel filter bank to obtain the Mel spectrum

4. Reverse spectral analysis is performed on the Mel spectrum (first take the logarithm of the energy in each Mel filter, then do the inverse transformation, the actual inverse transformation is generally achieved by the DCT discrete cosine transformation), and obtain Mel Frequency Cepstral Coefficient MFCC. This MFCC are characteristic of frame of speech

At this time, speech can be described by a series of MFCC vectors as shown in the Figure 4.3.

4.2.2 Fundamental Frequency as a feature

Many features of speech are used in Speaker Recognition systems. Fundamental frequency (F0) estimation , also referred to as pitch detection or pitch tracking, has been a popular research topic for many years, and is still being investigated today.
F0 is a very important parameter in determining the number of speech characteristics. F0 carries information about the speaker, such as gender, age, speech defect or emotional state. Two main sources of speaker characteristics are physical and learned [7]. As described in chapter 2 vocal folds is an important organ involved in the speech production and is part of the physical factor of speech. Some learned characteristics of speech production include speaking rate and dialect.

It is clear from results in several published studies that prosodic information can be used to effectively improve performance of and add robustness to Speaker Recognition systems [1, 25]. The motivation to use F0 as the feature was to asses how much speaker information is carried by this feature alone. F0 is an important feature in many research areas. For example in the voice privacy field, many speaker anonymization systems uses this kind of feature to generate anonymized

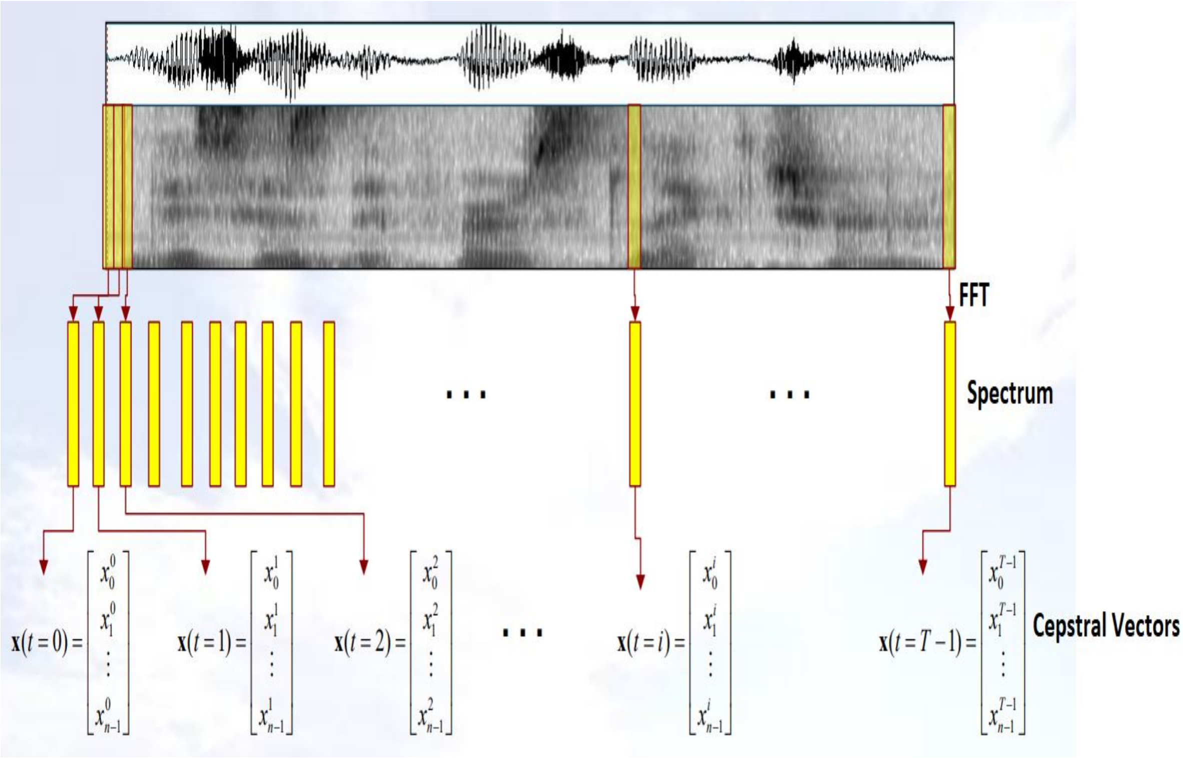
FFT
Spectrum
Cepstral Vectors
$\mathbf{x}(t=0)=[x_0^0, x_1^0, \dots, x_{n-1}^0]$
$\mathbf{x}(t=1)=[x_0^1, x_1^1, \dots, x_{n-1}^1]$
$\mathbf{x}(t=2)=[x_0^2, x_1^2, \dots, x_{n-1}^2]$
$\mathbf{x}(t=i)=[x_0^i, x_1^i, \dots, x_{n-1}^i]$
$\mathbf{x}(t=T-1)=[x_0^{T-1}, x_1^{T-1}, \dots, x_{n-1}^{T-1}]$

FIGURE 4.3: MFCC - Speech representation
Source - Presentation on Speech Recognition Speech signal descriptors by Krzysztof Ślot

voices, understanding it's implication in the modeling of the speaker identity is crucial for speaker anonymization researchers.

The traditional approach in estimating F0 is based on signal processing techniques. Time domain F0 estimation methods are based on the autocorrelation function YIN [9] or normalized cross-correlation function RAPT [48]. Frequency domain methods utilize the energy of the linear prediction residual harmonics SRH [11] or instantaneous frequency (e.g., TEMPO [23]). In addition to the methods that provide raw frame-level estimates of F0, multiple component methods have been developed for candidate F0 selection and/or postprocessing of the raw F0 estimates for improved robustness (e.g., pYIN [28], YAAPT [21], and Nebula [18]).

Recently in the literature, there exists a hybrid estimation approach which takes advantages of short-term temporal algorithms combined with short-term frequency-based algorithms such as the YAAPT algorithm which has been implemented lately. YAAPT is a widely used hybrid time and frequency domain method that can be run with varying levels of complexity. For our experiments we have used YAAPT. Figure 4.4 shows the signal in the amplitude time domain and Figure 4.5 shows it's corresponding F0 in the frequency time domain.

4.3 Toolkit for Speaker Recognition

Due to the advancement in technologies, various toolkits are readily available which has made our task a lot easier. These toolkits only requires speech datasets and few

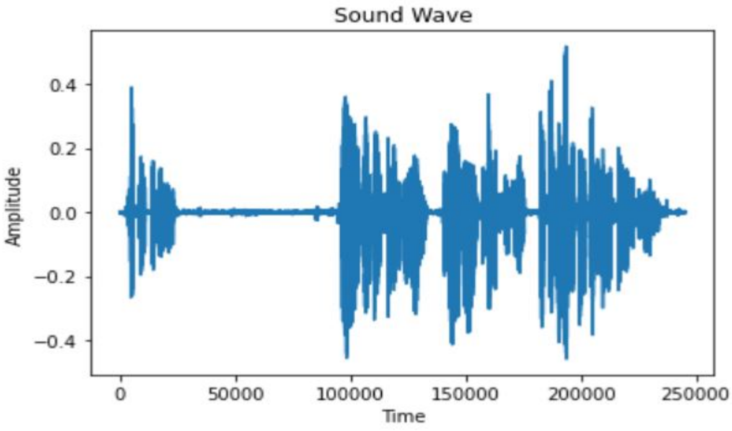

FIGURE 4.4: Signal in the amplitude time domain (sound wave - 14-208-0000.flac)

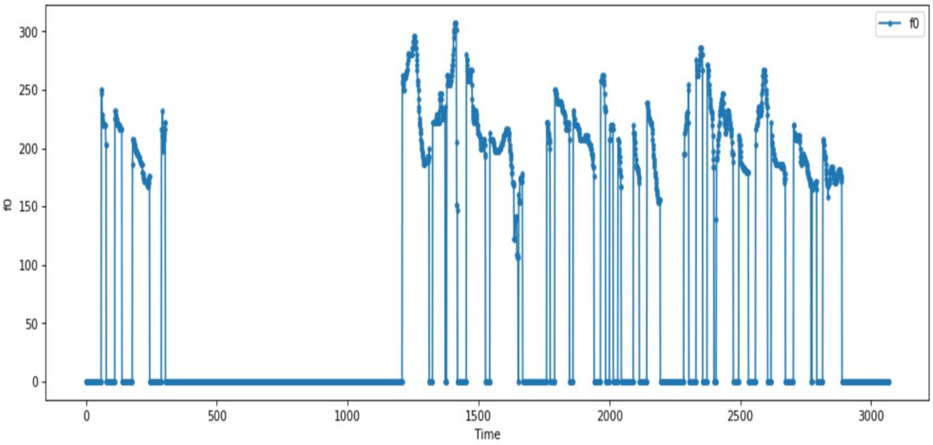

FIGURE 4.5: F0 extracted with YAAPT (sound wave - 14-208-0000.flac)

of the other parameters to perform experiments. We would be discussing about such toolkits which are available for Speaker Recognition.

SPEAR [24], an open source and extensible toolbox for Speaker Recognition which includes complete toolchain from the front-end feature extractor to the final steps of decision and evaluation. It is implemented in Python and C++. Modifying the C++ layer could be time consuming. ALIZE [6] is an open source software package for text independent Speaker Recognition. It is the closest competitor to Spear but does not contain an integrated feature module. SIDEKIT(Speaker IDEentification toolKIT) [26] is also an open source toolchain aims at providing an end-to-end tool chain with almost no dependencies on other tools. It is completely written in Python. Although, the above three toolkits have non restrictive licenses to be used in commercial applications, SIDEKIT is the most integrated solutions among these three toolkits. While the other two relies on some external packages or written in C++ which can make the system a bit complex. Kaldi [34] is a popular open source toolkit for speech recognition. It is is written in C++, thus would require knowledge in setting up the environment for executing the experiments. MSR [38] provides a high-level interface for Speaker Recognition experiments but it has a restrictive license.

4.3.1 Speaker Recognition in SIDEKIT

SIDEKIT is a toolkit for end to end Speaker Recognition system. It is completely written in Python. It provides a simple interface with a very minimum dependency on external tools, is easily understandable, requires minimum efforts to install, provides the compatibility with other tools. Most importantly, my supervisor's (Pierre and Hubert) worked on the toolkit. Having such experts in the team is an added advantage as it helped me right from the installation, adding pieces of code to executing my experiments. Because of these aspects, it was an easy to go toolkit.

Chapter 5

Implementation

This chapter describes the Speaker Verification model and it's related components in detail.

5.1 Speaker Recognition System

Figure 5.1 shows the block diagram of the verification system used in this thesis. Before performing speaker verification, one has to build a claimed speaker model through enrollment. During training phase, model learns how to identify the speaker of a training dataset.

Speaker verification consists of two phases enrollment and authentication.

- In enrollment phase, X-vector (explained in section 5.3) is extracted from one or more enrollment utterances spoken by the speaker whose identity is being claimed (Here, Bob identity is being claimed).

- In authentication phase, the X-vector extracted from the utterance of a claimed identity (called trial utterance) is compared with the X-vector of the speaker whose identity is being claimed (Bob) and a cosine similarity score is computed and then Automatic Speaker Verification system makes a decision by comparing the score with the threshold value. If the score is less than the threshold, Bob is not recognized as a claimed identity. If score is more than the threshold, Bob is recognized as a claimed identity.

5.2 Features

As described in chapter 4, we have used two features which serves as an input to TDNN (Time Delay Neural Network).

- MFCC - 80 features per frame is used as an input.

- F0 - We used single feature per frame as an input. This F0 value includes both voiced and unvoiced frames with voiced frames having some value and unvoiced frames has zero value. As shown in the Figure 4.5, the nonzero areas refers to voiced frames while the zero area refers to unvoiced or silence. It is represented in Hertz.

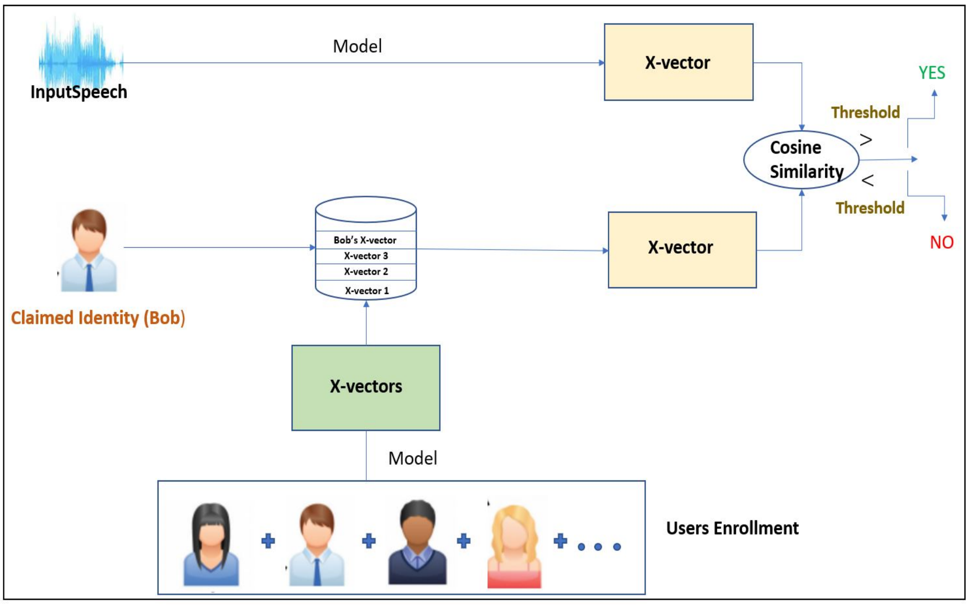

FIGURE 5.1: Speaker Verification System

5.3 Network Architecture

In Automatic Speaker Recognition, Time Delay Neural Network (TDNN) [33] has been proven to be an efficient network architecture because of it's strong ability in context modeling. Speaker's information is not stationary, it's information is contained in more than one frame. So, to capture speaker information, the model has to be able to model the dependency between the frames to deduce speaker's information. The TDNN is a nice architecture as it's property is to capture a large context. In addition, it is faster to train a TDNN because it is a feed-forward architecture, unlike recurrent neural network and also because of it's sub-sampling feature.

5.3.1 What is Speaker Embedding?

Speaker embeddings is a simple method of representing each speaker's identity in a compact way i.e it converts variable length audio into single dimensional vector which captures speakers discriminational characteristics. This is called as X-vector embeddings [45]. They are extracted using a neural network approach. In contrast to I-vector, which uses a statistical approach.

5.3.2 Model Architecture

Figure 5.2 shows the network architecture. The architecture consists of layers which captures the frame level representations of the input speech, statistical pooling layer that aggregates these representations together to generate a single vector representation of the speaker identity for the segment-level layers. Finally a softmax output layer. We will see the functionality of each of these layers below.

First two layers in the Figure 5.2 is TDNN layer which extracts frame level characteristics from utterances. The input represented here $x1, x2, x3, ...xT$ is the sequence of frame MFCC vectors, or frame F0 values. So, T is the speech length, expressed as a number of frames. The last layer output representations (sequence

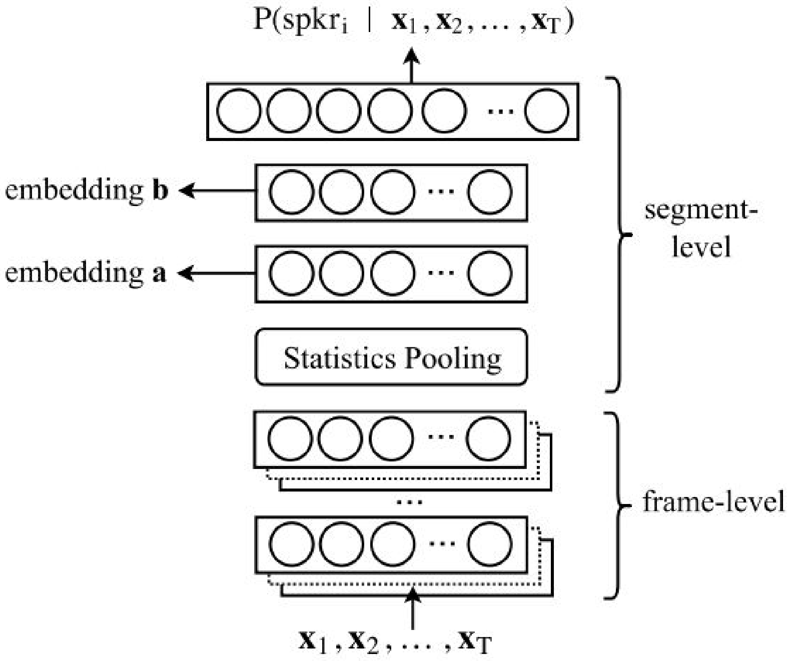
P(spkr$_i$ | $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$)
embedding **b**
embedding **a**
Statistics Pooling
segment-level
frame-level
$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$

FIGURE 5.2: Network Architecture

of frame level features) from TDNN is fed into the statistic pooling layer. This layer computes the mean and standard deviation of sequence of vectors of frames 1 to T. The statistical pooling layer generates an utterance level representation from many frames (mean operation). The additional layers are here to learn additional information. The last layer is the softmax layer which predicts the probability of the speaker for that particular utterance.

5.3.3 TDNN Basics

Time Delay Neural Network is based on context windowing and sub-sampling. It has connected layers of perceptrons and implemented as a feed-forward neural network.

In speech, as speech frames are correlated to each other, it is better to have a system that captures temporal context for prediction at a later stage. TDNN is based on context windowing which looks into the left and right features of the speech. Each layer cell of the TDNN takes outputs of the previous (lower) layer over a context window. The first layer processes input from narrow contexts of the speech signal. The deeper layers will process input by slicing the output of the hidden activations from the previous layer in order to learn wider temporal relationships. Context width increases as we go to upper layers. The notion t-13 and t+9 at the first layer in the Figure 5.3 means that the input has been spliced at the the current frame minus 13 and the current frame plus 9 [44]. Frames in red are evaluated.

It uses the property of sub-sampling [33] to improve efficiency. Sub-sampling is a method of allowing gaps between feature frames at each layer. It helps in decreasing the parameters and increasing computational efficiency.

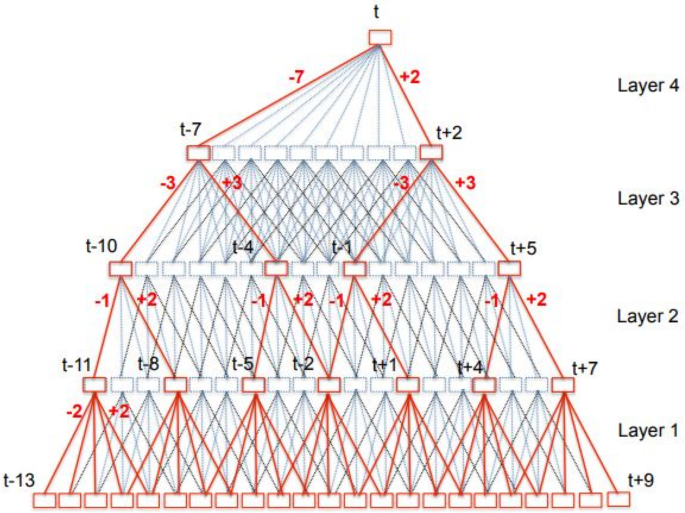

FIGURE 5.3: Computation in TDNN

5.3.4 Statistic Pooling

Neural network pooling layers are most commonly used in networks aggregates multiple output vectors into a single vector for example, in image classification there are many pooling techniques available like max pooling, average pooling etc. But, in speech processing, we use statistic pooling as we want to get the representations for the whole the utterances and these utterances may have thousands of frames and we want to get a single representation for all the utterances.

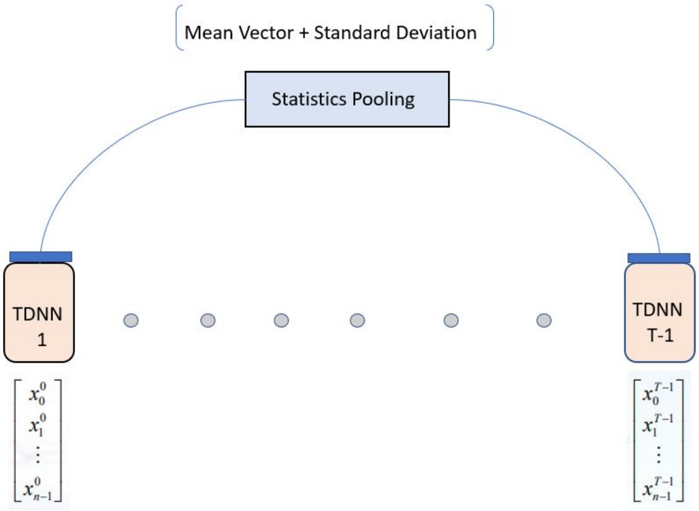
Mean Vector + Standard Deviation
Statistics Pooling
TDNN 1
TDNN T-1
$\begin{bmatrix} x_0^0 \\ x_1^0 \\ \vdots \\ x_{n-1}^0 \end{bmatrix}$
$\begin{bmatrix} x_0^{T-1} \\ x_1^{T-1} \\ \vdots \\ x_{n-1}^{T-1} \end{bmatrix}$

FIGURE 5.4: Statistic Pooling

TABLE 5.1: Network Configuration

Layer	Input X Output
TDNN Layer 1	1 X 512 (F0), 80 X 512 (MFCC)
TDNN Layer 2	512 X 512
TDNN Layer 3	512 X 512
TDNN Layer 4	512 X 512
TDNN Layer 5	512 X 1536
Statistic pooling	3072 X 256
Segment6 (Softmax Layer)	256 X 921 (921 = Number of Speakers)

In Speaker Verification where input utterances can have variable lengths, an average pooling layer has been introduced to aggregate frame-level speaker feature vectors to obtain an utterance-level feature vector, i.e., speaker embedding, with a fixed dimension. Statistic pooling was proposed as an extension to average pooling in the paper [44] where the author calculated not only the mean (aggregates the frame level features) but also the standard deviation (helps in capturing temporal variability over long contexts) of frame-level features. Figure 5.4 shows the Statistic pooling mechanism. Here, the inputs are MFCC vectors or F0 values. TDNN generates multiple vectors for each utterance as it operates on small segments. Statistic pooling generates a single vector.

5.3.5 Softmax/Cross-Entropy

The network is optimized using cross-entropy over all speakers in the training set. After training, the softmax layer is discarded, and the penultimate layer is used as the speaker representation.

5.3.6 Network Configuration

In our experiments, MFCC vectors and F0 values are used as features to encode speaker identity because they have been shown to be effective in speaker verification and recognition systems [43].

Five TDNN layers are used in sequence. These layers are followed by a ReLU and then BatchNorm layer. Sub-sampling is done at each layer. X-vectors are extracted from the segment6 layer just after the pooling layer. Cosine similarity is used as a scoring method during testing. Table 5.1 shows the configuration of the network.

5.3.7 Model Training

Using the embedding architecture described above, we conducted multiple experiments. Table below shows the parameters that were needed for model training. Each model was trained on 2 GPU's and 16 CPU's using Adam optimizer with a learning rate of 0.001. The learning rate value was modified during training to improve the model convergence using the CyclicLR scheduler policy.

TABLE 5.2: Hyperparameters

Parameters	Value
Batch size	200
Loss	aam
learning rate	0.001
Optimizer	adam
Scheduler	CyclicLR
Epoch	150

5.4 Speaker Verification test

Several methods have been proposed to measure the performance of the system using Equal Error Rate (EER), Detection Error Trade-off (DET) Curve, Receiver Operating Characteristic (ROC) curve as well as Detection Cost Function (DCF). We have used EER in our experiments.

It is possible to classify the errors according to the types of the system used [12] (identification, verification, open-set, closed-set). In Identification task, error types vary according to closed-set and open-set whereas in verification task, there are two types of errors, False rejection when system falsely rejects a true speaker and in False acceptance, when an imposter is accepted incorrectly. We will be discussing the error types of verification system in below section.

5.4.1 Cosine Similarity

X-vector is the embedding or vector that is extracted from the speech utterance during enrollment and verification. It is this embedding on which the cosine similarity [10] is evaluated over to predict a match or mismatch between the speaker in the enrollment and verification utterance.

Cosine similarity has the best ratio of simplicity and performance. In this approach, the cosine of the angle between the enrolled speaker embedding and that of the test utterance embedding is used as the decision score as shown in Figure 5.5. Scoring is performed directly using the enrollment and claimed X-vectors in the X-vector space.

5.4.2 Equal Error Rate

Classification of errors and deciding the threshold is crucial for the system performance. The main purpose of Speaker Verification systems is to verify whether a claimed identity is true or false. State-of-the-art Speaker Verification systems are trained to recognize the identity of speakers, and during Speaker Verification testing they make two kinds of error. There are two types of decision errors:

- FAR (False Acceptance Rate) - is related to accepting an impostor speaker (person who pretends to be someone else in case of fraudulent activities)
- FRR (False Rejection Rate) - is related to the incorrect rejection of a genuine speaker [29]

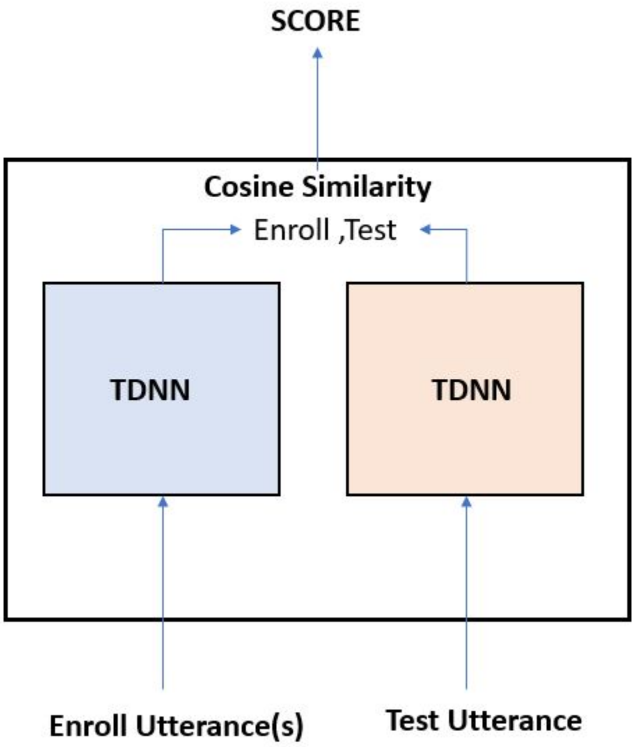

FIGURE 5.5: Cosine Similarity
Reproduced from [46]

EER stands for Equal Error Rate and used in determing the system performance. In this thesis, we calculated the EER and estimated the interval of confidence using the FEERCI toolkit [15]. As seen in the previous section, system provides a score of the test utterance against the speaker model of the claimed identity. This score is used to calculate EER.

The test data consists of both impostor and genuine speakers. Let's first take an example of imposters . If we consider left part in the Figure 5.6, we see that these are the scores obtained by different imposters and using Gaussian distribution, the scores are normalized by taking the mean and standard deviation as represented as "imposter mean" in the figure. This score is then compared to the threhhold . If the score of all impostors are falsely accepted, the value of FAR is one, and if none of the impostors are accepted, then FAR is zero. While the right part shows the values of the FAR for the score distribution as per the left image for varying threshold. An incorrect decision will occur in the system when it falsely accepts an utterance that is not from the claimed speaker. We can define FAR as per Equation 5.1.

$$\text{FAR} = \frac{\text{Number of falsely accepted utterances}}{\text{Total number of speaker trials of incorrect speakers}} \tag{5.1}$$

Same applies to the the genuine speaker or client as shown in figure Figure 5.7. Scores are normalized. This score is then compared to the threshold. It's value lies between zero and one. The image on the right shows the FAR for a varying threshold for the score distribution shown in the image on the left. An incorrect decision will occur if the system falsely rejects an utterance that is from the claimed speaker. We

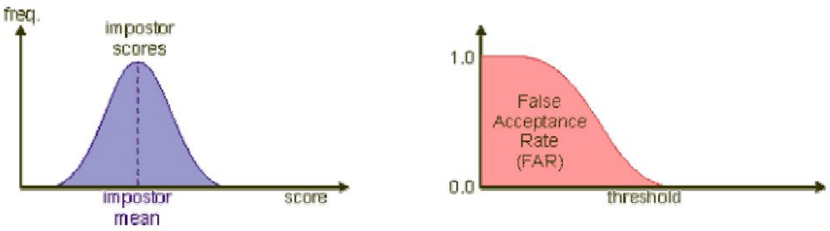

FIGURE 5.6: Imposter mean score distribution

can define FRR as per Equation 5.2.

$$\text{FRR} = \frac{\text{Number of falsely rejected utterances}}{\text{Total number of trials of correct speakers}} \tag{5.2}$$

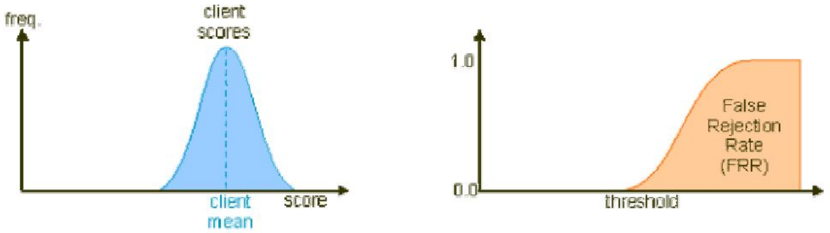

FIGURE 5.7: Client (genuine) mean score distribution

When the value of the FAR and the FRR are same. This point is called as Equal Error Rate (EER) as shown in the right part of the Figure 5.8.

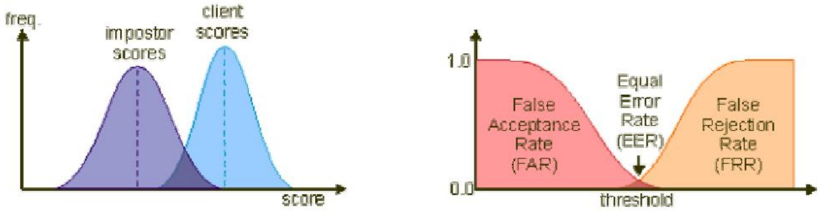

FIGURE 5.8: Equal Error Rate
Source - Taken from Technical Document About FAR, FRR and EER (SYRIS Technology Corp)

To compare our results together, we estimated the confidence interval for all of the EER results. The FEERCI package [15] performs this estimation using a Bootstrap Sampling method.

Chapter 6

Experimental Setup

This chapter provides details of experiments that were performed. It also lists out the dataset and environment under which these experiments were carried out.

6.1 Librispeech Dataset

We used LibriSpeech corpus [32] as the main training and testing datasets. The LibriSpeech corpus is a collection of approximately 1,000hr audiobooks in English, that are a part of the LibriVox project.

The training data is split into three partitions of 100hr, 360hr and 500hr sets while the dev and test data are split into ′clean′ and ′other′ categories, respectively. Therefore, it is divided into five sections namely, dev–clean, test–clean, dev–other, test–other, train–clean–100, train–clean–360, train–clean–500. Table 6.1 displays information about the LibriSpeech dataset.

We have used "train–clean–360" for training set and "test–clean" for testing set. Figure 6.1 shows the distribution of Female and Male data from the training dataset. We see that the Male and Female duration distributions are similar and centered around 15 seconds. It's more distributed between 3 to 15 seconds.

6.2 Experiment Settings

The experiments were conducted using SIDEKIT toolkit for Speaker Verification task and YAAPT for F0 extraction. We analysed the input signal to extract F0.

TABLE 6.1: Librispeech Dataset

subset	hours	per-spk minutes	female spkrs	male spkrs	total spkrs
dev-clean	5.4	8	20	20	40
test-clean	5.4	8	20	20	40
dev-other	5.3	10	16	17	33
test-other	5.1	10	17	16	33
train-clean-100	100.6	25	125	126	251
train-clean-360	363.6	25	439	482	921
train-other-500	496.7	30	564	602	1166

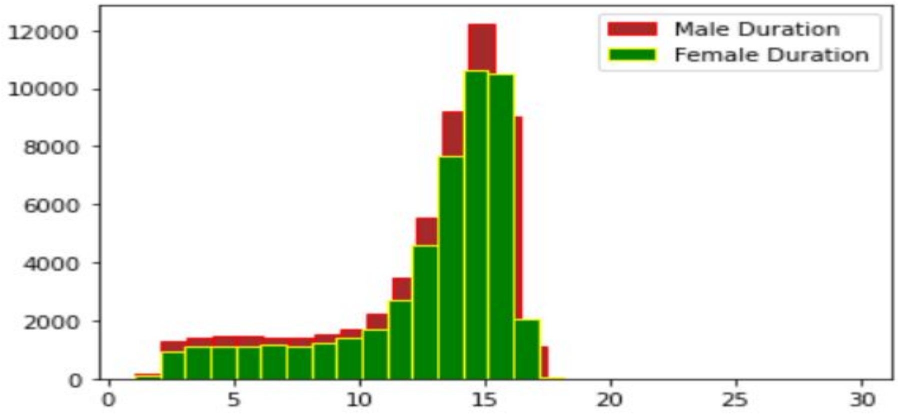

FIGURE 6.1: Distribution of Male and Female data with respect to speech duration (training dataset)

First, model was trained on "train_clean_360" data subset. This subset contains 921 speakers and is divided into train and validation such that 98% utterances of speaker are used for training while 2% are used for validation. The TDNN model is trained over the train split, while the validation set is used for monitoring the performance of the system. The Speaker Verification system is then used to verify unseen speaker from the test dataset from "test_clean" subset. According to the gender, Male, Female and Both were studied independently. The results of each of these experiments were measured by calculating EER and confidence interval using FEERCI.

6.3 Experimental Environment

The experiments were executed on *GRID5000*[1], which is a large-scale testbed for carrying out experiments when we want to use GPU to run machine-learning tasks or when our dataset size is too large to host it on personal machine or to run tasks that is time-consuming in a way that it occupies 90% of the CPU. GPU stands for "Graphical Processing Unit", designed to process parallel operations on multiple sets of data. Also, Inria had provided access to a shared filesystem for the duration of my internship period.

6.4 Evaluation Method

The dataset consists of 20 Male and 20 Female speakers selected from the test-clean database. All of the speech signals are sampled at 16000 Hz. Speaker Verification performance are reported using Equal Error Rate (EER) as described in section 5.4.

In this thesis, the different settings of F0 are studied experimentally by extracting fixed-length segments of durations 2 seconds, 3 seconds, 4 seconds, 5 seconds and by anonymizing F0 values which is then compared against the baseline feature MFCC and compared between different experiments as mentioned in section 6.5.

[1] https://www.grid5000.fr/w/Grid5000:Home

6.5 Results and Discussions

The main objective of this study is to determine the amount of speaker related information contained in the F0. To measure this information we utilized Speaker Verification. The following experimental techniques and conditions were studied:

1. **Experiment One - Speaker Verification for MFCC and F0 features**

   Similar to MFCC, F0 value was directly used as an input feature with a segment duration of 3 seconds. Table 6.2 provides a comparison between the results obtained using MFCC features and results obtained using Fundamental Frequency features for the same segment duration.
   As we can see from the table, EER of F0 **(20.0±1.2)** is higher for all (Male, Female, Both) than that of MFCC **(4.3±0.5)**. Performance in case of F0 is lower, compared to performance obtained with MFCC features. This is expected because MFCC has 80 features while F0 has 1 feature. The high EER also shows that F0 contains some speaker information. Female EER is higher than Male in both the cases.

TABLE 6.2: EER comparison between MFCC and Fundamental Frequency features

Feature	**Both** EER	**Male** EER	**Female** EER
MFCC	4.3±0.5	1.1±0.5	5.6 ±1.0
F0	20.0±1.2	17.8±1.7	21.8 ±1.7

2. **Experiment Two - Fundamental Frequencies of different segment durations**

   In this experiment, we studied the impact of segment duration when training the speaker recognition systems. In this context, we divided the training segments into segments of short durations (2, 3, 4, 5 seconds). This experiment was performed on training set and not during evaluation. Also, we had modified the training dataset in this experiment by removing all speech having less than 5 seconds duration from the training dataset in order make it uniform. This training dataset was then used for all four types of segment durations cases. Because of this, there is a little difference between the EER of F0 as shown in Table 6.2 and Table 6.3.

   The comparison has been studied based on the segment durations. The results are shown in Table 6.3. For 2 seconds segment duration, the performance is the worst with EER **24.3±1.0** as it doesn't carry much of the speaker's information due to the length of the frames while for 3 seconds, the performance has improved to some extent with EER of **20.8±1.2**. The performance improvement is impressive when we have 4 seconds duration for each training speech segment as EER has decreased to **18.5±1.0** thereby increasing the performance in comparison with 2 and 3 seconds while performance has decreased for 5 seconds **19.6±1.3** when compared with 4 seconds. This leads us to the conclusion that

F0 segment with 4 seconds performs better. At 4 seconds , F0 value is around 25% of EER which means that it carries some information. Also, Female EER is more than that of Male EER like on the original F0.

We can say that we should have short duration to have many segments, but not too short as it does not contain enough speaker information, not too big as it encapsulates other paralinguistic informations and reduces the number of total minibatches (with the increase in duration, number of segments for each speech decreases thereby decreasing the minibatches).

At the end of this experiment, our main focus was to analyze the duration having less EER but for our next experiments we would want to have an algorithm capable of modifying F0 value such that the EER is higher with least perturbating system and achieves privacy. As discussed in previous sections, the main aim of Deep Privacy project is to share an anonymized speech signals (do not carry speaker information). Current approaches are based on voice conversion systems (i.e., modifying a speech signal so that it sounds as if pronounced by another speaker). In our next experiments, we would be evaluating the performance by using different anonymization schemes.

TABLE 6.3: Fundamental Frequencies for different segment durations

Segment Durations (secs)	**Both** EER	**Male** EER	**Female** EER
2	24.3±1.0	23.6±1.7	25.0±1.5
3	20.8±1.2	17.7±1.9	23.0±1.6
4	18.5±1.0	16.5±1.8	19.5±1.5
5	19.6±1.3	18.5±1.8	20.6±1.6

3. **Experiment Three - Privacy preserving F0 transformation**

In this experiment, we have used noise and normalization schemes to modify F0 values.

- **Noise**
  We altered the original F0 values with noises (15 and 30 dB) as we wanted to find the right value to get anonymized speech while not being too destructive. White noise might remove some speaker information while still maintaining a usable F0 shape for other task such as voice conversion or speech to speech speaker anonymization. We evaluated the performance of the model under these noisy conditions as shown in Table 6.4. We see that there is some difference between the 15dB **(30.9±1.4)** and 30dB **(36.0±1.5)** noise. When compared with original F0 **(20.0±1.2)**, there is a huge difference in terms of performance (performance has decreased in anonymized speech). This also means that noise is able to affect the value of the frames as random noises are applied to each frames and is able to hide the speaker's information. Also, Female EER is more than that of Male EER like on the original F0.

TABLE 6.4: F0 with normalization and white noise

Type of Anonymization	**Both** EER	**Male** EER	**Female** EER
Original F0	20.0±1.2	17.8±1.7	21.8 ±1.7
Noise(15dB)	30.9±1.4	29.2±2.2	32.0 ±2.0
Noise(30dB)	36.0±1.5	32.6±2.2	37.9 ±2.1
Norm	19.9±1.1	18.3±1.8	21.1 ±1.5
Norm with Noise(15dB)	44.2±9.1	43.3±11.8	46.3 ±8.2
Norm with Noise(30dB)	46.0±7.3	44.3±10.6	47.5 ±6.8

- **Normalization**
  Also, we computed the mean and standard deviation of the Fundamental Frequency between enrollment and test utterances [8] for normalization. After estimating the F0 values for all training utterances, mean F0 and standard deviation was calculated from those F0 values. During verification, the F0 of the test utterances from either the genuine speaker or the impostor was normalized to target the mean and variance. F0 normalization is done based on the paper [16] which is computed using the Equation 6.1 as mentioned below:

  $$\hat{x}_t = \frac{1}{\sigma_x}(x_t - \mu_x) \tag{6.1}$$

  x_t = the source F0
  $\hat{x}_t$ = the normalized F0
  μ_x and σ_x: the mean and standard deviation statistics

  The experiment where F0 was only normalized "Norm", we see that result is not that impressive. We hypothesize that F0 normalization alone may not improve privacy as EER is almost same as original F0. It completely stays in the confidence interval of the original F0. The results are shown in Table 6.4. If we compare the results of noises (Noise (15dB) and Noise (30dB)) with normalization alone, we see huge differences between the two. Noise (15dB) and Noise (30dB) performed better in concealing the speaker's information. We further investigated by adding both the anonymization schemes (noise and normalization) together.

- **Normalization and noise together**
  We are able to see a huge difference between these experiments. "Norm with Noise (15dB)" is almost same as "Norm with Noise (30dB)" but performance has decreased when compared with the original F0. We hypothesize that "Norm with Noise (15dB) and "Norm with Noise (30dB)" is able to affect the local variance of the frame while normalization is affecting the global variance thereby removing most of the speaker's information, which in turn means better privacy. There is one drawback with "Norm with Noise (30dB)", it loses the utility for speech to speech conversion as it is too destructive.

When it comes to gender, we see that privacy is more in case of Female as on original F0. Please refer to the Appendix B for graphical representation of these experiments.

4. **Experiment Four - Voiced frames extracted from F0 value**

   In this experiment, the voiced information from the extracted F0 has been flattened (by assigning the values of voiced frames of speech to 1, while unvoiced is equal 0) to conceal the speakers identity. The result is shown in the Table 6.5. On it's own, voiced and unvoiced (which is a binary information, voiced =1 and unvoiced = 0) is not sufficient to capture speaker's information. EER rate is higher which is a good goal but it is not a viable speaker anonymization transformation as it is completely destructive. So, this experiment has not been presented as a privacy transformation but more as a specific evaluation. Also, EER has not hit the maximum of 50% which means that it does contain very little speaker information, especially for Female in this case. Female performs better unlike other experiments.

TABLE 6.5: EER for voiced segments of speech

Feature	**Both** EER	**Male** EER	**Female** EER
Voiced	44.8±3.8	44.2±2.5	41.5 ±2.4

6.6 Summary

We saw that, EER of F0 is high as compared to MFCC. One of the reasons being is the differences in the number of input features.

Moreover, we wanted to assess the segment duration at which F0 performs better so we extracted F0 values from segments with durations of 2 seconds, 3 seconds, 4 seconds and 5 seconds from the training set and compared the results. The results showed that the EER of 4 seconds was better than other cases.

Having found out which segment duration performed better, our next experiment was based on anonymization scheme (noise and normalization). In case of noise, we tampered F0 value at frame level thereby affecting the local variance while in normalization the whole utterance is tampered, which means that the global variance of the utterance on whole is getting changed. Performance for the experiments "Noise (15dB)" and "Noise (30dB)" has changed and we see huge difference when compared with original F0. In short, these anonymization scheme is able to hide speaker's information. F0 with noise 30dB performed better than 15dB in concealing the speaker information. The main motivation of this experiment was to assess which normalization scheme performs better by concealing speaker's information and in turn increase the privacy. Normalization alone didn't perform well. Our next experiments were based on adding both the anonymization scheme together. For experiments "Norm with Noise (15dB)" and "Norm with Noise (30dB), EER is high in these cases which means that the system performed well by concealing speaker's information thereby increasing the privacy. We can say based on the results that the use of F0 is most effective in terms of privacy when the F0 is represented with normalization and noise together.

The experiment where the voiced frames of F0 were flattened, showed that this anonymization transformation was completely destructive yet EER didn't reach maximum of 50%. This means that it contains little speaker information especially in case of Female.

When it comes to gender, Male EER is comparatively low than the Female EER and overall EER is dominated by Female.

Chapter 7

Conclusion and Future Work

In this thesis, we presented a study of F0 and it's usage as a feature in Speaker Verification tasks. We evaluated this F0 modification and the performance was measured in EER to measure privacy.

7.1 Conclusion

We observed that original F0 retains information about the speaker when compared to MFCC. The experiments shows that modifying F0 conceals speaker information and allows better privacy. The results also show that the performance is gender specific. The performance of Speaker Verification systems also depends on the segment duration. Keeping too low and too high can degrade the performance of the system. As per the results, we conclude that the F0 transformations can be used to anonymize the speech which can then be used for different algorithms for improving speech technologies.

7.1.1 F0 Extraction

We fixed our goal on the use of the parameters deduced from the the Fundamental Frequency. The in-depth study of all the techniques that we did during our research was useful to us in determining the most appropriate F0 extraction technique, which is the YAAPT algorithm. We have conducted experiments on one of the largest publicly available speech datasets (Librispeech) to measure the impact on Speaker Verification on different kinds of anonymiztion schemes on trained speaker embeddings.

7.1.2 F0 for Speaker Recognition

MFCC feature is known to have a high performance in speaker recognition. In our experiments, the equal error rate for MFCC is **4.3±0.5**. While the equal error rate for Fundamental Frequency feature for the same segment duration is **20.0±1.2**. F0's EER is higher than MFCC. As per the results obtained, we see that F0 contains speaker information.

7.1.3 F0 Anonymization

We anonymized F0 values by adding noise and normalizing it to verify if F0 can add useful information for the recognition by concealing speaker information and increasing the privacy of the system. We were able to conceal the speaker's information by modifying F0 values. This anonymized value can be useful in other speech processing tasks as the speaker's privacy is maintained.

7.2 Future Work

As per the results, F0 transformation is able to conceal speaker's information. So, it would be very interesting to perform further experiments to check the utility of speech by calculating word error rate (WER) on anonymized speech. If the results show low WER and high EER, it would indicate that the speech generated with this approach can be shared, stored and used in voice-based services, while protecting the speaker's identity.

Currently, we have only executed our experiments based on F0 value. But, in future we would like to add additional information like phonetic along with F0 value to test the privacy of the system.

7.3 Final Words

The work presented in this thesis shows that it is possible to preserve privacy of the speech through different transformations on the input feature. We managed to perform different experiments in a given timespan and could study the results. However, there are open ways in which we can further investigate the transformations to study the privacy preserving schemes on the speech data.

Appendix A

Terminologies

Please find the descriptions of the key terminologies used in the report:

- Training Data - The process of capturing voice for the purpose of extracting speaker′s representations and generating speaker model is known as training. The data collected from the process is the training data
- Test Data - The unseen utterance for verification task is called as Test data
- Feature Extraction - The process of extracting speakers representation from the voice signal
- Model Database - The repository consisting of speaker′s representation which are later used for verification task
- Enrollment - The process when a speaker enroll their voice in order to create it′s representaion in order to create a reference which can then be used during verification
- Threshold - The point that must be reached for a speech signal to be considered a match
- Feature Vectors - It is the speaker representations or characteristics which is in the form of vectors, that is collected from the speech samples

Appendix B

Experiments demonstrated by graph

The following section shows the graphical representation of the experiments that we carried out.

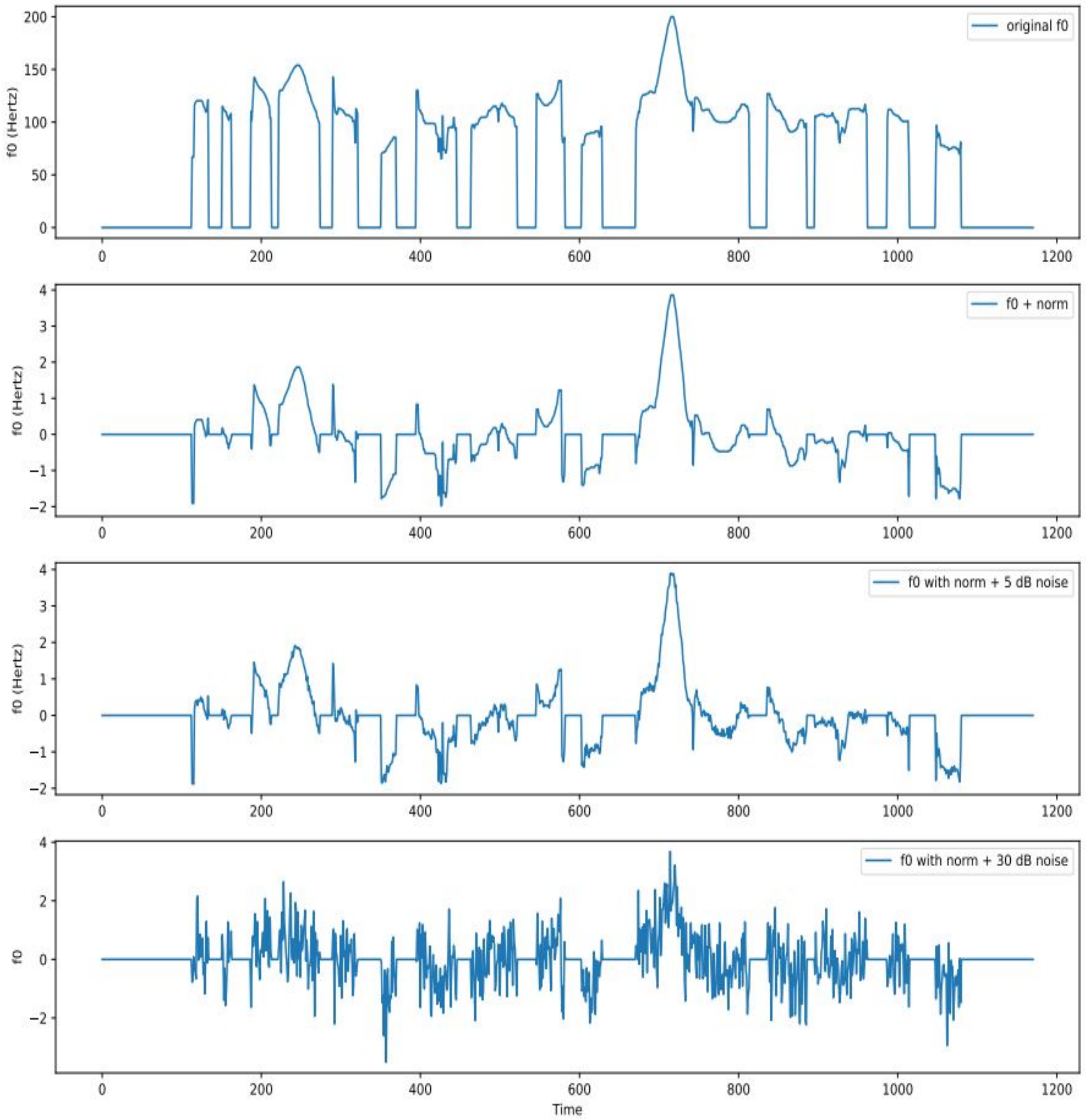

FIGURE B.1: F0 graphs for different experiments

We can see different variations in the graph. First graph shows the original F0. Second graph shows when normalization is applied to F0 and we can see that it is trying to bring the values to 0. Third graph shows F0 with normalization and noise of 5 dB and we can see little difference than the previous one. While we can see a huge difference here when F0 is normalized and noise of 30dB is added.

Bibliography

[1] A.G. Adami, R. Mihaescu, D.A. Reynolds, and J.J. Godfrey. "Modeling prosodic dynamics for speaker recognition". In: *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).* Vol. 4. 2003, pp. IV–788. DOI: 10.1109/ICASSP.2003.1202761.

[2] Sadeen Alharbi, Muna Alrazgan, Alanoud Alrashed, Turkiayh Alnomasi, Raghad Almojel, Rimah Alharbi, Saja Alharbi, Sahar Alturki, Fatimah Alshehri, and Maha Almojil. "Automatic Speech Recognition: Systematic Literature Review". In: *IEEE Access* 9 (2021), pp. 131858–131876. DOI: 10.1109/ACCESS.2021.3112535.

[3] G. Aneeja and B. Yegnanarayana. "Extraction of Fundamental Frequency From Degraded Speech Using Temporal Envelopes at High SNR Frequencies". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.4 (2017), pp. 829–838. DOI: 10.1109/TASLP.2017.2666425.

[4] J van den Berg. "Myoelastic-aerodynamic theory of voice production." In: *Journal of speech and hearing research* 1 3 (1958), pp. 227–44.

[5] Mouaz Bezoui, Abdelmajid Elmoutaouakkil, and Abderrahim Beni-hssane. "Feature extraction of some Quranic recitation using Mel-Frequency Cepstral Coeficients (MFCC)". In: *2016 5th International Conference on Multimedia Computing and Systems (ICMCS)*. 2016, pp. 127–131. DOI: 10.1109/ICMCS.2016.7905619.

[6] Jean-François Bonastre, Nicolas Scheffer, Matrouf Driss, Corinne Fredouille, Anthony Larcher, Alexandre Preti, Gilles Pouchoulin, Nicholas Evans, Benoit Fauve, and John Mason. "ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition". In: *Proc. Odyssey IEEE Workshop* 5 (Jan. 2008).

[7] J.P. Campbell. "Speaker recognition: a tutorial". In: *Proceedings of the IEEE* 85.9 (1997), pp. 1437–1462. DOI: 10.1109/5.628714.

[8] M.J. Carey, E.S. Parris, H. Lloyd-Thomas, and S. Bennett. "Robust prosodic features for speaker identification". In: *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*. Vol. 3. 1996, 1800–1803 vol.3. DOI: 10.1109/ICSLP.1996.607979.

[9] Alain de Cheveigné and Hideki Kawahara. "YIN, a fundamental frequency estimator for speech and music." In: *The Journal of the Acoustical Society of America* 111 4 (2002), pp. 1917–30.

[10] Najim Dehak, R. Dehak, Patrick Kenny, Niko Brummer, Pierre Ouellet, and Pierre Dumouchel. "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 1 (Jan. 2009), pp. 1559–1562.

[11] Thomas Drugman and Abeer Alwan. "Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics". In: Jan. 2011, pp. 1973–1976.

[12] Figen Ertas. "FUNDAMENTALS OF SPEAKER RECOGNITION". In: *Pamukkale University Journal of Engineering Sciences* 6 (2011).

[13] Fuming Fang, Xin Wang, Junichi Yamagishi, I. Echizen, Massimiliano Todisco, Nicholas Evans, and Jean-François Bonastre. "Speaker Anonymization Using X-vector and Neural Waveform Models". In: Sept. 2019, pp. 155–160. DOI: 10.21437/SSW.2019-28.

[14] Fuming Fang, Junichi Yamagishi, Isao Echizen, Md Sahidullah, and Tomi Kinnunen. "Transforming acoustic characteristics to deceive playback spoofing countermeasures of speaker verification systems". In: *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. 2018, pp. 1–9. DOI: 10.1109/WIFS.2018.8630764.

[15] Erwin Haasnoot, Ali Khodabakhsh, Chris G. Zeinstra, Luuk J. Spreeuwers, and Raymond N. J. Veldhuis. "FEERCI: A Package for Fast Non-Parametric Confidence Intervals for Equal Error Rates in Amortized O(m log n)". In: *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)* (2018), pp. 1–5.

[16] Zdeněk Hanzlíček and Jindřich Matoušek. "F0 transformation within the voice conversion framework". In: vol. 1. Aug. 2007, pp. 1961–1964. DOI: 10.21437/Interspeech.2007-549.

[17] Pearson Highered. *Production and Classification of Speech Sounds 3.1 Introduction*. https://documents.pub/document/production-and-classication-of-speech-sounds-pearson-and-classication-of.html?page=5.

[18] Kanru Hua. *Nebula: F0 Estimation and Voicing Detection by Modeling the Statistical Properties of Feature Extractors*. Oct. 2017.

[19] David G. Hanson. Jack Jiang Emily Lin. "VOCAL FOLD PHYSIOLOGY". In: 33.1 (2000). ISSN: 0030-6665. URL: https://doi.org/10.1016/S0030-6665(05)70238-3.

[20] Maëva Garnier Joe Wolfe and John Smith. *Voice Acoustics: an introduction*. http://newt.phys.unsw.edu.au/jw/voice.html. 2009.

[21] Kavita Kasi and Stephen A. Zahorian. "Yet Another Algorithm for Pitch Tracking". In: *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing* 1 (2002), pp. I–361–I–364.

[22] Bartkova Katarina, David Le Gac, Delphine Charlet, and Denis Jouvet. "Prosodic parameter for speaker identification". In: Sept. 2002. DOI: 10.21437/ICSLP.2002-325.

[23] Hideki Kawahara, Alain de Cheveigné, and Roy D. Patterson. "An instantaneous-frequency-based pitch extraction method for high-quality speech transformation: revised TEMPO in the STRAIGHT-suite". In: *5th International Conference on Spoken Language Processing (ICSLP 1998)* (1998).

[24] Elie Khoury, Laurent El Shafey, and Sébastien Marcel. "Spear: An open source toolbox for speaker recognition based on Bob". In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2014, pp. 1655–1659. DOI: 10.1109/ICASSP.2014.6853879.

[25] Pavel Labutin, Sergey Koval, and Andrey Raev. *Speaker identification based on the statistical analysis of f0*. Jan. 2007.

[26] Anthony Larcher, Kong Aik Lee, and Sylvain Meignier. "An extensible speaker identification sidekit in Python". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2016, pp. 5095–5099. DOI: 10.1109/ICASSP.2016.7472648.

[27] Leena Mary and B. Yegnanarayana. "Extraction and representation of prosodic features for language and speaker recognition". In: *Speech Communication* 50 (Oct. 2008), pp. 782–796. DOI: 10.1016/j.specom.2008.04.010.

[28] Matthias Mauch and Simon Dixon. "PYIN: A fundamental frequency estimator using probabilistic threshold distributions". In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2014, pp. 659–663. DOI: 10.1109/ICASSP.2014.6853678.

[29] Victoria Mingote, Antonio Miguel, Dayana Ribas, Alfonso Ortega, and Eduardo Lleida. "Optimization of False Acceptance/Rejection Rates and Decision Threshold for End-to-End Text-Dependent Speaker Verification Systems". In: Sept. 2019, pp. 2903–2907. DOI: 10.21437/Interspeech.2019-2550.

[30] Ansar MK. *Speech Production mechanism*. https://www.academia.edu/3052872/Speech_Production_mechanism.

[31] Tumisho Billson Mokgonyane, Tshephisho Joseph Sefara, Thipe Isaiah Modipa, Mercy Mosibudi Mogale, Madimetja Jonas Manamela, and Phuti John Manamela. "Automatic Speaker Recognition System based on Machine Learning Algorithms". In: *2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/ Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)*. 2019, pp. 141–146. DOI: 10.1109/RoboMech.2019.8704837.

[32] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. "Librispeech: An ASR corpus based on public domain audio books". In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

[33] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. "A time delay neural network architecture for efficient modeling of long temporal contexts". In: *INTERSPEECH*. 2015.

[34] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Vesel. "The Kaldi speech recognition toolkit". In: *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* (Jan. 2011).

[35] Shamsul Mohamad Rafizah Mohd Hanifa Khalid Isa. "A review on speaker recognition: Technology and challenges". In: *Computers Electrical Engineering* 90 (2021).

[36] Ravi Ramachandran, Kevin Farrell, Roopashri Ramachandran, and Richard Mammone. "Speaker recognition - General classifier approaches and data fusion methods". In: *Pattern Recognition* 35 (Dec. 2002), pp. 2801–2821. DOI: 10.1016/S0031-3203(01)00235-7.

[37] Douglas A. Reynolds. "An overview of automatic speaker recognition technology". In: *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 4. 2002, pp. IV–4072–IV–4075. DOI: 10.1109/ICASSP.2002.5745552.

[38] Seyed Omid Sadjadi, Malcolm Slaney, and Larry Heck. *MSR Identity Toolbox v1.0: A MATLAB Toolbox for Speaker Recognition Research*. Tech. rep. MSR-TR-2013-133. 2013. URL: https://www.microsoft.com/en-us/research/publication/msr-identity-toolbox-v1-0-a-matlab-toolbox-for-speaker-recognition-research-2/.

[39] Zia Saquib, Nirmala Salam, Rekha Nair, Nipun Pandey, and Akanksha Joshi. "A Survey on Automatic Speaker Recognition Systems". In: *Communications in Computer and Information Science* 123 (Jan. 2010), pp. 134–145. DOI: 10.1007/978-3-642-17641-8_18.

[40] Ali Shahin Shamsabadi, Brij Mohan Lal Srivastava, Aurélien Bellet, Nathalie Vauquier, Emmanuel Vincent, Mohamed Maouche, Marc Tommasi, and Nicolas Papernot. *Differentially Private Speaker Anonymization*. 2022. DOI: 10.48550/ARXIV.2202.11823. URL: https://arxiv.org/abs/2202.11823.

[41] Rohini Shinde. *Study of Feature Extraction and Pattern Comparison Techniques for Speech Recognition*. Nov. 2012.

[42] Nilu Singh, Prof. Raees Khan, and Raj Shree Pandey. "Applications of Speaker Recognition". In: *Procedia Engineering* 38 (Dec. 2012), pp. 3122–3126. DOI: 10.1016/j.proeng.2012.06.363.

[43] David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. "Spoken Language Recognition using X-vectors". In: June 2018, pp. 105–111. DOI: 10.21437/Odyssey.2018-15.

[44] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur. "Deep Neural Network Embeddings for Text-Independent Speaker Verification". In: *INTERSPEECH*. 2017.

[45] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. "X-Vectors: Robust DNN Embeddings for Speaker Recognition". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, pp. 5329–5333. DOI: 10.1109/ICASSP.2018.8461375.

[46] David Snyder, Pegah Ghahremani, Daniel Povey, Daniel Garcia-Romero, Yishay Carmiel, and Sanjeev Khudanpur. "Deep neural network-based speaker embeddings for end-to-end speaker verification". In: *2016 IEEE Spoken Language Technology Workshop (SLT)*. 2016, pp. 165–170. DOI: 10.1109/SLT.2016.7846260.

[47] *Speech Signal Processing (4) Mel Frequency Reciprocal Coefficient (MFCC)*. https://blog.csdn.net/zouxy09/article/details/9156785.

[48] David Talkin. *A Robust Algorithm for Pitch Tracking (RAPT)*. 2005.

[49] "The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection". In: ISCA, 2017, pp. 2–6. DOI: 10.21437/INTERSPEECH.2017-1111.

[50] Vibha Tiwari Tiwari. "MFCC and its applications in speaker recognition". In: *Int. J. Emerg. Technol.* 1 (Jan. 2010).

[51] Ville Vestman, Tomi Kinnunen, Rosa González Hautamäki, and Md Sahidullah. "Voice Mimicry Attacks Assisted by Automatic Speaker Verification". In: *Computer Speech Language* 59 (2020), pp. 36–54. ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2019.05.005. URL: https://www.sciencedirect.com/science/article/pii/S0885230818303863.

[52] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md Sahidullah, and Aleksandr Sizov. "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge". In: *Sixteenth annual conference of the international speech communication association*. 2015.