Machine Learning Based Listener Classification and Authentication Using Frequency Following Responses to English Vowels for Biometric Applications

by

Bijan Borzou

Ph.D., University of Ottawa, 2016

M.Sc., Amirkabir University of Technology, 2009

B.A.Sc., Amirkabir University of Technology, 2007

Thesis Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Applied Sciences in Electrical and Computer Engineering

in the

School of Electrical Engineering and Computer Science

Faculty of Engineering

University of Ottawa

Abstract

Auditory Evoked Potentials (AEPs) have recently gained attention as a biometric feature that may improve security and address reliability shortfalls of other commonly-used biometric features.

The objective of this thesis is to investigate the accuracy with which subjects can be automatically identified or authenticated with machine learning (ML) techniques using a type of AEP known as the speech-evoked frequency following response (FFR).

Accordingly, the results show more accurate discrimination between FFRs from different subjects than what has been reported in past studies. The accuracy improvement is searched either by optimized hyperparameter tuning of the ML model or extracting new features from FFRs and feeding them as inputs to the model. Finally, the accuracy of authenticating subjects using FFRs is investigated using a “sheep vs. wolves” scenario.

The results of this work shed more light on the potential of use of speech-evoked FFRs in biometric identification and authentication systems.

Acknowledgements

Deciding to pursue a second M.Sc. program after completing my PhD was a turning point in my life and professional career, which I will always remember. Moving forward on this path was impossible without the support and help of my supervisors. Therefore, I would like to express my sincere gratitude to Dr. Martin Bouchard and Dr. Hilmi Dajani for giving me this opportunity and guiding my research. Their fantastic guidance and cooperation throughout my M.Sc. studies were invaluable.

Once again, similar to my previous Ph.D., M.Sc., and B.A.Sc. theses, I would like to express my heartfelt thanks to my best friend forever, Mohammad Alavirad, whom I cannot call anything but my brother. His help, presence, and dedication during the last ten years of my life cannot be expressed in words. Without him, it would have been impossible to overcome the challenges and complete this long journey.

My parents have played an essential role in helping me achieve our dreams. I am blessed to have parents who have supported and encouraged me during the lowest of lows and the highest of highs, fueling my every moment.

Table of Contents

List of Figures

List of Tables

List of Acronyms

ABR	Auditory Brainstem Response
AEP	Auditory Evoked Potential
ASSR	Auditory Steady State Response
CABR	Complex Auditory Brainstem Response
CAP	Compound Action Potential
DT	Decision Tree
eFFR	envelope Frequency Following Response
ENT	Ear Nose Throat
ERB	Equivalent Rectangular Bandwidth
FFR	Frequency Following Response
KNN	K-Nearest Neighbors
LDA	Linear Discriminant Analysis
LLAEP	Long Latency Auditory Evoked Potential
LR	Logistic Regression
MLR	Middle Latency Response
RBF	Radial Basis Function
RF	Random Forest
sFFR	spectral Frequency Following Response
STFT	Short-Time Fourier Transform
SVM	Support Vector Machine

Chapter 1 Introduction

1.1 Overview

The interest in using biometric technologies in identification and authentication applications in various industries and sectors has substantially increased in recent years. From securing access to buildings, devices, and financial transactions to enhancing the accuracy and efficiency of personal identification, biometrics is changing the way we interact with technology. With the development of advanced technologies such as facial recognition, fingerprint analysis, and iris scanning, biometrics is playing a significant role in increasing security and convenience in various aspects of our lives. Accordingly, it is worthwhile to shed light on the accuracy and possibility of using appropriate forms of biometrics and their accuracy in identification and authentication applications.

1.2 Biometrics

1.2.1 Definition

Biometrics are automated methods of recognizing a person based on a physiological or behavioral characteristic [1]. Biometric technologies have become the foundation of many high security identification solutions. Some of these physiological characteristics include but are not limited to facial characteristics, iris recognition and finger or hand images. However, behavioral characteristics are distinguishing quality characteristics that are either learned or acquired. Dynamic signature verification and speaker verification are examples of behavioral characteristics.

Biometric verification is comparing a registered biometric sample against a newly captured sample i.e., a fingerprint or an image of the user's face captured by their smart phone camera during a log in. In the identification mode, the biometric system identifies a person from the entire enrolled population by searching through a database for a match solely based on the biometrics. For instance, an entire database is searched to verify whether a person has applied for entitlement benefits under two different names or not. This is referred to as one-to-many matching [2].

However, in the authentication or verification mode, the biometric system verifies a person's claimed identity from their previously enrolled pattern. This is also referred to as one-to-one matching. Examples of this could be found in most computer access or network access environments. Users enter their username, but instead of entering a password, a simple touch with a finger or a glance at a camera authenticates the user [2].

1.2.2 Biometric Identification and Authentication Techniques

As mentioned above, the most common types of the biometrics currently in use are:

- Fingerprints
- Face recognition
- Iris recognition
- Hand and finger geometry
- Signature verification
- Speaker recognition.

Fingerprints are patterns of local furrows and ridges on the fingertip surface that are formed and determined in the fetal period [3]. Their validity as appropriate biometrics for identification has been established by Jain et al. [4].

Facial images are the most frequently used biometric characteristic. Subject identification approaches using face images are usually based on features such as gender, skin color and shape of attributes like eyes, nose, mouth, chin, and forehead. Also, the location and spatial relationship of these facial attributes are important factors in subject identification [5].

Face recognition systems have been reported to obtain very promising results at least in recent years. Accuracies as high as 97% using deep neural network models have been reported in the literature [6].

Iris is the colored area of the eye that surrounds the pupil. Iris patterns are also thought to be unique and therefore are used as a biometric feature. These patterns are recorded through image acquisition systems. Iris recognition systems have been shown to work well even in the presence of contact lenses or eyeglasses in both verification and identification modes for individuals from different ethnic groups and nationalities [2].

Hand recognition is another personal verification method which is well established and has been available for the past couple of decades. In this method, measures of physical characteristics of either the fingers or the hands are used. The physical characteristic measures include length, width, thickness, and the surface area [1]. Hand recognition has been used in a variety of applications such as commercial and residential applications, time, and attendance systems, and also in general personal authentication applications [2].

Signature verification uses the dynamic analysis of a signature to verify a subject. The verification is based on measuring the pressure, speed and the angle used by the subject while signing [1]. Signature verification has been used in many applications such as e-business applications and any other application where signature is the accepted method of personal verification [2].

In speaker recognition, the acoustic features of speech that have been found to differ between individuals are used. These acoustic patterns are influenced by anatomy (shape and size of the mouth and throat) and behavioral patterns such as speaking style and voice pitch [1]. Various models have been mentioned in the literature to perform speaker recognition such as hidden Markov models, neural networks, pattern-matching algorithms, decision trees etc. [1].

Identification and authentication of subjects based on the above-mentioned biometric methods have been tried for long time in different applications and good results and accuracies have also been reported in the state of the art. For instance, the study by Chen et al. [7] explored the performance of fingerprint recognition algorithms on a real-world database containing 10,000 fingerprint images, achieving an impressive identification accuracy of 99.2%.

Also, Taigman et al. [6] conducted a comprehensive evaluation of facial recognition algorithms and achieved an overall accuracy of 99.63% on the LFW (Labeled Faces in the Wild) dataset using the DeepFace architecture. Also, in the field of biometric identification using signatures, some impressive results have been reported in the literature. For instance, the study conducted by Leghari et al. [8], proposed a multi-feature fusion approach for signature verification [2]. The authors combined multiple low-level and high-level features, such as shape, texture, and motion, extracted from the signature images to create a robust feature representation. By employing a support vector machine classifier, their approach achieved impressive accuracies of 99.1% and

90.8%, respectively. Finally, for biometric identification using iris recognition, accuracies as high as 99.5% are reported in the literature [9].

1.2.3 Critiques

Although all the above-mentioned biometric techniques have been very frequently used in different applications as explained, there are some shortfalls associated with each of them that makes them not very secure as well. For example, fingerprint identification, as a form of biometric authentication, has been subject to various challenges and problems. The accuracy of fingerprint recognition systems can be negatively impacted by factors such as aging, injury, or the presence of dirt or grease on the finger [10]. Also, it is possible to obtain and replicate a person's fingerprints using materials such as play-doh, resin, or even high-resolution images [11]. Dong [12] found that digital replicas of fingerprints can be created using commercially available fingerprint sensors and 3D printing technology. These findings suggest that the authenticity and uniqueness of fingerprints, as a form of biometric authentication, can be compromised.

Facial recognition technology, despite its widespread use as a form of biometric authentication, has also been shown to be vulnerable to being fooled by high-resolution photographs. According to a study performed by Shan [13], facial recognition systems can be deceived by highly detailed, high-resolution images, which can be used to impersonate a target individual. Athalye et al. [14] also found that facial recognition systems can be easily misled by "adversarial examples," or carefully crafted images designed to trick the technology into recognizing a false identity.

Also, in the last several years as biometric scanners have become cheaper and more prevalent, the challenge to biometric privacy has become critical [15], [16]. Advocates have been sounding the alarm about biometric privacy and large companies such as Microsoft was calling for greater regulation of facial recognition technology by 2018 [17], [16]. Along with such an increased concern, a wave of litigation has been generated against companies that use facial recognition to identify people in photographs and employers that use fingerprint biometric scanners for employee timekeeping [16] [18]. For instance, Facebook has faced more than $30 billion in liability for biometric privacy laws violation [16], [19].

Even with a more private biometric feature such as iris, the security of databases has been a concern, as there is a risk of the sensitive biometric data being stolen by hackers. According to Johnson [20], iris recognition databases are vulnerable to cyber attacks, and once breached, the sensitive information can be easily misused or sold on the black market. Wu [21] also found that iris recognition systems can be hacked using high-resolution images or videos obtained through surreptitious means, such as a hidden camera.

Speaker recognition has also been shown to be susceptible to failure due to the possibility of creating a mathematical model of a person's speech production system and mimicking their voice. Smith [22] reported that speaker recognition systems can be easily fooled by speech synthesized from a mathematical model of a target individual's speech production system. Nguyen [23] also found that speaker recognition systems can be circumvented by using audio recordings or digital samples of a target individual's voice, which can be used to create a synthetic version of their speech.

These findings highlight the limitations of the above-mentioned biometric technologies and the need for ongoing research and improvement to ensure their reliability and security.

1.2.4 Listener Recognition

Based on the above-mentioned critiques, there is a high demand for biometric methods with both higher security and reliability level. Accordingly, listener recognition could be investigated as a biometric-based technique that can address those shortfalls.

In listener recognition, Auditory Evoked Potentials (AEPs) are used as biometric features. The AEPs are brain neural responses to sound stimuli that are time-locked to some event or to the omission of a stimulus [24]. The stimuli are audio signals including but not limited to speech sounds, clicks, tones and noise [24] and the response is a small electrical potential that can be recorded from the scalp non-invasively. Characteristics of the AEP signals such as amplitude and energy have been recently used along with various models like deep neural networks, decision trees, etc. for listener identification applications [24], [25], [26].

One of the AEP types that can address the security and reliability issues with previous biometric techniques is called the Frequency Following Response (FFR). The FFR is a short latency, scalp-recorded neuro response reflecting phase-locked activity from the human auditory system [27], [28].

Since FFRs are recorded with electrodes, the signals cannot be easily captured. This could address many of the security and reliability difficulties that methods such as fingerprints, facial and speaker recognition were suffering from. In addition to this, there is a possibility of refreshing the stored FFRs database by changing the stimuli on a periodic basis which increases the dataset security compared to fixed traits like facial features or fingerprints. Accordingly, even if the system is hacked and the stored responses are copied, the database can be refreshed, and the hacker cannot produce FFRs in response to new speech stimuli.

As to date, many researchers have worked on characterizing the frequency following response [27]- [29]. Also, the aforementioned advantages of FFRs have raised increasing interest in the use of speech FRR for identification applications. In a recent study, Sun et al. [25], [26] showed that their extracted features from the speech evoked FFR were stable and representative, and thus effective for constructing a solid biometric identification system in laboratory environment. However, investigations on the accuracy of a listener recognition system based on FFR features is still an open question and is where the state of the art could be improved.

1.2.5 FFR Classification and Identification with Machine Learning

Machine learning models have recently been used to perform FFRs classification due to the increasing interest in application of speech FFRs for identification. Within this context, Heffernan [24] characterized the FFR in normal hearing adults subjected to 4 short synthetic vowels at different sound presentation levels, in terms of the spectral content and frequency peaks of the FFR. He then used a machine learning approach to classify the responses to these vowels. Also, he studied the effect of gender on the amplitude spectra and major waveform peaks of the responses. In this study, various machine learning models such as Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbors Algorithm (KNN), Decision Tree (DT) and Random Forest (RF) were used, and the best performance reported was 71.7% for vowel classification using the Random Forest Model.

Also, Sun [25], [26] used a variety of classification algorithms for listener identification on the same FFR database used by Heffernan [24] based on spectral and spectrogram features. The evaluated machine learning algorithms included an SVM with linear kernel, radial basis function kernel and polynomial function kernel, XGBoost, and a convolutional neural network. The best accuracy reported was 84.09% obtained by an SVM model with radial basis function kernel. To check the validity of his results, Sun also investigated the stability of the responses between Test and Retest for each subject.

1.3 Thesis Goal

Given the above attempts in using machine learning in subject, vowel, and sound level classification in listener identification application, there is still a need to both increase their accuracy and add the subject authentication aspect to such work. Accordingly, by adopting the same database as Sun [26] and Heffernan [24], this study focuses on investigating additional extracted features than those used by Sun and Heffernan with the goal of achieving better accuracies. An SVM classifier is the adopted machine learning model since it had the best classification accuracies reported by Sun [26]. In addition, the possibility of performing listener authentication with machine learning models and using the adopted database is explored.

1.4 Contributions

The following contributions resulted from this work:

1- Classification of different subjects based on their FFRs was performed with better accuracies compared to what has been reported in past similar studies. Here, a very recent study in the field performed by Sun [25], [26] that employed a variety of classification algorithms for listener identification on an FFR database recorded by Heffernan [24] was adopted as a reference to compare the results with. The best accuracy reported for the classification task was 84.09% obtained by an SVM model with RBF kernel.

   By adopting the same dataset, analogous scenarios and the same input features as in [25], [26], performing a systematic hyperparameter tuning for the SVM model and using higher spectrogram resolutions as input features, better accuracies were obtained in this thesis compared to [25], [26]. For instance, the best accuracy of classification obtained in this study was 93.18% which is a 9.09% improvement compared to [25], [26].

2- The accuracy of authenticating subjects using FFRs was investigated using a "sheep vs. wolves" scenario. The sheep vs. wolves scenario was adopted to classify the listeners into two groups of trusted and untrusted subjects. In this case the best average accuracy of 86.36% was obtained. However, due to the limitations of the adopted dataset size, implementation of other authentication scenarios such as one vs. all was not possible, but still the accuracies obtained for sheep vs. wolves experiments were promising. The results of this work shed more light on the potential of use of speech-evoked FFRs in biometric identification and authentication systems.

1.5 Thesis Structure

This thesis consists of five chapters that are arranged as follows:

1) Chapter 1 reviews the background of the biometric identification and authentication techniques, motivations, aims and provides a brief introduction of the structure of the thesis.

2) Chapter 2 introduces the Frequency Following Response (FFR) evoked by speech, the dataset used in the thesis and the experimental setup used for recording the data.

3) Chapter 3 reports on the extracted features from the FFRs such as spectrograms and gammatonegrams.

4) Chapter 4 reports on the results of the machine learning model used to perform subject classification on the extracted features described in Chapter 3.

5) Chapter 5 reports on the results obtained for a case study of authenticating a group of listeners against the other group defined as the sheep vs. wolves case.

6) Finally, Chapter 6 includes a discussion on the major findings of the thesis in comparison to previous studies, some limitations of the work and future recommendations.

Chapter 2 The Frequency Following Response

2.1 Overview

This chapter introduces the Frequency Following Response (FFR) evoked by speech, the dataset used in this thesis and the experimental setup used for recording the data.

2.2 Auditory Evoked Potentials

Auditory Evoked Potentials (AEPs) are a subclass of event-related potentials - neural responses that are time-locked to some event (e.g., a sensory stimulus or a mental event such as the recognition of a target stimulus) or the omission of a stimulus [30]. For AEPs, the response to the sound is represented as very small electrical voltage potentials originating from the brain. These electrical potentials are typically recorded by placing surface electrodes on the scalp of human beings. The stimuli could also be of different types, but they are most typically tones, clicks, speech or modulated noise. For instance, Figure 2.1 shows an AEP as a response to clicks.

As neural activity in response to a stimulus ascends the auditory pathway from the brainstem to the cortex, it is processed by the neural system. The difference in the layout of the neural architecture along the pathway, constitutes differences in latency of the response to the stimulus. Accordingly, AEPs to brief stimuli are often divided into three temporal groups – early, middle, and late latency responses [30], where latency is defined as the time between the onset of stimulus and the onset of the response. To have a more detailed understanding of the classification of AEPs, Table 2.1 provides a more detailed classification schema with description and examples to be looked at along with Figure 2.1.

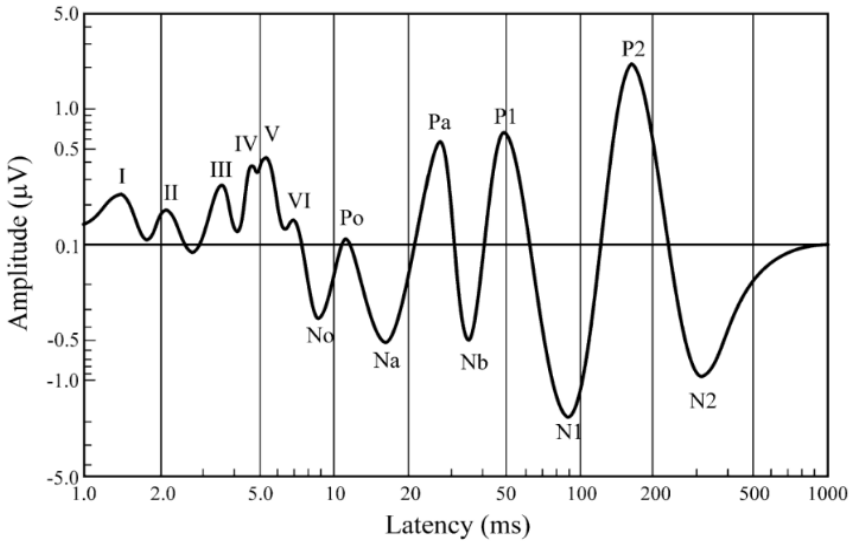

Figure 2.1 Transient auditory evoked potentials to a 60 dB nHL (above normal adult hearing threshold) click [30].

Early responses include the first and fast responses. The first response is the Compound Action Potential (CAP) of the auditory nerve. The fast response is also described as the Auditory Brainstem Response (ABR). In Figure 2.1 waves labelled as I-VI are the early response. Middle-Latency Responses (MLR) are the waves from No to Nb in Figure 2.1. Finally, P1-N2 waves represent the late or slow response, usually called Long Latency Auditory Evoked Potential (LLAEP).

In addition to classification based on latency, AEPs can also be classified based on their type of stimulation as transient, steady state and sustained. Transient AEPs occur following a change in stimulus such as an onset or offset, while Auditory Steady State Responses (ASSR) are evoked by a regularly changing stimulus. Sustained AEPs are also evoked via continuous stimulation.

2.2.1 Frequency Following Response

The Frequency Following Response (FFR) is a potential originating from the human auditory system typically recorded from the scalp in response to a brief tone. If the stimulus is speech,

Table 2.1 Classification of auditory evoked potentials by latency and type [30].

Latency	Transient	Steady-State	Sustained
First (0-5 ms)	Cochlear Nerve Compound Action Potential (CAP: N1, N2)	Cochlear Microphonic (CM)	Summating Potential (SP)
Fast (1-10 ms)	Auditory Brainstem Response (ABR: I-VII)	Frequency Following Response (FFR); Fast (>70 Hz) Auditory Steady-State Response (ASSR)	Pedestal of Frequency Following Response
Middle (10-50 ms)	Middle Latency Response (MLR: Na, Pa, Nb)	40-Hz Potential	
Slow (30-500 ms)	Vertex Potential (P1, N1, P2, N2)	Slow (< 30 Hz) Auditory Steady-State Response (ASSR)	Cortical Sustained Potential (SP)
Late (200-1000 ms)	Mismatch Negativity (MMN); Processing Negativity. Late Positive Waves		Contingent Negative Variation (CNV)

they are called as speech-evoked frequency following response. The speech sound can evoke both transient and sustained responses in the human brainstem and cortex. For instance, a pure vowel stimulus can evoke a transient brainstem response [31].

In the literature, a variety of terminologies have been used to represent the FFR concept. For instance, Skoe & Kraus [32] have referred to FFR as CABR (Complex Auditory Brainstem

Response) while describing how to record and analyze it. However, in many of the recent papers the concept is simply referred to as ‘FFR’, ‘human FFR’ or the ‘scalp recorded FFR’.

2.2.1.1 Envelope and Spectral FFR

Skoe & Kraus [32] divided FFRs (referred to as CABRs) into two components, namely the envelope response and the spectral response. As the name implies, in the case of speech stimuli, the envelope FFR (eFFR) refers to the measurement of the response of the auditory system to changes in the amplitude, or loudness, of an auditory stimulus [33]. On the other hand, spectral FFR (sFFR) assesses the response of the auditory system to changes in the frequency, or pitch, of an auditory stimulus [31]. A summary of possible combinations of the evoked responses to the original stimuli is provided in Table 2.2.

Table 2.2 Average response nomenclature. A summary of possible combinations of the evoked responses to the original stimuli and their opposite polarity, as well as the components contained within the response signal. Adapted from (Aiken & Picton, 2008) [31].

Response	Derivation	Component
+ +	Average together all responses to original stimuli	Envelope FFR
		Spectral FFR
		Cochlear microphonic
		Stimulus artifact
+ −	Average together an equal number of responses to original stimulus and responses to inverted stimulus	Envelope FFR
− −	Subtract responses to inverted stimulus from an equal number of responses to original stimulus and to the inverted stimulus. Divide by the total number of responses	Spectral FFR
		Cochlear microphonic
		Stimulus artifact

To obtain the envelope FFR, the envelope of the evoked electrical response of the auditory system to a continuous amplitude-modulated tone is extracted and its phase and amplitude are

compared to the phase and amplitude of the stimulus. The magnitude of the eFFR is calculated as the ratio of the amplitude of the response to the amplitude of the stimulus.

To calculate the spectral FFR, a continuous frequency-modulated tone is presented to the participant, and the evoked electrical response is recorded in a similar manner as for the eFFR. The magnitude of the sFFR is calculated as the ratio of the phase of the response to the phase of the stimulus [34], [35]. The issue here again is the terminology challenge. In the literature both envelope FFR and spectral FFR are sometimes also ambiguously referred to as the FFR. However, in this thesis they are differentiated by using the eFFR and sFFR terms, respectively.

2.3 Vowels as the stimuli for speech evoked FFR

As mentioned earlier, the speech sound can evoke both transient and sustained responses in the human brainstem and cortex [31].

Accordingly, many studies have been conducted using pure vowels as a stimulus to evoke responses. For instance, in several studies Krishnan [36], [37] analyzed the FFR of 10 normal hearing adults in response to three vowels (/a/, /ɔ/ and /u/) at four different sound levels (55, 65, 75, 85 dB nHL, i.e., above normal adult hearing threshold). His observations confirmed that using vowels as stimuli established a clear effect on the spectral amplitudes at both first and second formants.

In another study, Yi et al. [29] used vowel evoked FFRs to perform assessment of human auditory function in thirty-eight young adults. They used two native English speakers to generate two vowels (/a/ and /u/) to evoke the FFRs. Scalp-recorded responses were projected onto a low-dimension spectral feature space from the same vowels produced by 40 other speakers. They found that the ability to extract interpretable features with such trials provides significant potential for the assessment of human auditory function.

Yellamsetty and Bidelman [38] also used two steady-state vowels (/a/ and / ɛ/) for speech identification by human subjects through FFR. The vowels were presented diotically in both quiet and noise-degraded (+5dB SNR) conditions and their fundamental pitch frequencies F0 were different by zero or four semitones (0ST, 4ST). They also performed double vowel identification. Their results confirmed that FFR F0 amplitudes are more robust for single vowels

compared to double vowels and that they have a weaker response in noise. The FFR F0 amplitudes, however, failed to predict the listeners' identification performance. In contrast, the first formant FFR F1 was found to be related to faster reaction times when limited by noise. The F0 or fundamental frequency here refers to the lowest frequency of a periodic waveform, which is the rate at which the waveform repeats itself over time.

The approach from the above-mentioned studies using vowels as a stimulus for speech evoked FFRs was also adopted in the current study to perform listener classification and authentication. Accordingly, a dataset recorded from an earlier study [24] including the response of normal hearing adults to English vowel sounds, was used. All the four different vowel stimuli (100ms /a/, /ɔ/, /U/, /u/ vowels) were selected for this study. The data were recorded in test and retest sessions to offer the possibility of investigating the reliability of responses for the biometric identification task. Further details of the recording method can be found in Heffernan [24] but a summary is also given in the next section.

2.4 Experimental Setup

To record the FFRs, the Bio-logic BioMARK system was used (version v7.0.2), which used non-disposable silver chloride electrodes for recording the signals. The stimulus was presented to the right ear by a Bio-logic foam-tip phone insert. The left ear was kept excluded throughout the duration of the experiment. The active electrode was placed at the vertex (Cz) while the reference was placed on the right earlobe and the ground electrode was on the left earlobe. To keep the impedance under 6 kΩ, a mild abrasive and a conductive electroencephalogram (EEG) paste were applied on the participant’s skin. The EEG signals were amplified with a gain of 100,000 and bandpass filtered from 100 to 3000 Hz. To suppress the myogenic contamination, an artifact rejection criterion of 23.8 microvolts was applied. The stimulus onset was synchronized with the recording and the stimulus presentation rate was set at 8.4/s. The recording system fixed the maximum 1024-points per epoch, which resulted in an epoch time of 106.6 ms, and a sampling rate of approximately 9606 Hz. Also, a 3.4 ms onset recording delay was applied to maximize the capture of the sustained response.

2.4.1 Stimulus Creation

In his study, Heffernan created 100 ms synthetic vowel sounds, namely /a/ as in ‘father’, /ɔ/ as in ‘call’, /U/ as in ’boot’, /u/ as in ‘who’ as the four vowel stimuli [24]. All four selected stimuli had a common fundamental frequency (F0) of 100 Hz. Also, a 100 ms duration was selected in order to have an integer number of periods at F0. To generate the stimuli, a simplified Klatt formant synthesizer with a sampling rate of 48kHz at 16-bit resolution was used. The synthesized vowel parameters are indicated in Table 2.3. All the stimuli were presented to each subject at the level of 85dB. The levels of the sound were calibrated with an insert earphone connected to a 2-cc Bruel & Kjaer DB0138 coupler with Bruel & Kjaer Type 4144 microphone and subsequently to a Bruel & Kjaer type 2235 sound level meter.

Table 2.3 Duration, fundamental (pitch) frequency, first, second and third formant frequencies, bandwidths and relative levels of created stimuli.

Vowel	Duration (ms)	F0 (Hz)	F1 (Hz)	F2 (Hz)	F3 (Hz)	BW1 (Hz)	BW2 (Hz)	BW3 (Hz)	A1 (dB)	A2 (dB)	A3 (dB)
/a/	100	100	700	1200	2600	130	70	160	-1	-5	-28
/ɔ/	100	100	600	900	2400	100	60	110	0	-7	-34
/U/	100	100	500	1200	2200	80	100	80	-1	-12	-34
/u/	100	100	300	900	2200	65	110	140	-3	-19	-43

2.4.2 Recording Session

In order to enable an analysis of test-retest stability of the responses, two recording sessions, test and retest, were conducted. For each vowel, the response at 85dB sound level was recorded in two consecutive 1500 sweep blocks, giving a total of 3000 sweeps per session. Consequently, a

total of 6000 sweeps were collected for each subject. The sweeps were presented in alternating polarity and the envelope FFR (eFFR) was extracted by averaging the summed responses in each of the two polarities. Also, the spectral FFR (sFFR) was extracted by averaging the difference between the responses from the two polarities [31].

2.4.3 Subjects

A total of twenty-two normal-hearing English speaking adult subjects (11 males and 11 females) ranging from 20 to 35 years of age participated in this study. As discussed earlier, the recordings took place over two sessions with a minimum of 3 days apart, and 23 days apart on average. All the subjects provided consent and all the recordings were performed according to the requirements of University of Ottawa's Ethics Board.

Before recording, each subject underwent an audiometric test to confirm that their hearing thresholds were 20dB HL or less at 250, 500, 750, 1000, 2000, 4000 Hz in both ears. During the recordings, participants were seated comfortably on a reclining chair in a sound attenuating booth that passes the ANSI standard for background noise level, ANSI S3.1-1999 (R2013). They were encouraged to keep awake and still during the recording session.

During the process of recording, two additional female subjects were excluded as one failed the audiometric threshold testing and the other one, despite passing the audiometric testing, had been very recently referred to an ENT by an audiologist. Consequently, the index of subjects ranges from 1 to 25 in this work but misses the numbers 5, 10, and 24, as will be shown in the data analysis section.

2.5 Data Preprocessing

All recorded responses were exported to ASCII text files using the Bio-logic "AEP to ASCII" software version 1.2.1. Later, to facilitate the analysis using Python, the ASCII text files were converted into a Pandas library DataFrame data structure. In the time domain, all the signals were detrended by removing the DC offset as it is a potential source of distortion [39]. Then, a Hamming window was applied so as to suppress the spectral leakage after calculating the Fourier transform. Since zero padding is also helpful for obtaining a more densely interpolated frequency domain representation, this technique was also used to facilitate estimating correctly the amplitude at a given frequency (for example, at the fundamental frequency of the FFR of 100 Hz). Accordingly, a series of 86454 zeros was added to the end of the time domain signals. Finally, a normalization using the root mean square as the normalizing term, was also performed to make different FFRs comparable.

Figure 2.2 Data preprocessing flowchart.

2.6 Data Analysis

This section presents the analysis of the recorded data for the twenty-two subjects in both time and frequency domains. In order to perform the analysis in this section, it is useful to review the definitions of fundamental frequencies and higher order harmonics.

The fundamental frequency (F0) is the lowest frequency of a periodic waveform, which is the rate at which the waveform repeats itself over time, as previously mentioned. It is often referred to as the first harmonic and is the primary determinant of the perceived pitch of a sound. In addition to F0, there are higher harmonics that are integer multiples of the F0, including the second harmonic (H2) and the third harmonic (H3), which correspond to frequencies that are twice and three times the F0, respectively. These harmonics are also important in determining the timbre or quality of a sound, as they contribute to the complex waveform of a sound signal.

Also, the frequencies beyond the first three harmonics of the fundamental frequency are often referred to as higher order harmonics. These harmonics are integer multiples of the fundamental frequency and are represented by H4, H5, and so on. For example, the fourth harmonic (H4) would correspond to a frequency that is four times the fundamental frequency, while the fifth harmonic (H5) would correspond to a frequency that is five times the fundamental frequency. Looking at the amplitude of the eFFRs at these fundamental and higher order harmonics can be a useful feature in the data analysis.

In this regard, Figures 2.3, 2.4, 2.5 and 2.6 show the amplitude spectra of the eFFR for each of the /a/, /ɔ/, /U/ and /u/ vowels, respectively. Also, Figures 2.7, 2.8, 2.9 and 2.10 show the amplitude spectra for the sFFR for each of the /a/, /ɔ/, /U/ and /u/ vowels, respectively. The figures include both the test and retest conditions, pertaining to tests performed on different days. In these figures, each curve therefore represents the average of two sets of recordings (each of 1500 trials on the same day). The amplitude spectra of the retest and test signals from 0 to 1300 Hz are color coded with blue and orange lines, respectively.

Figure 2.3 shows that for the /a/ vowel's FFR, there is a relatively similar tendency across the subjects, based on the amplitude of the peak at F0, and H2 to H6. It can also be seen that the spectral amplitude of H2 is often higher than that of F0, with usually a decreasing trend from H3 to H6. Also, comparing between the subjects, most of them have a stable response between the test and retest conditions. However, a much different pattern at certain harmonics can also be seen in some of the subjects. For instance, subject 9 has substantially different amplitudes at F0, H5 and H7 in test and retest conditions. Also, the spectral FFRs for the /a/ vowel presented in Fig. 2.7, shows relatively clear harmonics for most of the subjects at F1 (700Hz). Most subjects seem to have a clear peak at F2 (1200 Hz) as well. However, the spectral amplitudes in the higher frequencies are relatively smaller. It can also be seen that at lower frequency components such as at F0, it is more difficult to identify the response components.

Figure 2.4 also shows the averaged envelope FFRs for the /ɔ/ vowel. Here also clear peaks at F0 and at H2 to H6 can be observed. In most of the subjects, similar spectral amplitudes at F0 and H2 are observed. However, in certain subjects such as subject 8 and 18, the peaks at F0 appear to be relatively suppressed. Again, comparison between subjects shows that the spectral amplitude of H2 is most often the highest one compared with the other harmonics. Also, clear

peaks around F1 (600 Hz) can be observed in Figure 2.8, where the spectral FFRs for the /ɔ/ vowel are presented. However, the peaks don't look very stable for subjects 2, 3, 7 and 21. The response at F2 (900 Hz) is also difficult to visually observe. Also here, comparing the subjects seems to be more difficult than in eFFRs, particularly at lower harmonics.

Figure 2.5. shows the envelope FFRs for the /U/ vowel. Here, some slight changes in pattern compared with the responses to vowels /a/ and /ɔ/ are observed. In the responses to vowel /U/, F0 shows more stability than other harmonics. It is also observed that the spectral amplitude at 500 Hz (H5) is generally slightly higher than the amplitude at 400 Hz (H4). Again, comparing the subjects shows a relatively high similarity in the waveforms from H4 to H6 except for subject 2, where the spectral amplitudes of H4, H5, and H6 are substantially different. Also, in most of the subjects the figure shows very good matching between the F0 and H2 to H3 except subjects 21 and 23. Meanwhile, the spectral FFRs for the /U/ vowel in Figure. 2.9. show various patterns at F1 (500 Hz) and F2 (1200 Hz). However, good matching between test and retest conditions is usually maintained in most of the subjects.

Finally, Figure 2.6 shows the envelope FFRs for the /u/ vowel. In this case, a clear pattern compared to the other three vowels is observed since the spectral amplitudes of F0 and H2 to H6 are clearly distinguished for most of the subjects. Indeed, except for subject 17 where the amplitudes of H3 to H6 are difficult to visually observe, most of the subjects show good matching between test and retest conditions. On the other hand, the spectral FFRs for the /u/ vowel in Figure 2.10 don't provide any clear pattern for F1 (300 Hz) and F2 (900 Hz). There is a poor matching between test and retest conditions, except for subject 22, who shows strong peaks at F0 and H2 to H6 and good similarity between test and retest conditions.

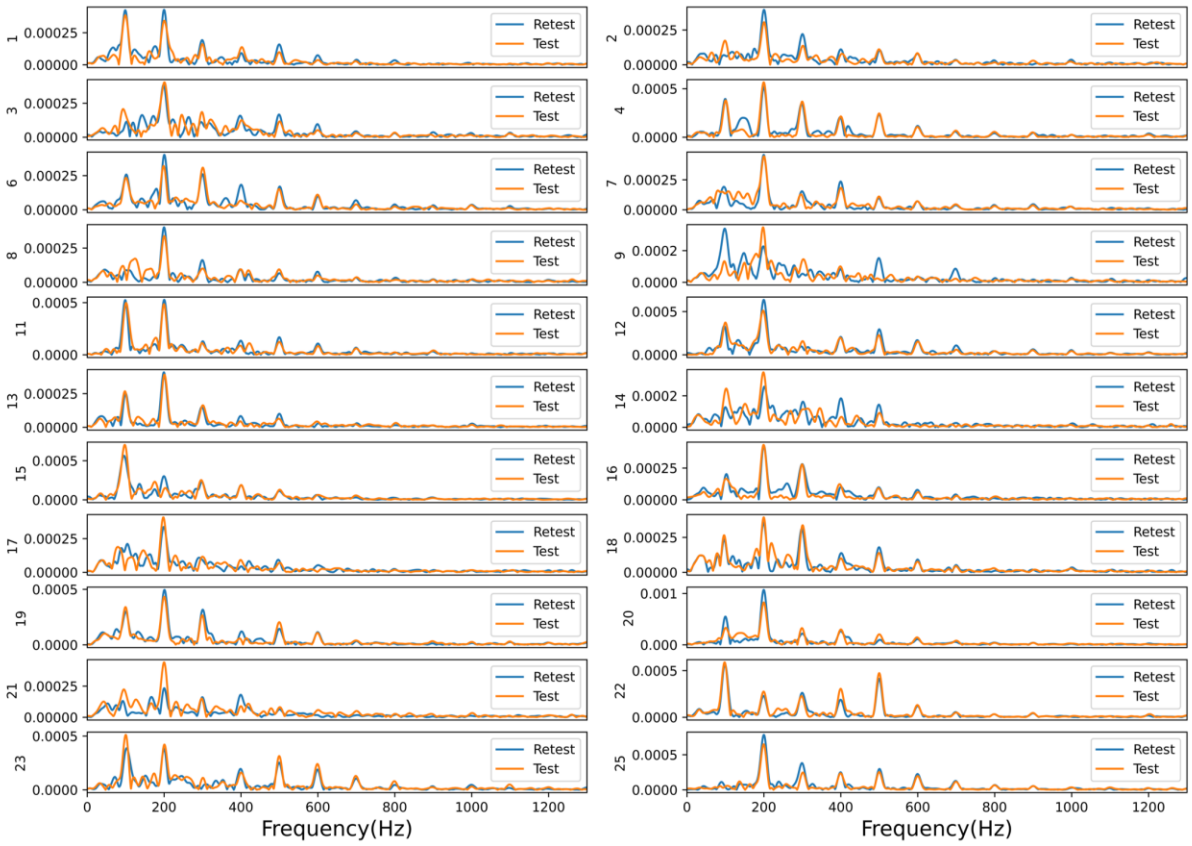

Figure 2.3 Amplitude spectrum of frequency components of envelope FFR for a 100 ms /a/ vowel stimulus with F0 = 100 Hz presented at 85 dB for all 22 subjects.

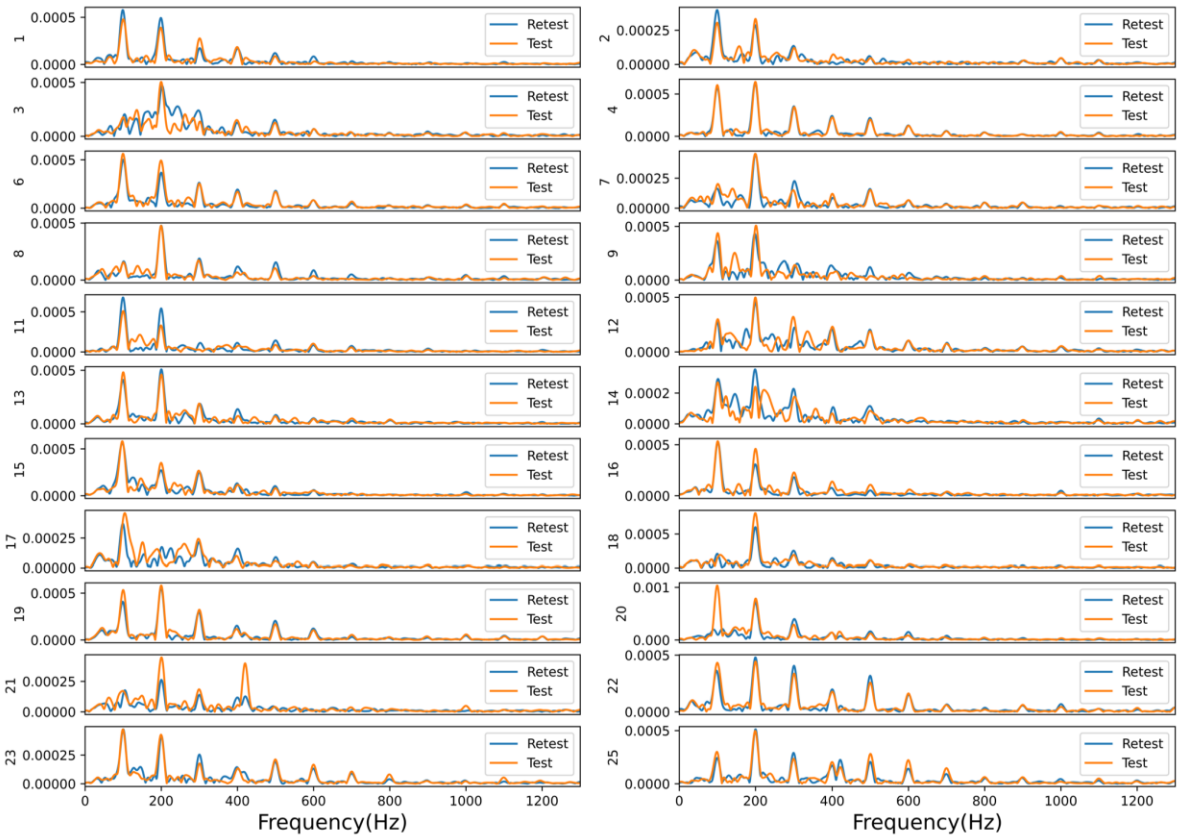

Figure 2.4 Amplitude spectrum of frequency components of envelope FFR for a 100 ms /ɔ/ vowel stimulus with F0=100 Hz presented at 85 dB for all 22 subjects.

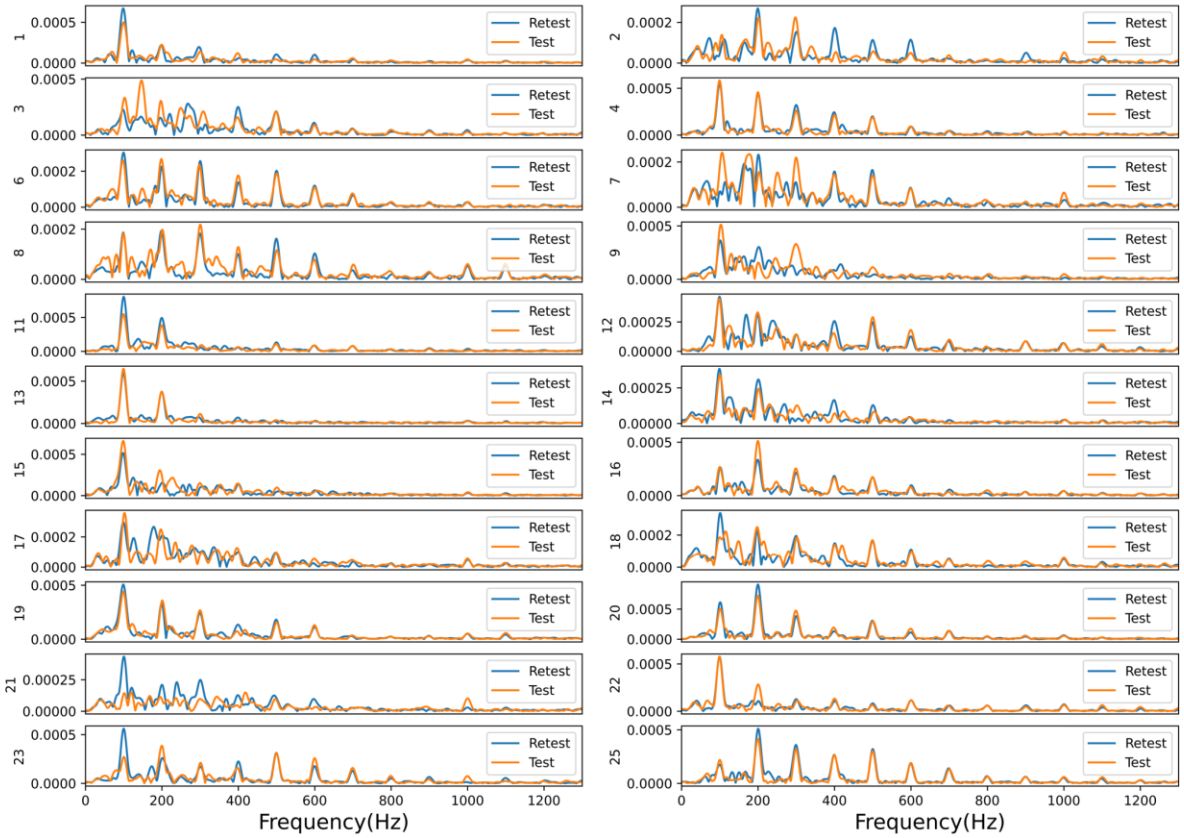

Figure 2.5 Amplitude spectrum of frequency components of envelope FFR for a 100 ms /U/ vowel stimulus with F0=100 Hz presented at 85 dB for all 22 subjects.

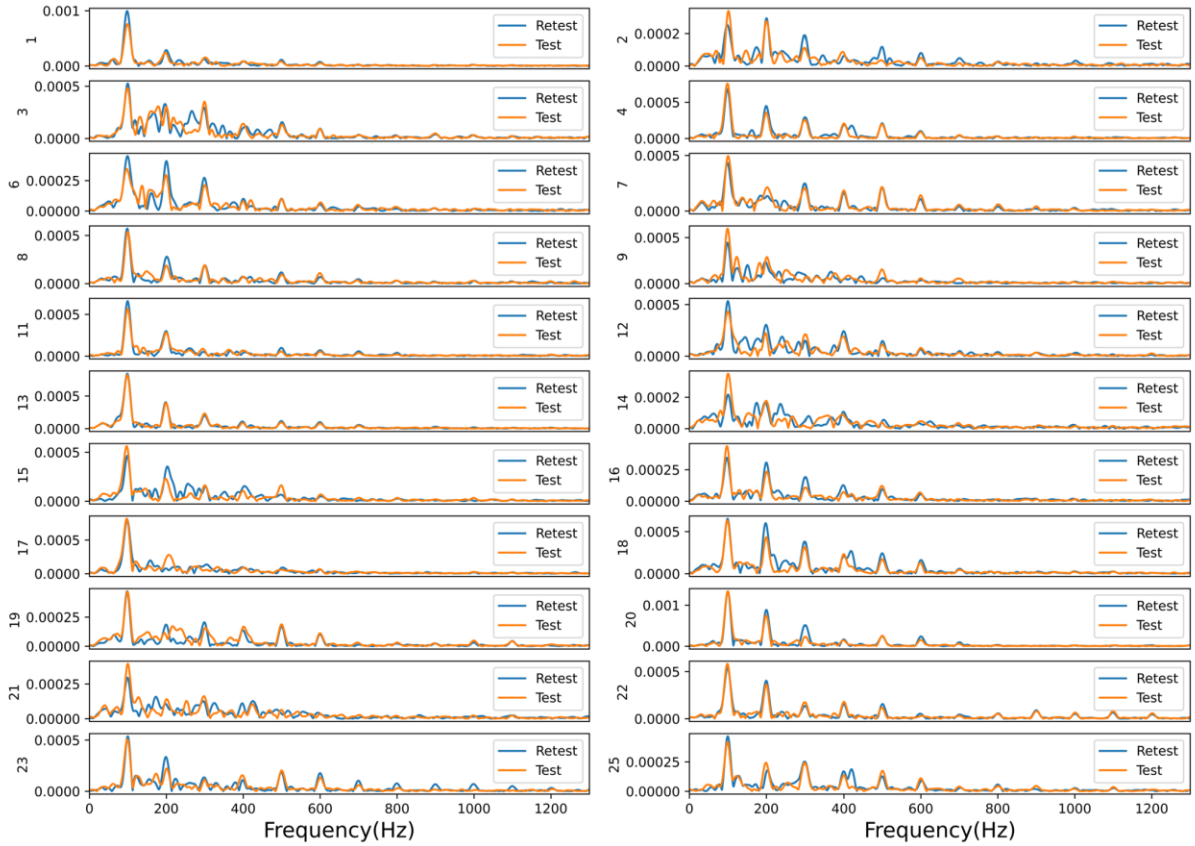

Figure 2.6 Amplitude spectrum of frequency components of envelope FFR for a 100 ms /u/ vowel stimulus with F0=100 Hz presented at 85 dB for all 22 subjects.

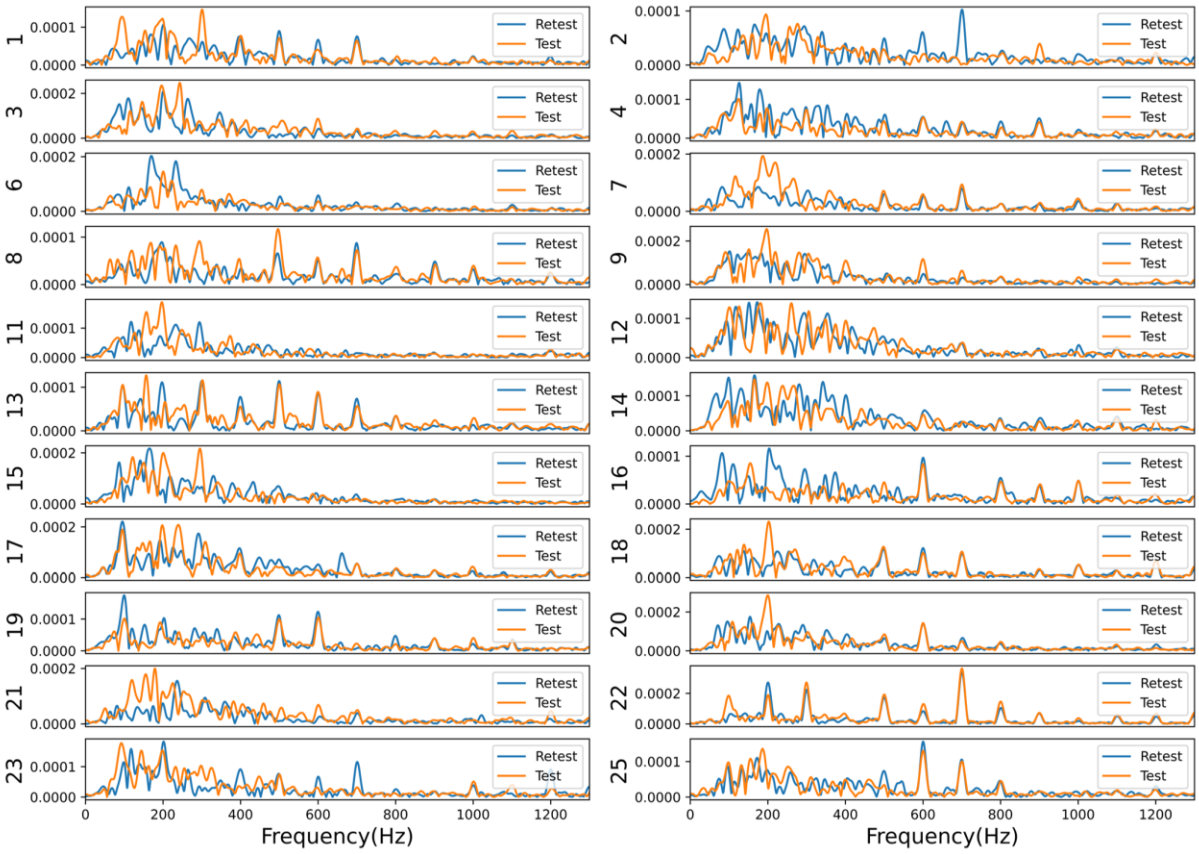

Figure 2.7 Amplitude spectrum of frequency components of spectral FFR for a 100 ms /a/ vowel stimulus with F0=100 Hz presented at 85 dB for all 22 subjects.

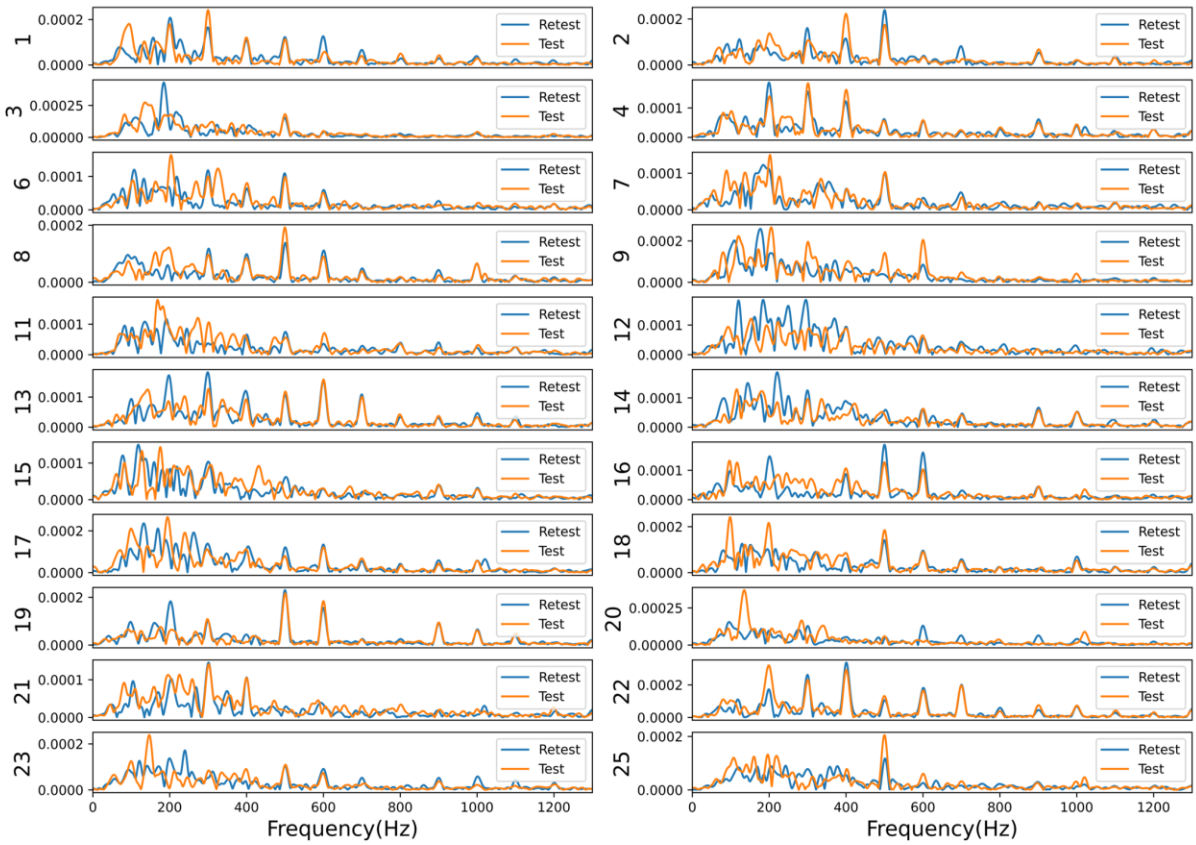

Figure 2.8 Amplitude spectrum of frequency components of spectral FFR for a 100 ms /ɔ/ vowel stimulus with F0=100 Hz presented at 85 dB for all 22 subjects.

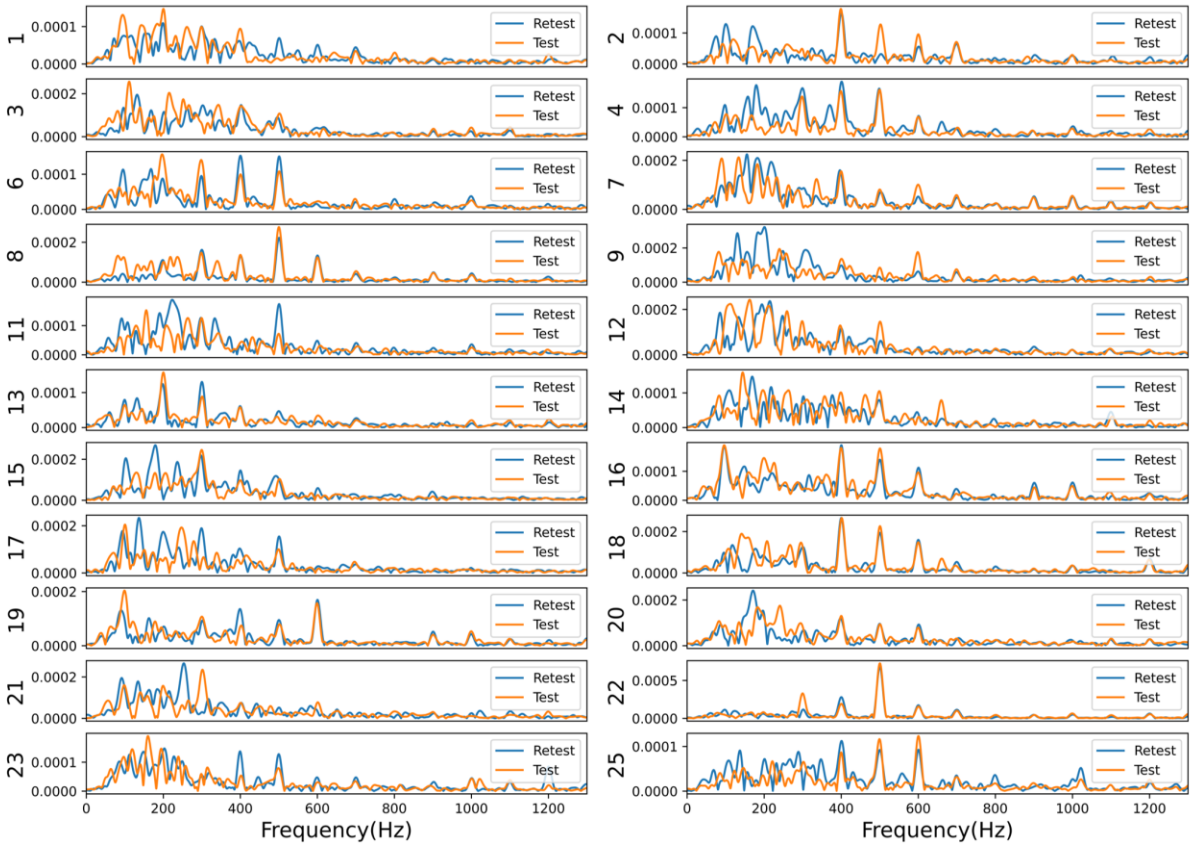

Figure 2.9 Amplitude spectrum of frequency components of spectral FFR for a 100 ms /U/ vowel stimulus with F0=100 Hz presented at 85 dB for all 22 subjects.

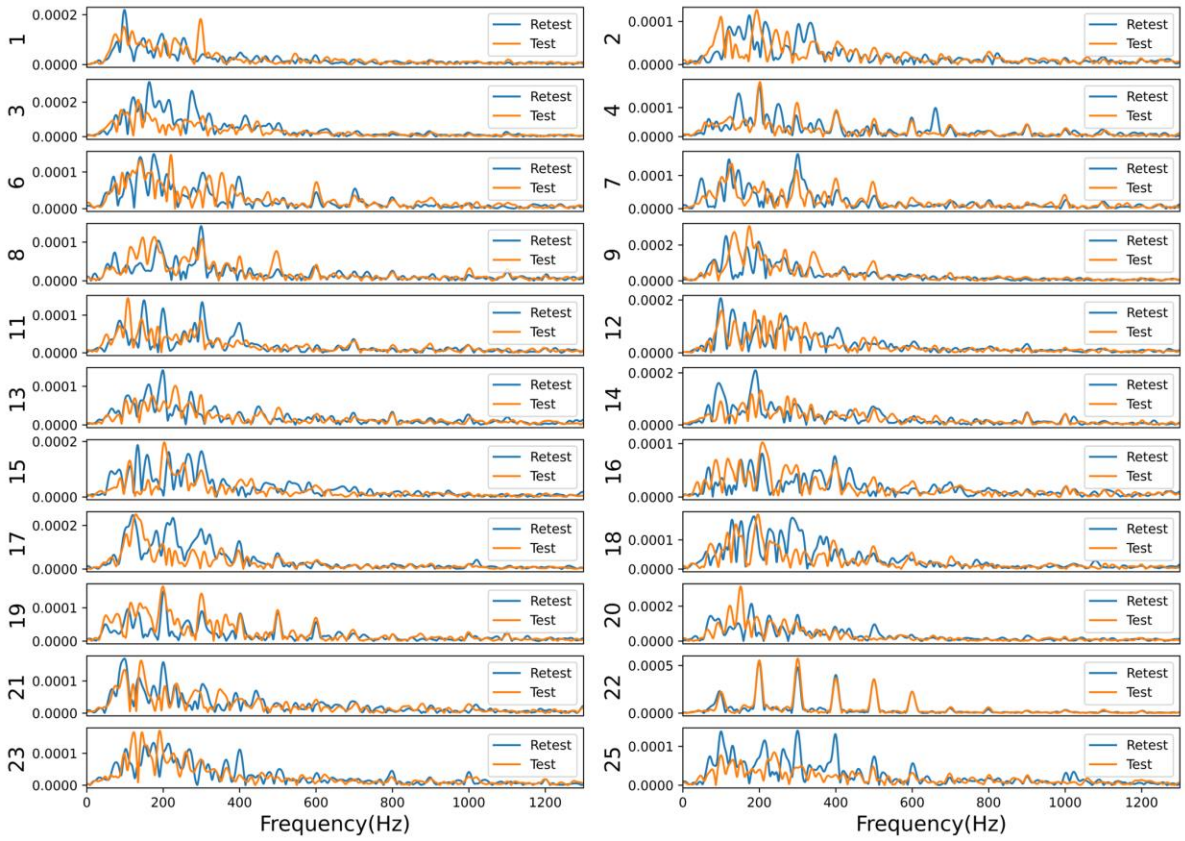

Figure 2.10 Amplitude spectrum of frequency components of spectral FFR for a 100 ms /u/ vowel stimulus with F0=100 Hz presented at 85 dB for all 22 subjects.

Chapter 3 Feature Extraction

3.1 Overview

This chapter reports on the extracted features from the FFRs such as spectrograms and gammatonegrams.

3.2 Extracting Features for Signal Classification

In order to train machine learning models, data should be inputted to the model in an appropriate form. This input data comprises features, which are basically independent measurable variables, properties, or characteristics of the input data [40]. Choosing informative, discriminating, and independent features is a crucial element of effective algorithms in machine learning applications such as classification, regression, etc. [41]. Features are usually numeric, but structural features such as graphs and strings are also used in some applications such as pattern recognition.

In any specific machine learning problem, what is required to be learned is a set of these features as independent variables, coefficients of these features, and parameters for coming up with appropriate functions or models (also termed hyperparameters) [40], [42]. To provide the features to the model, first the input dataset should be prepared to be compatible with the machine learning algorithm requirements. Also, since the performance of machine learning models depends on the quality of the features, they should be selected to be as informative, discriminating, and independent as possible [41].

Accordingly, this chapter focuses on the appropriate feature selection from the FFRs as the inputs of the machine learning models in this study. On this basis, two different types of features have been extracted from the FFRs: spectrograms and gammatonegrams.

3.2.1 Spectrograms

The spectrogram is a visual representation of the time-frequency-intensity content of a signal and is popularly used with audio signals [43] [44]. There are three dimensions in a spectrogram, two are frequency (y-axis) and time (x-axis), and the third dimension is magnitude or squared magnitude (intensity). A sample spectrogram of an eFFR recorded in this study calculated by

the spectrogram module of the Python Scipy library [45] is shown in Figure 3.1. As expected, bands of high amplitude (yellow stripes) can be seen at 100Hz, 200Hz, 300Hz, 400Hz, 500Hz and 600Hz. Also, it can be seen that as the frequency increases, the amplitude of the peaks (intensity of the yellow color) decreases, which is in agreement with what was observed in Figures 2.2 to 2.5.

Spectrograms are calculated from the time domain signal using the Fourier Transform. They can

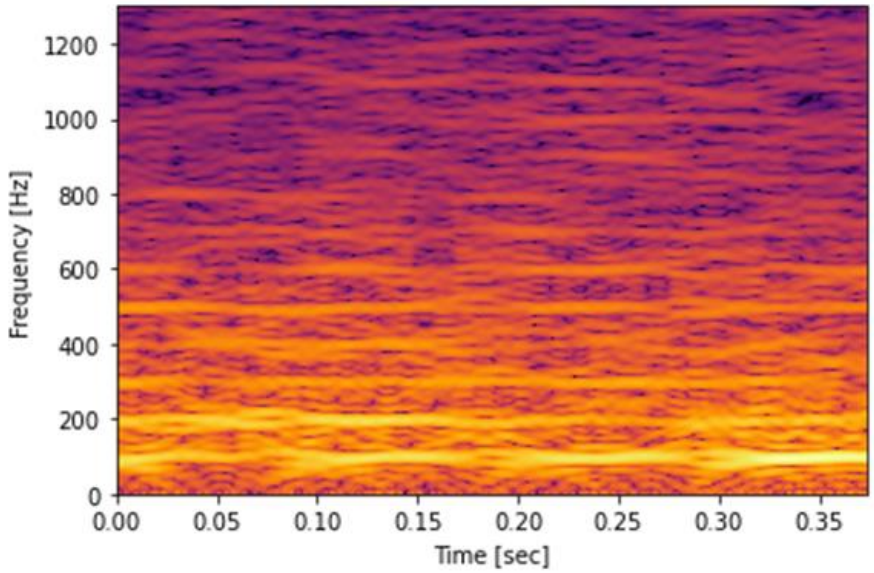

Figure 3.1 A sample spectrogram extracted from an FFR.

be defined as an intensity plot (usually on a log scale, such as dB) of the Short-Time Fourier Transform (STFT) magnitude. The STFT is simply a sequence of FFTs of windowed data segments, where the windows are usually allowed to overlap in time, typically by 25-50% [46]. Consequently, to calculate spectrograms, first the signal is sampled in the time domain and then it is segmented into several overlapping windows. Then, the Short-time Fourier Transform (STFT) is used to calculate the frequency spectrum for each window, where each window represents a vertical line in the spectrogram image. These parts are put together side by side and the size and the overlap of the windows define the computational resolution of the spectrogram (which depends on the number of "pixels" or data points in the spectrogram). The process of creating a spectrogram essentially involves the calculation of the squared magnitude of the STFT of the signal for a particular window location.

On this note, it is worthwhile mentioning that the computational resolution of spectrograms mentioned above, and the physical frequency resolution of spectrograms (or STFTs) are two different aspects and should not be confused together. The computational resolution of a spectrogram image depends on the number of "pixels" or data points computed to produce the spectrogram image, and it is a property of the image itself. It can affect its visual appearance and result in a more detailed representation but may also increase the file size and processing time required for the image.

On the other hand, the physical frequency resolution refers to the smallest change in frequency that can be accurately represented in the spectrogram, which is determined by the length of the analysis window. It measures how two sinusoidal time signals with closely located frequency components could be distinguished in the spectrum and how much a spectrum is smoothed in general as a consequence of windowing and finite length observation. The physical frequency resolution (in Hz) represents the amount of observable fine details in a continuous-scale frequency spectrum, and can be defined by $\frac{FWHM}{W_n} \times f_s$ where $FWHM$, W_n and f_s are the Full-Width-Half-Maximum measure (in Hz, bandwidth measured from the frequency response of a 1-sec. window), the window size (in samples) and the sampling frequency (in samples/sec.), respectively. Here, a larger window length W_n leads to better physical frequency resolution.

Spectrograms are an important representation of audio data since human hearing is based on a kind of real-time spectrogram encoded by the inner ear or the cochlea [47]. Spectrograms have been used extensively in the field of computer music as a guide during the development of sound synthesis algorithms.

3.2.1.1 Investigation of Spectrograms of the FFRs

As discussed above, in this study spectrograms have been selected as an appropriate type of features to be used in machine learning models to serve the purposes of classification and authentication defined for the study. For instance, in the classification application, the spectrograms of the FFRs are fed into machine learning models to classify the 22 different listeners. Accordingly, in one case study, the FFRs recorded for each of the four vowels (106.6 ms signal) are concatenated together at each sound level. The spectrograms of the formed concatenated four-vowel FFR (a 426.4 ms signal) are then used for subject classification at a

specified sound level (i.e., 85 dB). For instance, Figure 3.2 shows the windowing process for the concatenated four-vowel signal.

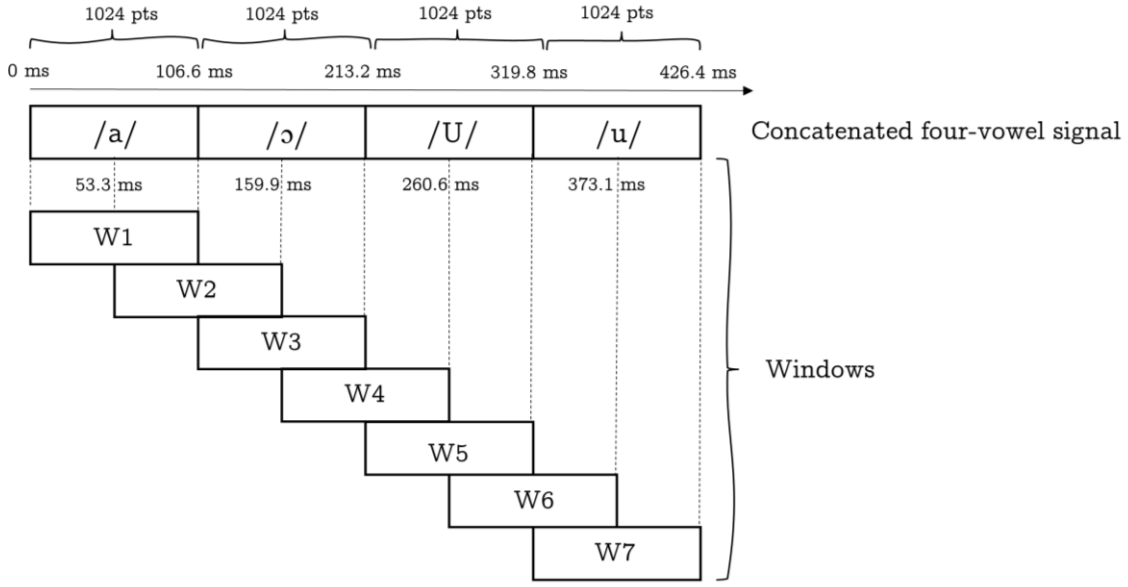

Figure 3.2 Windowing process on the four-vowel signal with 50% overlap of windows.

Based on the procedure discussed above and shown in Figure 3.2, an increase in overlap of the windows will result in an increase in the computational resolution of the spectrogram. Figures 3.3, 3.4, 3.5 and 3.6 show the spectrogram of a concatenated four-vowel FFR for a window size of 512 points (half-length of each of the vowels) and overlaps of 0, 256, 384 and 511 points, respectively. It can be seen that as the overlap increases, the resolution of the spectrogram images increases as well. This would suggest that higher resolution spectrograms would contain more information for training the machine learning models. However, it should be kept in mind that training the model with high resolution inputs would be more expensive and time consuming.

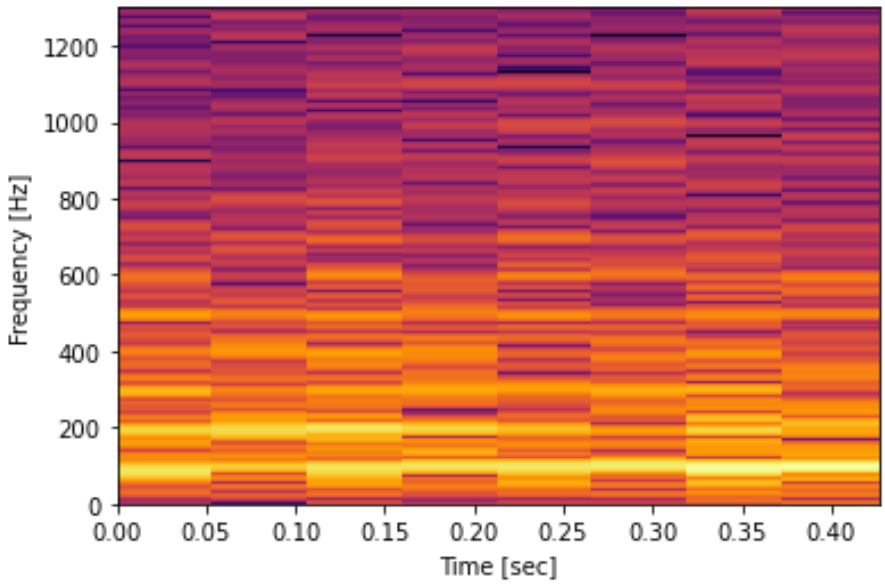

Figure 3.3 Spectrogram of a concatenated four-vowel FFR with window size of 512 points and overlap of 0 point.

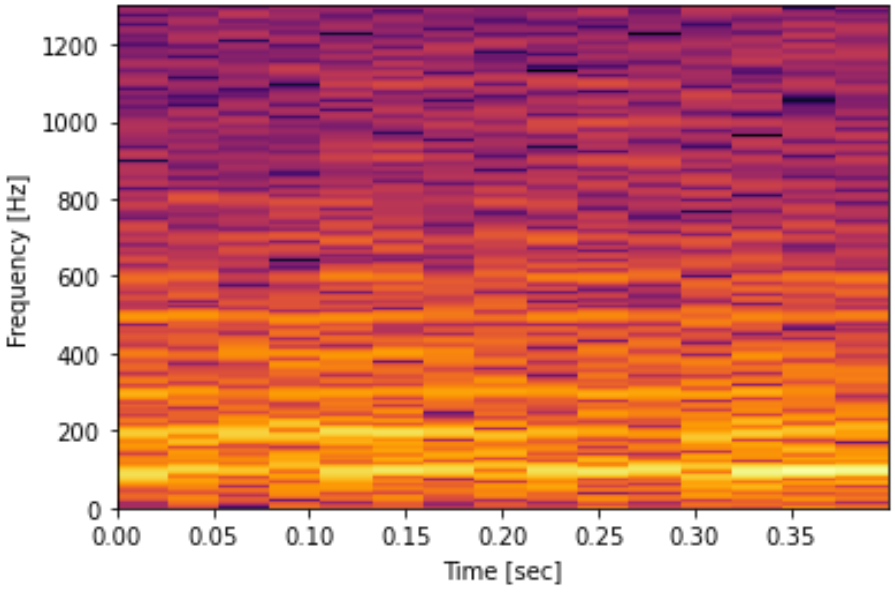

Figure 3.4 Spectrogram of a concatenated four-vowel FFR with window size of 512 points and overlap of 256 points.

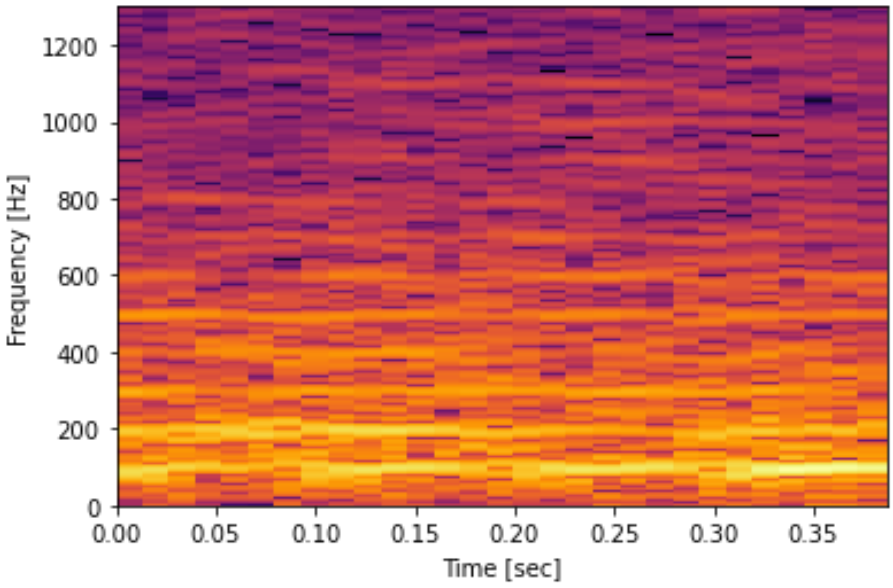

Figure 3.5 Spectrogram of a concatenated four-vowel FFR with window size of 512 points and overlap of 384 points.

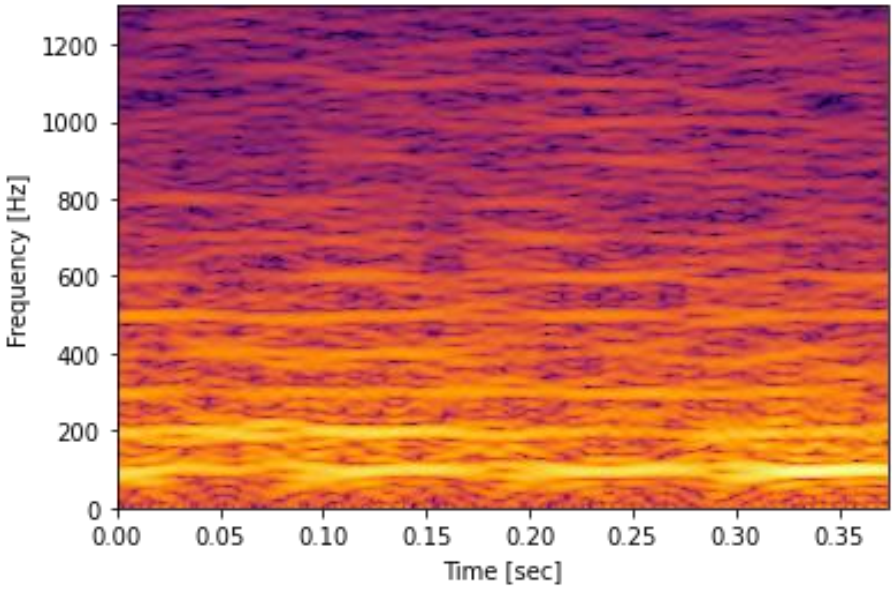

Figure 3.6 Spectrogram of a concatenated four-vowel FFR with window size of 512 points and overlap of 511 points.

3.2.2 Gammatonegrams

The spectrograms discussed earlier are the traditional time-frequency visualization of the sound. However, there are some fundamental differences between the spectrogram representations and how sound is actually analyzed by the ear, most significantly that the ear's frequency subbands get wider at higher frequencies, whereas the spectrogram has a constant bandwidth across all frequency channels. Accordingly, there have been many signal processing approximations proposed for the frequency analysis performed by the ear that have inspired other types of visualizations. One of the most popular approximation is the gammatone filter bank originally used by Patterson et. al. [48], [49]. Gammatones are actually wavelets-like representations with better frequency resolution at low frequencies and better time resolution at high frequencies.

The gammatone filter bank is a popular linear approximation of human peripheral auditory filtering. Patterson showed that gammatone functions provide a good fit to the experimentally determined auditory responses among various auditory filter bank models [48]. Gammatone filters have a repeated pole structure leading to an impulse response that is the product of a Gamma envelope $g(n) = t^n e^{-t}$ and a sinusoidal tone (Figure 3.7).

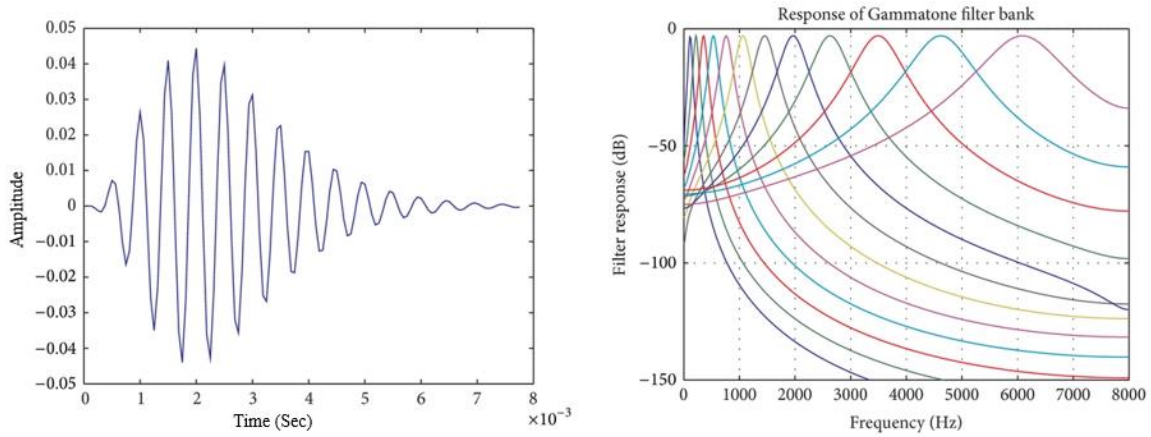

Figure 3.7 Gammatone impulse response function of a cochlear filter (left) and gammatone filter bank with 25 filters in the frequency domain (right). Adopted from [70].

Based on what was mentioned above, it is very natural to visualize sound as a time-varying distribution of energy in frequency since this is one way of describing the information our brains get from our ears via the auditory nerve. Therefore, gammatonegrams or gammatone-like spectrograms are constructed as an alternative to spectrograms that are more consistent with the hearing process of human beings. To construct gammatonegrams, a signal is first processed by all the filters in a gammatone filter bank. In order to convert this into a time-frequency visualization, the energy is then summed up within regular time windows. The successive windows are usually allowed to overlap in time. The distance in time between successive windows referred to as hop, affects the computational temporal resolution of the analysis. A smaller hop size will provide more detail in the temporal domain but requires more

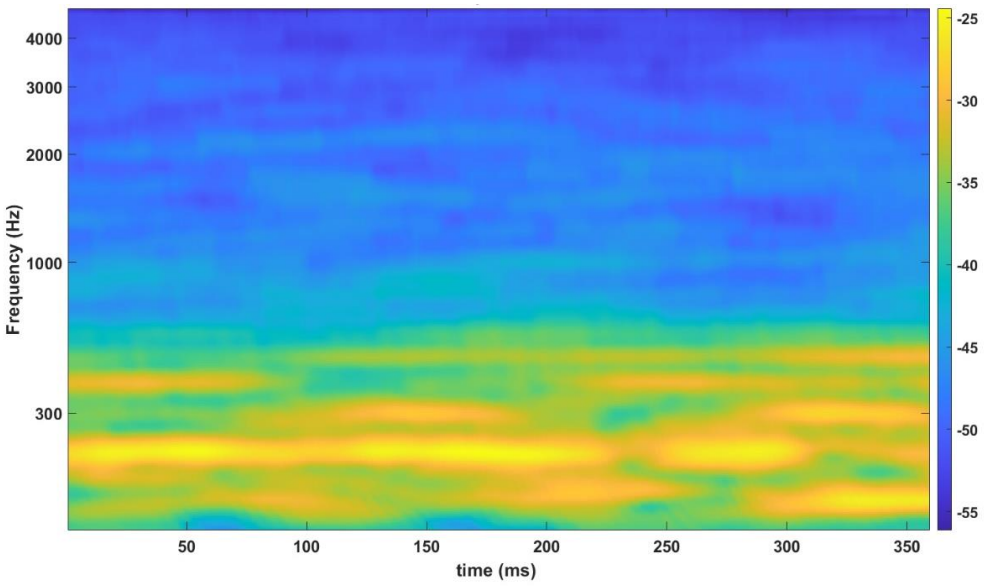

Figure 3.8 Gammatonegram of a sample concatenated four-vowel eFFR with 256 filters.

computational resources, while a larger hop size will reduce the amount of detail but require fewer resources. For instance, Figure 3.8 shows a gammatonegram constructed from a sample concatenated four-vowel eFFR using 256 gammatone filters, a window size of 53.3 ms (512 samples) and hop of 1 ms (10 samples). The window size has been selected as 53.3 ms (512 samples) in order to be analogous with the set of spectrograms that had been constructed earlier with the same window size of 512 samples. However, by changing the hop size or in other words the windows overlap, it's effect on the computational temporal resolution of gammatonegrams was investigated.

Accordingly, Figures 3.9, 3.10, 3.11, 3.12 and 3.13 show the gammatonegram of a concatenated four-vowel FFR for a window size of 53 ms (half-length of each of the vowels) and hopping sizes (window shifts) of 53, 26, 10, 5 and 1ms, respectively. It can be seen that as the hopping size decreases (overlap increases), the computational temporal resolution of the gammatonegram images increases as well. This would suggest that higher temporal resolution gammatonegrams contain more information for training the models. However, it should be considered that training the model with high resolution inputs would be more expensive and time consuming. Along with spectrograms, gammatonegrams have been used as inputs for the machine learning models selected for this study, and the results obtained are reported in Chapter 4.

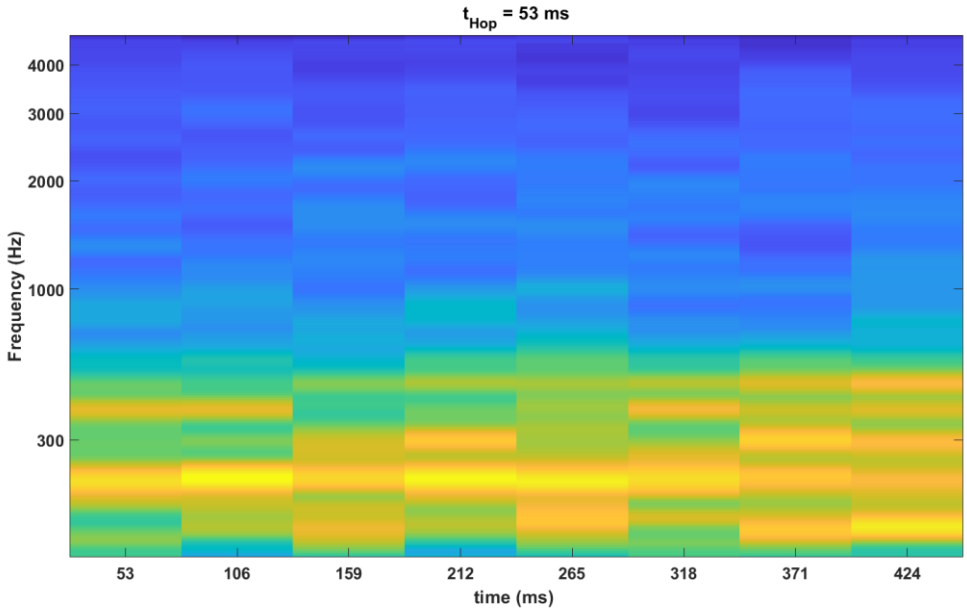

Figure 3.9 Gammatonegram of a sample concatenated four-vowel eFFR with 256 filters, window size of 53ms and t_{Hop} = 53ms.

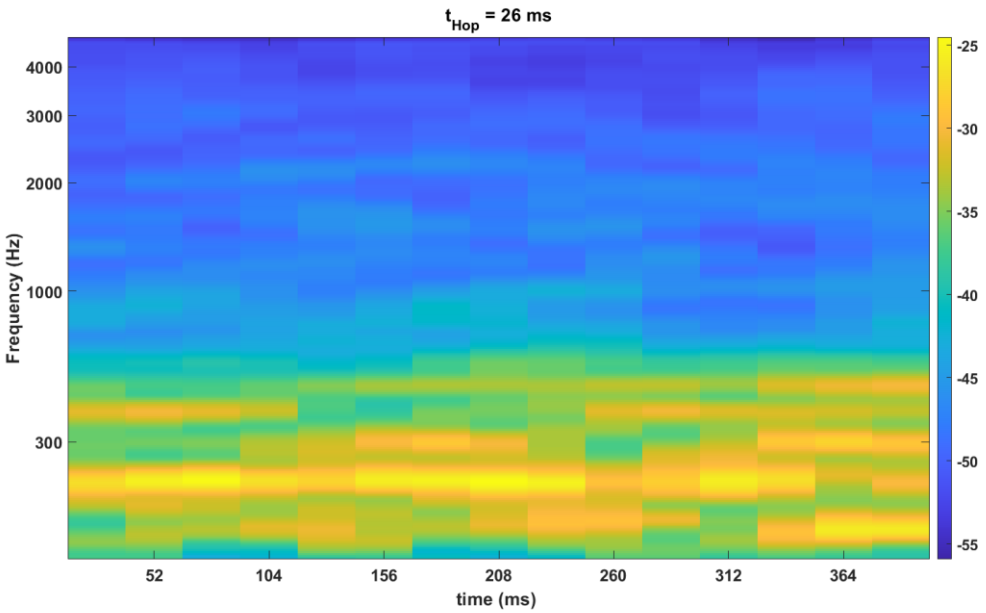

Figure 3.10 Gammatonegram of a sample concatenated four-vowel eFFR with 256 filters, window size of 53ms and t_{Hop} = 26ms.

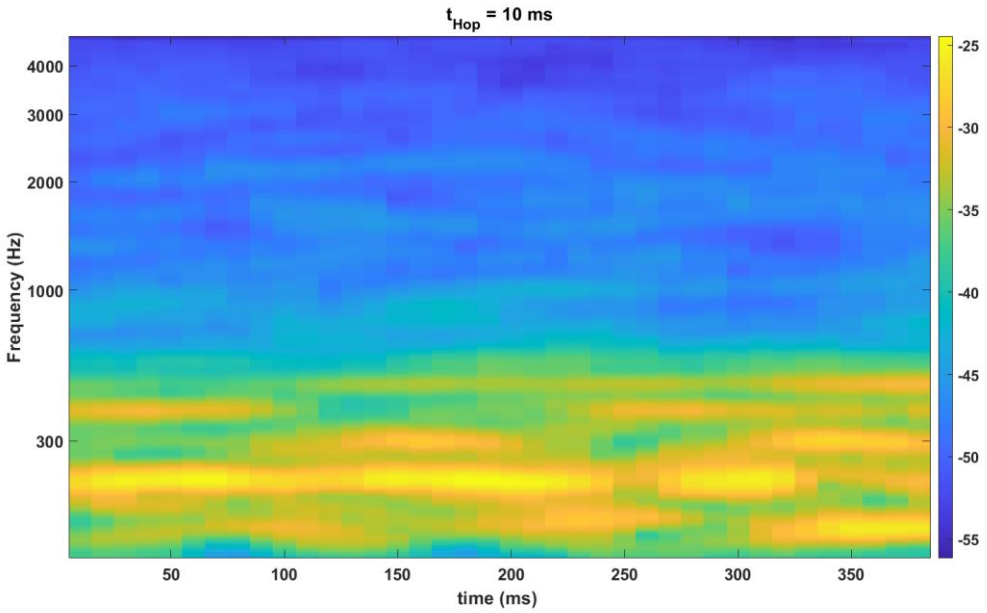

Figure 3.11 Gammatonegram of a sample concatenated four-vowel eFFR with 256 filters, window size of 53ms and t_{Hop} = 10ms.

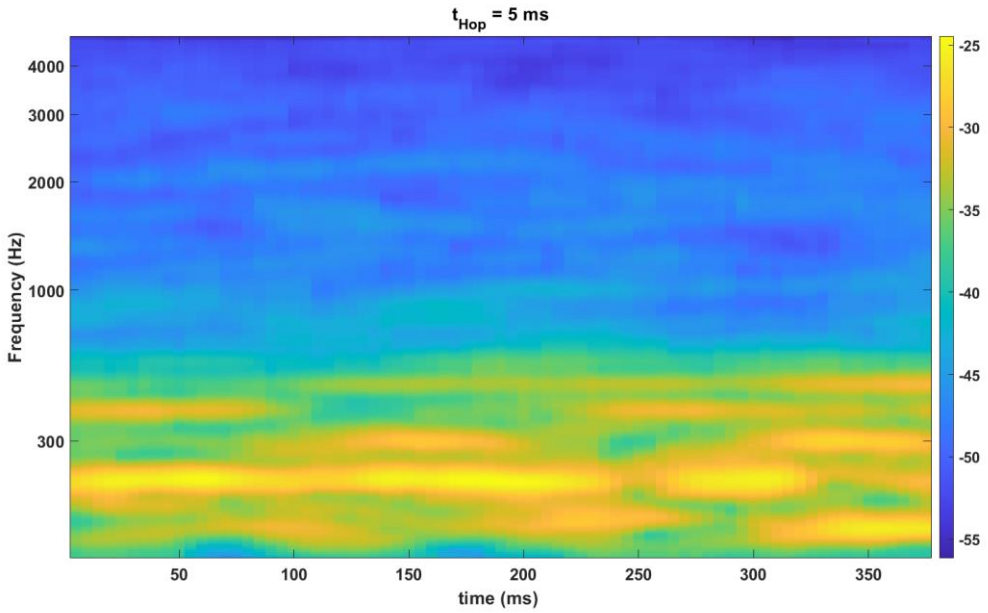

Figure 3.12 Gammatonegram of a sample concatenated four-vowel eFFR with 256 filters, window size of 53ms and t_{Hop} = 5ms.

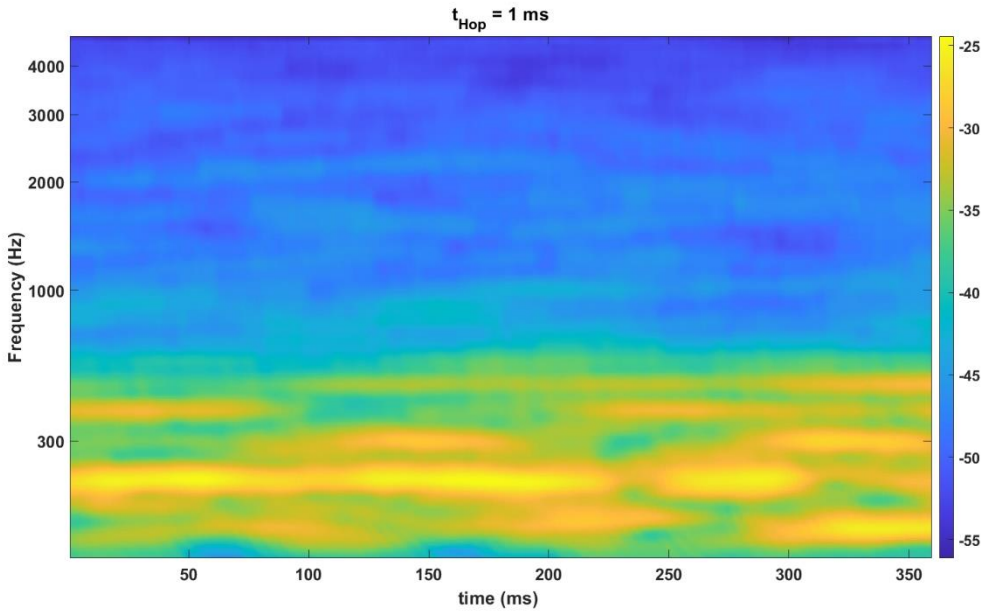

Figure 3.13 Gammatonegram of a sample concatenated four-vowel eFFR with 256 filters, window size of 53ms and t_{Hop} = 1ms.

Chapter 4 Listener Classification

4.1 Overview

This chapter reports on the results of the machine learning models used to perform subject classification on the extracted features described in Chapter 3.

4.2 Subject Classification

After extracting appropriate features from the FFRs as explained in Chapter 3, these features are next used as inputs to various machine learning models to see if it is possible to discriminate between FFR responses from different subjects with good accuracy.

In a previous study, Sun [25] used a variety of classification algorithms for listener identification on the same FFR database recorded by Heffernan [24] based on spectral and spectrogram features. The best accuracy reported was 84.09% obtained by an SVM model with radial basis function kernel. To check the validity of his results, Sun [25] also investigated the stability of the responses between the Test and Retest sessions for each subject.

Given the above and earlier reviewed attempts in using machine learning in subject classification with FFR data, there is still a need to increase their accuracy. Accordingly, by adopting the same database as Sun [25] and Heffernan [24], this chapter of the thesis focuses on investigating the possibility of increasing the classification accuracy of machine learning models which previously used spectrograms as input features. Also, gammatonegrams are tested as additional extracted features, with the goal of achieving better accuracies.

4.3 Classifier Method

A support vector machine (SVM) classifier method has been adopted in this study, since it has been the algorithm that provided the best performance among the ones used by Sun [25].The support vector machine classifier offers several advantages specially when it comes to training on small datasets It is particularly found to be effective in scenarios where the number of features is high compared to the number of samples, making it suitable for small datasets with complex and high-dimensional data similar to the dataset adopted for this thesis. Additionally, SVM

employs a regularization parameter that helps prevent overfitting, which can be crucial when working with limited training samples. It is also robust to outliers, as the model aim to maximize the margin between different classes, thereby reducing the influence of individual data points. This model is capable of handling both linearly separable and non-linearly separable data by using kernel functions, allowing for more flexible and accurate modeling even with limited training samples.

However, the SVM model has its own pitfalls as well. For instance, the training time and memory requirements can be significant, making the model less practical for extremely large datasets. Furthermore, SVMs are sensitive to the choice of hyperparameters, such as the kernel type and regularization parameter, and selecting the optimal values can be challenging. Considering the above-mentioned pros and cons, the SVM model can be a good selection given the size of the dataset used in this thesis and therefore has been adopted to perform the classification tasks. The algorithm is implemented using the scikit-learn python library.

It is worth mentioning that other classification methods, including random forest and XGBoost, were initially experimented with but not extensively utilized. This decision was made to concentrate on enhancing the results reported by Sun [25], potentially employing the same classifier method. Instead, the strategy was defined as enhancing classification performance through either increasing the spectrogram resolutions or methodically tuning the hyperparameters of the SVM model.

4.3.1 Support Vector Machine (SVM)

The SVM is a well-known classification method first proposed by Cortes and Vapnik [50]. The method was then extended as a non-linear classifier through using a variety of kernel functions to maximize the margin of its classification hyperplanes [51].

4.3.1.1 Binary Classification

Consider that there is a dataset consisting of d feature vectors as (X_1, y_1), (X_2, y_2), …, (X_n, y_n) where $X_i \in R^d$, y_i is a scalar and $y_i \in \pm 1$. Assuming that the data from these two classes are linearly separable, any hyperplane separating the classes can be written as:

$$X^T W + b = 0 \tag{4.1}$$

where W is the hyperplane normal vector and b is a bias term. Figure 4.1 shows a linearly separable two-class example, with samples on the margin called the support vectors.

Since the training data is linearly separable, two parallel hyperplanes separating the two labels can be drawn. The distance between the two hyperplanes should be as large as possible. The room (space) that is bounded by the two hyperplanes is defined as the margin, and the maximum-margin hyperplane is the one that is equidistant between them. When normalizing the dataset, the two hyperplanes can be described as

$$X^T W + b = 1 \tag{4.2}$$

and

$$X^T W + b = -1 \tag{4.3}.$$

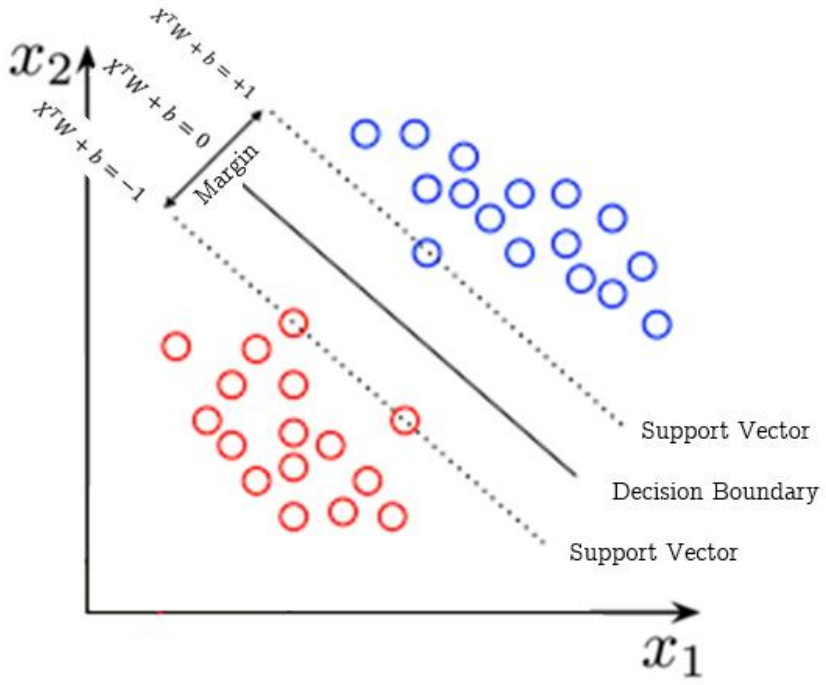

Figure 4.1 Two linearly separable classes, with samples on the margins called the support vectors.

Any data on and over the first boundary is labeled as 1 belonging to one class, while any data on or under the second boundary belongs to another class and is labeled as -1. The goal is to find the "maximum-margin hyperplane" to separate the group of points with label $y_i = 1$ from the group of points with label $y_i = -1$, so that the distance between the nearest point X_i from either group can be maximized.

Accordingly, the loss function that helps maximize the margin between the data points and the hyperplane is the hinge loss function defined as

$$C(X, y, X^T W + b) = \begin{cases} 0, & if\ y\ (X^T W + b) \geq 1 \\ 1 - y\ (X^T W + b), & else \end{cases} \quad (4.4)$$

The cost is 0 if the predicted value and the actual value are of the same sign. If they are not, a loss value is calculated. Also, a regularization parameter can be added to the cost function which makes the cost function be expressed as

$$C(X, y, X^T W + b) = min_W\left(\lambda \|W\|^2 + \sum_{i=1}^{n}\left(1 - y\ (x_i, W)\right)\right) \quad (4.5)$$

where λ is a regularization term. For the sake of easiness in use and interpretation, the inverse of this λ parameter is defined as C, which controls the regularization of the cost function:

$$C = \frac{1}{\lambda} \quad (4.6).$$

In other words, the C parameter adds a penalty for each misclassified data point. If C is small, the penalty for misclassified points is low, so a decision boundary with a large margin is chosen at the expense of a greater number of misclassifications. If C is large, SVM tries to minimize the number of misclassified examples due to high penalty, which results in a decision boundary with a smaller margin.

Finally, in a more generic case, if the data from two classes are not linear separable, the original data can be mapped into a higher dimension space where the mapped data is linearly separable. The mapping transformation is performed by a kernel function defined as:

$$K(x_i, x_j) = K(x_i^T x_j) = \phi(x_i)^T \phi(x_j) \quad (4.7)$$

where $\phi(x_i)$ is a mapping function. In this thesis, the linear, polynomial, and Gaussian radial basis kernel functions are used and defined as follows. For the linear kernel:

$$K(x_i, x_j) = x_i^T x_j \quad (4.8).$$

For the polynomial kernel function:

$$K(x_i, x_j) = (x_i^T x_j + b)^d \ , \ b \geq 0 \quad (4.9)$$

where *d* is the polynomial degree and *b* is a free parameter used to trade off the influence of higher-order versus lower-order terms in the polynomial.

Finally, for Gaussian radial basis kernel function, we have:

$$K(x_i, x_j) = exp\left(-\gamma \|x_i - x_j\|^2\right)^d, \quad \gamma \geq 0 \quad (4.10).$$

4.3.1.2 Multi-Class Classification

To extend the binary classification algorithm explained in Section 4.3.1.1 to a multi-classification algorithm, the problem is broken into smaller subproblems, all of which are binary classification problems. In multiclass prediction we would like to learn a function $h: X \to y$ without loss of generality where $y_i \in \{1, 2, \dots, k\}$.

The popular methods which are used to perform multi-classification on the problem statements using SVM algorithm are as follows:

- one vs. all (one vs. rest)
- one vs. one.

In the one vs. all also known as one vs. rest method, k binary classifiers are trained each of which discriminates between one class and the rest of the classes [52]. That is, given a training set vectors as $S = \{(X_1, y_1), (X_2, y_2), \dots, (X_n, y_n)\}$ k binary training sets $S_1, \dots, S_k$ are constructed,

where $S_i = \{(X_1, (-1)^{1[y_1 \neq i]}), (X_2, (-1)^{1[y_2 \neq i]}), \ldots, (X_n, (-1)^{1[y_n \neq i]})\}$. In words, S_i is the set of instances labeled 1 if their label in S was i, and -1 otherwise. For every $i \in [k]$ a binary predictor $h_i: X \to \{\pm 1\}$ is trained. Finally, given $h_1, \ldots, h_k$, a multiclass predictor is constructed using:

$$h(x) \in argmax_{i \in [k]} h_i(x) \quad (4.11).$$

Accordingly, a multiclass SVM classifier tries to find hyperplanes to separate the classes. This means that the separation takes all points into account and then divides them into two groups in which there is a group for the one class points and the other group for all other points. Figure 4.2 shows a linearly separable multiclass example with the hyperplanes obtained based on one vs. all method.

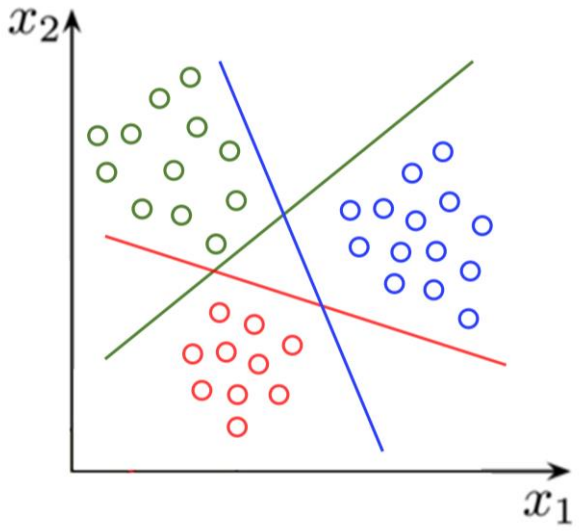

Figure 4.2 Hyperplanes for multi linearly separable classes using one vs. all method.

On the other hand, in the one vs. one approach all pairs of classes are compared to each other [52]. Formally, given a training set $S = \{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)\}$, where every y_i is in $[k]$, for every $1 \leq j \leq k$ a binary training sequence, $S_{i,j}$, is constructed containing all examples from S whose label is either i or j. For each such example, the binary label in $S_{i,j}$ is set to be $+1$ if the multiclass label in S is i and -1 if the multiclass label in S is j. Next, a binary classification algorithm is trained based on every $S_{i,j}$ to get $h_{i,j}$. Finally, a multiclass classifier is constructed by predicting the class that had the highest number of wins. An example of one vs. one binarization technique for decomposing the multi-class problem is shown in Figure 4.3.

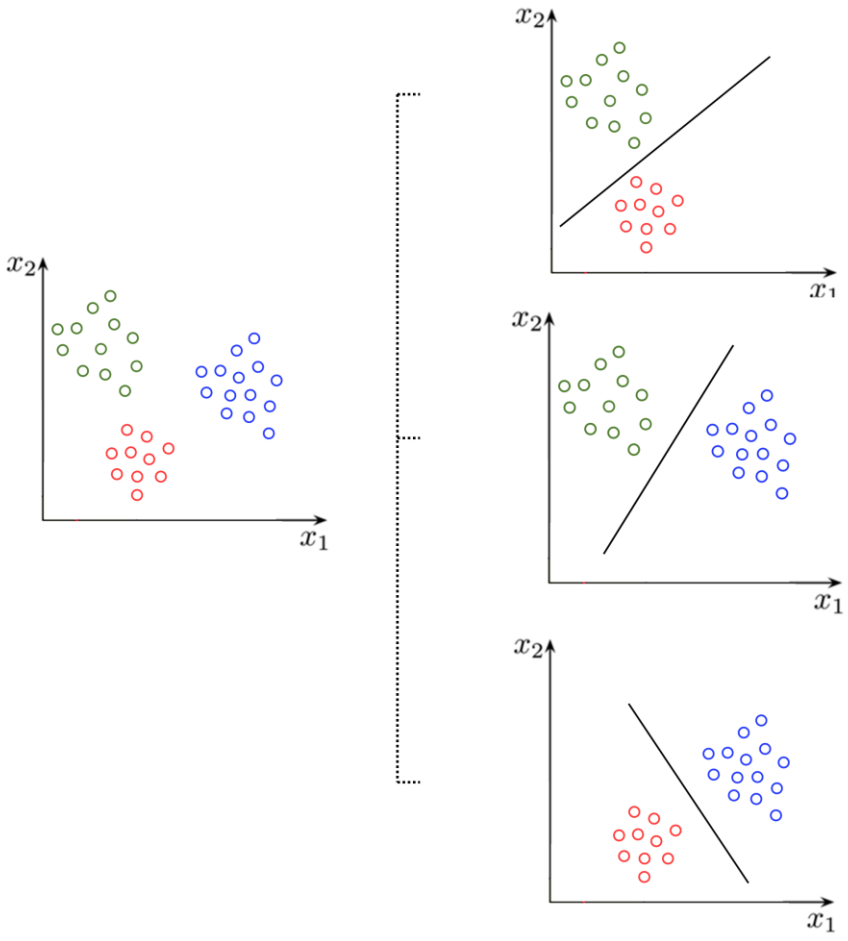

Figure 4.3 An example of one vs. one decomposition of a three-class problem into three two-class problems.

4.4 Classification Results

4.4.1 Classification with Spectrograms

The subject classification scenario is defined here as the correct classification of the listener (22 classes) based on the extracted features of the four-vowel concatenated FFRs in a specified sound level. Accordingly, the spectrograms and gammatonegrams with a variety of resolutions are inputted to SVM classifiers with Linear, Polynomial and Radial Basis Function (RBF) kernel functions to search for the best performance. As a first step and to benchmark the performance of the SVM classifier itself, it was first trained and tested on exactly the same dataset of spectrograms, and as expected, an accuracy score of 100% was achieved.

Figure 4.4 shows a spectrogram of a sample four-vowel concatenated eFFR record of this dataset at sound level of 85dB, which is used as an input data point to the SVM model for subject classification. As can be seen, the most informative and discriminative part of the spectrogram falls below the frequency of 1000 Hz. Therefore, to input the spectrograms to the SVM model, they are further cropped up to the frequency of 1300 Hz.

On this basis, the spectrogram of the four-vowel concatenated eFFRs at the sound level of 85dB are inputted into SVM classifiers with linear, polynomial and RBF kernels to classify for the 22 classes of listeners. To investigate the effect of the resolution of the spectrograms on the accuracy of classification, various window sizes and overlaps are tested (respectively defining the physical and computational frequency resolution of spectrograms). Also, to make sure the parameters of the machine learning model are selected to offer the best performance, hyperparameter tuning is performed. Being the most important hyperparameters of the SVM model, the C and γ parameters defined in equations (4.6) and (4.10) are varied in applicable ranges (grid search) and the resulting accuracy of the models are investigated.

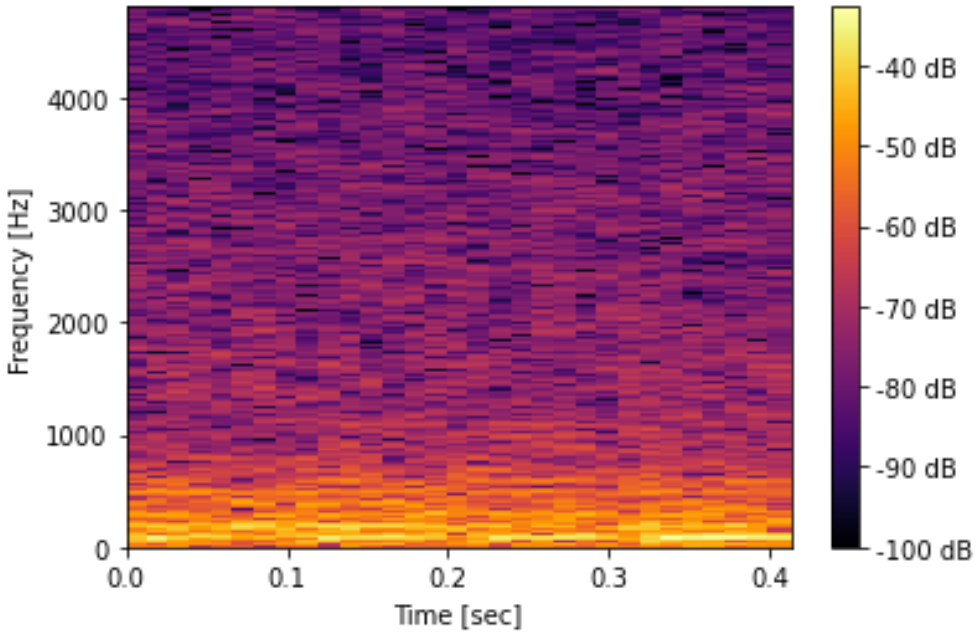

Figure 4.4 A sample spectrogram of a concatenated four-vowel eFFR at 85dB with window size of 256 points and overlap sizes of 128 points (50%).

Figure 4.5 shows an example of tuning the hyperparameters of C and γ for a dataset of spectrograms with a window size of 256 points and overlap size of 128 points (50%) for the SVM classifier with RBF kernel. The result of this parameter study suggests that for this specific case study, the optimized values of $C = 10.0$ and $\gamma = 1 \times 10^{-5}$ offer the best classification accuracy with the value of 86.4%.

To perform the classification, from the total 1408 records of the eFFR dataset, a concatenation of four vowels was performed and one of the four different sound levels (85dB) was selected, resulting a dataset of 88 spectrograms. In order to be able to perform a one-to-one comparison of the classification results with Sun [25], a similar train/test strategy was adopted. Accordingly, to train the SVM model, two different scenarios were considered based on two cases of the recordings (Test and Retest). In the first scenario, the classifier was trained on Test recordings and tested on the Retest recordings. For simplicity of referencing, this scenario is referred to as Test/Retest scenario in the tables of this chapter. In the second scenario, the classifier was trained on Retest recordings and tested on the Test recordings (Retest/Test scenario). In both cases, the

ratio of the train and test datasets is (0.5:0.5) and each of the sets includes 44 spectrograms of eFFRs.

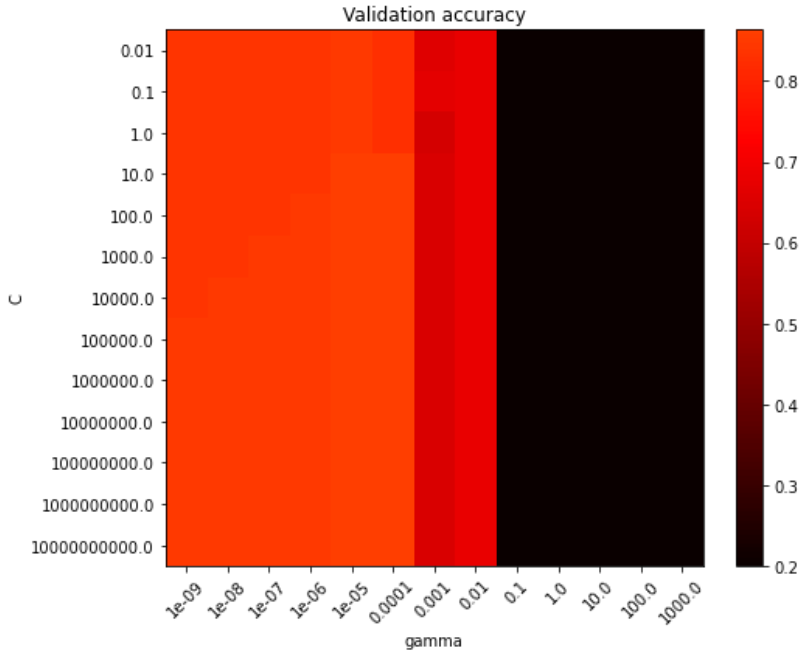

Figure 4.5 Hyperparameter tuning results of the SVM classifier with RBF kernel with spectrograms with window size of 256 points and overlap of 128 points (50%) as input.

Table 4.1 Subject classification accuracies performed by SVM model with linear kernel function using spectrograms with different resolutions

(Window, Overlap, Computational Resolution)	Test/Retest	Retest/Test
256, 8, 16 points	72.73%	72.27%
256, 64, 21 points	81.82%	81.82%
256, 128, 31 points	86.36%	81.82%
256, 250, 641 points	86.36%	86.36%
512, 511, 3585 points	90.91%	86.36%

Table 4.1 shows the accuracy of the subject classification performed for spectrograms with different resolutions using the SVM model with linear kernel function. The spectrograms are obtained for a sampling rate of 9606 Hz and an FFT length of 9606 points. It can be seen that

by increasing the resolution of the spectrograms, the accuracy of classification increases. Likewise, Tables 4.2 and 4.3 also show the classification results performed for spectrograms with different resolutions using the SVM model with polynomial and RBF kernel functions, respectively. The same trend of increase in accuracy with increasing the spectrogram resolutions is observed. It's noteworthy that the chance level classification accuracy for this 22-class scenario is $1/22$ or approximately 4.5%.

Table 4.3 Subject classification accuracies performed by SVM model with polynomial kernel function using spectrograms with different resolutions.

(Window, Overlap, Computational Resolution)	Test/Retest	Retest/Test
256, 8, 16 points	63.64%	72.37%
256, 64, 21 points	63.64%	68.18%
256, 128, 31 points	72.73%	79.54%
256, 250, 641 points	73.18%	84.09%
512, 511, 3585 points	75%	88.63%

Table 4.2 Subject classification accuracies performed by SVM model with RBF kernel function using spectrograms with different resolutions.

(Window, Overlap, Computational Resolution)	Test/Retest	Retest/Test
256, 8, 16 points	75%	77.27%
256, 64, 21 points	81.82%	81.82%
256, 128, 31 points	84.09%	84.09%
256, 250, 641 points	86.36%	88.64%
512, 511, 3585 points	93.18%	88.63%

Looking at the results presented in Tables 4.1 to 4.3 obtained by the three different kernel functions, it can be seen that the RBF kernel has provided the best accuracies among the three. The summary of the best results obtained by the RBF kernel function and highest resolution spectrograms with respective window size, overlap size and temporal resolution of 512, 511 and 3585 points is presented in Table 4.4. Consequently, the best accuracy reported here is 93.18% for the scenario of Test/Retest which is 9.09% higher than what reported as their best accuracy by Sun [25], which serves as a reference study.

Also, to compare comprehensively with all the analogous scenarios performed by (Sun, 2020), Table 4.5 collects and compares the results for all the three kernels for the only spectrogram resolution (256, 128, 31 points) performed in that study. It can be concluded that performing a methodical hyperparameter tuning plus increasing the resolution of the spectrograms as performed in this study have resulted in a 9.09% improvement in the best accuracy obtained. Such hyperparameter tuning had even more substantial effect on the results obtained by the polynomial kernel function. While Sun [25] reports 9.09% and 13.64% accuracies for Test/Retest and Retest/Test scenarios with polynomial kernel, the results obtained after hyperparameter tuning in this study are 72.73% and 79.54% for the same resolutions, respectively.

Table 4.4 Summary of the best subject classification accuracies performed by SVM model using spectrograms.

(Window, Overlap, Resolution)	Kernel function	Test/Retest	Retest/Test
512, 511, 3585 points	RBF	93.18%	88.63%

Table 4.5 Comparison of the subject classification accuracies performed by SVM model using spectrograms obtained in this study and Sun [26].

(Window, Overlap, Resolution)	Kernel function	Test/Retest	Retest/Test
256, 128, 31 points [Current Study]	Linear	86.36%	81.81%
256, 128, 31 points [Sun 2020]	Linear	70.45%	63.64%
256, 128, 31 points [Current Study]	Polynomial	72.73%	79.54%
256, 128, 31 points [Sun 2020]	Polynomial	9.09%	13.64%
256, 128, 31 points [Current Study]	RBF	84.09%	84.09%
256, 128, 31 points [Sun 2020]	RBF	84.09%	81.82%

4.4.2 Classification with Gammatonegrams

In this section, subject classification is performed using the other set of features extracted and explained in Chapter 3, the gammatonegrams. Accordingly, the gammatonegrams of the four-vowel concatenated eFFRs at the sound level of 85dB are inputted into SVM classifiers with linear, polynomial and RBF kernels to classify for the 22 classes of listeners. To achieve the best accuracy, these experiments are performed using the gammatonegrams with the highest computational frequency resolutions. Also, to make sure the parameters of the machine learning model are selected to offer the best performance, once again, hyperparameter tuning is performed. Again, the most important hyperparameters of the SVM model, the C and γ parameters defined in equations (4.6) and (4.10) are varied systematically in the applicable range and the resulting accuracy of the models is investigated. Figure 4.6 shows a gammatonegram of a sample four-vowel concatenated eFFR record at sound level of 85dB. The gammatonegram is extracted using the technique explained in Chapter 3 with a window size of 53 ms (half of a vowel length) and hopping size of 1 ms (overlap of 52ms). As can be seen, the most informative and discriminative part of the spectrogram falls below the frequency of 1000 Hz. Therefore, to input the spectrograms to the SVM model, they are further cropped up to the frequency of 1200 Hz.

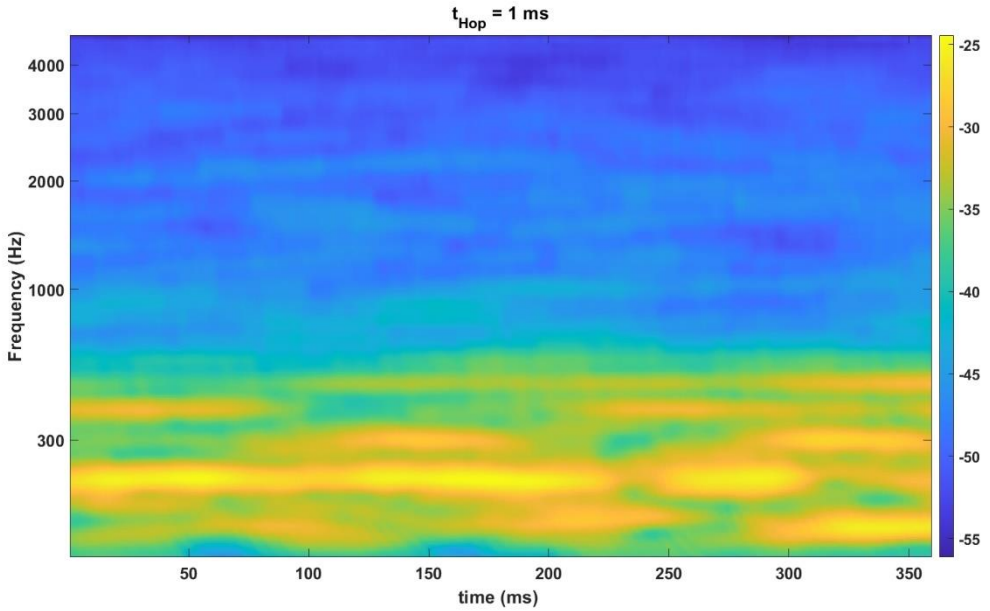

Figure 4.6 Gammatonegram of a sample concatenated four-vowel eFFR at sound level of 85dB with 256 filters, window size of 53 ms and t_{Hop} = 1 ms.

Figure 4.7 shows an example of tuning the hyperparameters of C and γ for a dataset of gammatonegrams with window size of 53 ms and hopping size of 1 ms for the SVM classifier with RBF kernel. The result of this parameter study suggests that for this specific case study, the optimized values of $C = 1000$ and $\gamma = 1 \times 10^{-7}$ offer the best classification accuracy with the value of 71.59%.

The experiments were performed on the same two scenarios that were done for the spectrograms, i.e., Test/Retest and Retest/Test scenario. Table 4.6 shows the accuracy of the subject classification performed for gammatonegrams using the SVM model with the three kernel

functions used. It was observed that among the three kernel functions, RBF provided the best results.

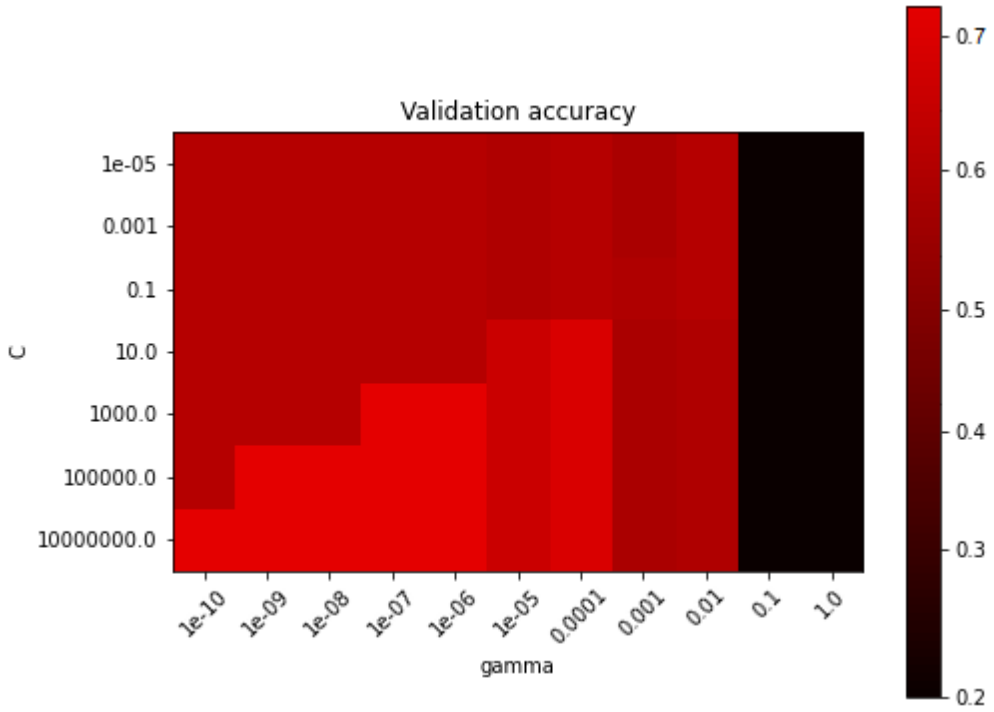

Figure 4.7 Hyperparameter tuning results of the SVM classifier with RBF kernel for gammatonegrams with window size of 53 ms and t_{Hop} = 1 ms as input.

Table 4.6 Subject classification accuracies performed by SVM model with different kernel functions using gammatonegrams.

Kernel Function	Test/Retest	Retest/Test
RBF	70.4%	67.05%
Linear	69.82%	67.65%
Polynomial	68.18%	65.4%

Comparing the results obtained for gammatonegrams and spectrograms, it can be seen that the maximum achieved classification accuracy was higher with spectrograms (93.18% compared to 71.59%). Here, comparing the physical frequency resolutions of the spectrograms and gammatonegrams could give a better understanding of the difference in the accuracies obtained. As introduced in the previous chapter, for the spectrogram the physical frequency resolution (in Hz) can be defined by $\frac{FWHM}{W_n} \times f_s$ where $FWHM$, W_n and f_s are the Full-Width-Half-Maximum measure (in Hz, bandwidth measured from the frequency response of a 1-sec. window), the window size (in samples) and the sampling frequency (in samples/sec.), respectively. Considering the FWHM of 1.207 for a Hamming window, the frequency resolution of the spectrograms with window size of 512 points and sampling frequency of 9606 Hz is obtained to be 34.05 Hz.

On the other side, the physical frequency resolution of a gammatonegram is determined by bandwidth of the gammatone filters, which depends on the number and spacing of the gammatone filters used in the filter bank. A higher frequency resolution in a gammatonegram means that the filters used in the filter bank have shorter bandwidths and are more closely spaced and thus can distinguish between more closely spaced frequency components in the signal. This can result in a more detailed representation of the signal in the frequency domain, particularly at lower frequencies. In a gammatone filter bank, the filters are spaced evenly on the ERB (Equivalent Rectangular Bandwidth) scale, which is a scale that approximates the frequency selectivity of the human auditory system (ERB) [53]. It is the bandwidth of a rectangular filter which approximates human hearing [54]. Glasberg and Moore [55] express the ERB as a function of centre frequency f_c of each filter as

$$ERB(f_c) = minBW + f_c/e_q \qquad (4.12)$$

where $minBW$ and e_q are the Glasberg and Moore parameters respectively set to 24.7 and 9.26449 derived from psychoacoustic experiments.

Based on equation 4.12, the bandwidth of the auditory filters in the gammatonegram at frequencies of 100, 200, 300, 400 and 500 Hz is obtained to be 35.49, 46.29, 57.08, 67.87 and 78.67 Hz, respectively. Comparing the 34.05 Hz physical frequency resolution of the

spectrograms with the higher values of the gammatonegram filters bandwidths may give a better understanding of why a higher accuracy was obtained with the spectrogram. This can also be observed by comparing the level of detail along the vertical (frequency) axis in Figure 3.6 (spectrogram with very good computational temporal resolution) and Figure 3.13 (gammatonegram with very good computational temporal resolution), where a higher level of detail is visible with the spectrogram.

Chapter 5 Listener Authentication

5.1 Overview

This chapter reports on the results obtained for authenticating one class of listeners against another. This scenario is defined within the framework known as the “sheep vs. wolves” case [56].

5.2 Biometric Authentication

5.2.1 Motivation

User authentication with biometrics is a method of verifying the identity of an individual based on their unique physical or behavioral characteristics. As discussed in detail in Chapter 1, biometric authentication can include various methods, such as fingerprint recognition, facial recognition, iris scans, speaker and listener recognition. Biometric authentication is becoming increasingly popular as it provides a highly secure and convenient method of authentication, as users do not need to remember passwords or carry physical tokens [57] [58].

Biometric authentication can be used in various applications, including physical access control, mobile device security, and financial transactions. For example, a company might use fingerprint recognition to allow employees to enter a secure area, or a hotel might use facial recognition to allow guests to enter their rooms.

Another application of user authentication with biometrics is in mobile device security. Biometric authentication can be used to unlock smartphones, tablets, and other mobile devices, providing a more secure and convenient alternative to traditional password or PIN authentication. It can also be used to secure mobile payments, allowing users to make transactions using their fingerprints or facial recognition. This is becoming an increasingly popular method of payment as it provides a faster and more secure alternative to traditional payment methods.

Finally, biometric authentication is also used in financial transactions, such as online banking and e-commerce. It can be used to verify the identity of users when making transactions, providing an additional layer of security beyond traditional authentication methods. This can

help prevent fraud and identity theft, as biometric authentication is more difficult to replicate or spoof than traditional authentication methods.

5.2.2 Listener Authentication with FFRs

Based on the discussion on the advantages and disadvantages of each of the discussed biometric features and the high demand for security and reliability in various applications, once again listener recognition using FFRs could address many of the shortfalls. Listener authentication can be classified as one vs. all, meaning that it is used to distinguish between one specific listener and all other listeners. The system compares the FFRs of the listener to a database of all authorized listeners and determines whether the listener's identity matches any of the identities in the database and decides based on the result.

5.3 Sheep vs. Wolves Scenario

In addition to the one vs. all authentication scenarios described, there is another user authentication scenario called sheep vs. wolves that has also been used in applications like cybersecurity [56]. In this scenario, the classification is based on the concept that there are two types of users - sheep who are legitimate users, and wolves who are attackers attempting to gain unauthorized access to sensitive information and resources. By analyzing the behavior and/or biometrics of a user, a sheep vs. wolves classification system can determine whether they are a genuine user or an attacker.

The sheep vs. wolves classification has been widely used in various security applications, including access control, intrusion detection, and authentication. In a study conducted by Chen et al. [59], a sheep vs. wolves classification system was proposed for detecting and blocking unauthorized access attempts to online services. The system used behavioral biometrics and machine learning techniques to classify users as either sheep or wolves based on their typing behavior, device usage, and location information.

Also, another study conducted by Zhang et al. [60] proposed a sheep vs. wolves classification system for intrusion detection in IoT networks. The system used anomaly detection and machine learning techniques to classify listeners as either sheep or wolves based on their behavior patterns

and network traffic. The results showed that the sheep vs. wolves classification system was highly effective in detecting and blocking unauthorized access attempts in IoT networks.

Overall, sheep vs. wolves classification is an effective method for protecting against cyber attacks by identifying and blocking unauthorized access attempts. By analyzing various factors and implementing multiple layers of security, organizations can ensure that only genuine users are granted access to sensitive information and resources, while attackers are blocked. This method helps to prevent data breaches, protect sensitive information, and maintain the security and integrity of systems and networks.

5.4 Case Study

Based on the motivations discussed above, the goal of this section of the thesis is to preliminarily investigate the use of eFFR for user authentication by performing user authentication with the adopted eFFR database using machine learning techniques. However, performing one vs. all classification scenario with the adopted dataset would result in a highly imbalanced distribution of trusted (one listener) vs. untrusted (twenty-one listeners) classes. This imbalance between the two classes cannot be addressed and resolved efficiently due to the limited number of records per listeners. Therefore, based on the reviewed usage of the sheep vs. wolves scenario in real-life authentication applications and since it can help to resolve the imbalance between the trusted and untrusted classes, a different scenario was selected for this study as outlined below.

Accordingly, two different classes of trusted (sheep) and untrusted (wolves) were formed among the listeners. For example, Figure 5.1 shows a possible distribution of different listeners in sheep and wolves classes. In this distribution, subjects 1, 2, …, 11 were considered as sheep while subjects 12, 13, …, 22 were considered as wolves class members.

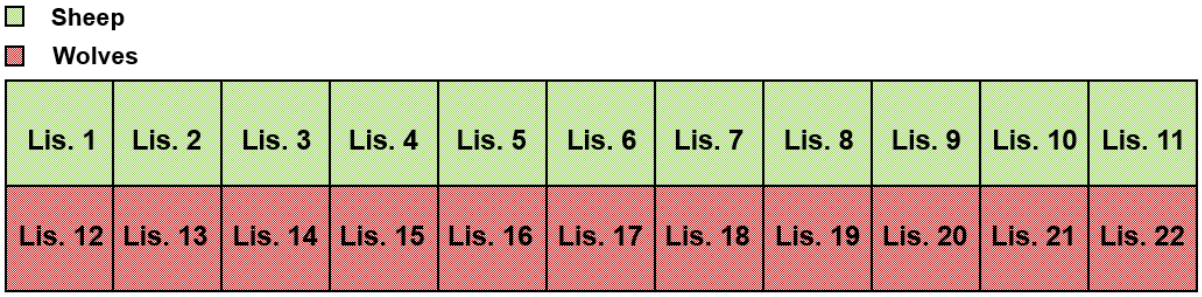

Figure 5.1 An example of distribution of listeners into sheep and wolves classes.

5.4.1 Implementation

Based on adopting the sheep vs. wolves scenario, the classification task here is defined as the correct classification of the groups of listeners (2 classes of sheep and wolves). The classification is once again performed based on the extracted features of the four-vowel concatenated FFRs at a specified sound level of 80 dB. Also, based on the previous good results obtained by the SVM classifier, this model is adopted once again to perform the classification task.

Accordingly, the spectrograms with a systematic variation in resolution are inputted to SVM classifiers with linear, polynomial and radial basis function (RBF) kernels to search for the best performance. As a first step and to benchmark the performance of the SVM classifier itself, it was first trained and tested on exactly the same dataset of spectrograms, and as expected, an accuracy score of 100% was achieved. The result of this benchmarking is shown in the confusion matrix shown in Figure 5.2. It can be seen that when training and testing on the same dataset, all the sheep and wolves have been classified correctly.

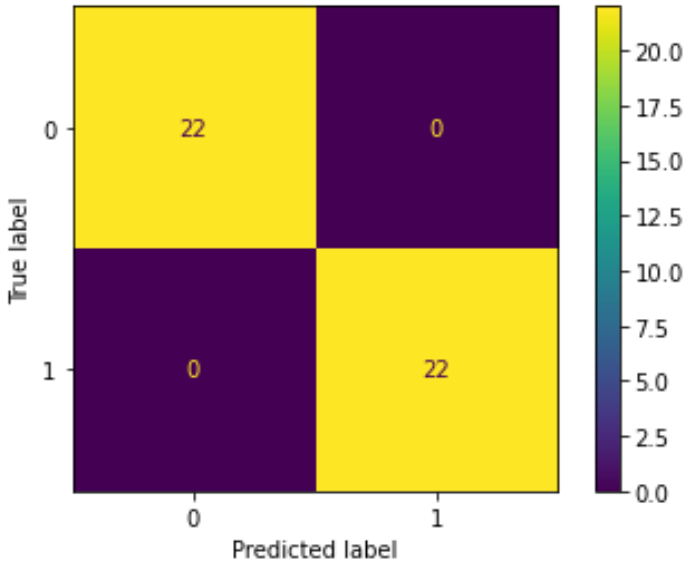

Figure 5.2 Confusion matrix obtained from training and testing the SVM model on the same dataset to benchmark the model.

Before performing the classification task, the hyperparameters of the SVM model are also tuned to guarantee the best performance possible on the dataset. Accordingly, while training the model, the C and γ values are varied systematically, and the validation accuracy obtained with each pair of parameters was recorded. Finally, the C and γ values offering the best validation accuracy were selected as the parameters to train the SVM model and perform the predictions. This grid search was performed on a wide range of values for C and γ parameters to ensure the best possible results. Figure 5.3 shows an example of this hyperparameter tuning result for the two-class classification scenario on spectrograms with a window size of 256 samples and an overlap of 128 samples.

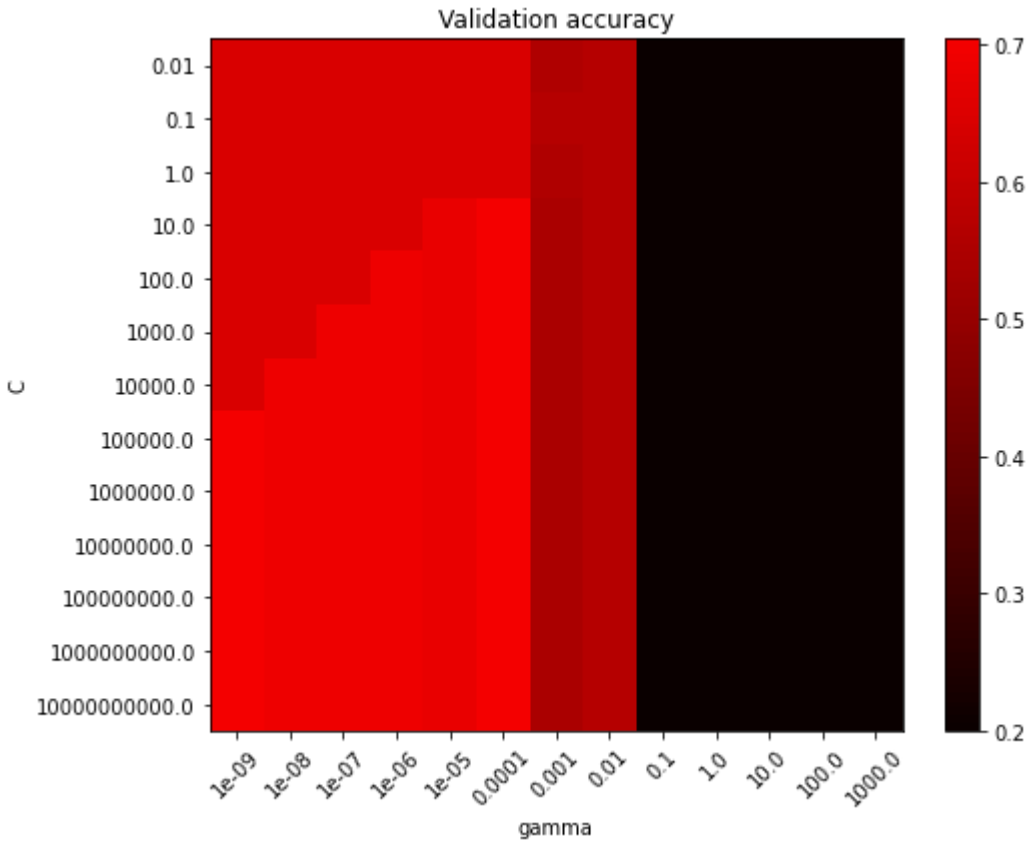

Figure 5.3 Hyperparameter tuning results of the SVM classifier with linear kernel for the two-class scenario using spectrograms with a window size of 256 samples and an overlap of 128 samples as input.

5.4.2 Results

In this study, 11 out of 22 subjects were selected to be members of the sheep class (trusted users) while the other 11 subjects were selected to be members of the wolves class (untrusted users). In order to make the classification accuracy independent of the selection of the subjects and their placement in each of the two classes, this experiment was repeated in five different cases. In each case, the selection of 11 trusted out of 22 subjects was varied randomly and the accuracy of classification was measured. Finally, the average of the classification accuracy for the five cases was adopted as the accuracy of authenticating a class of 11 subjects against the other 11.

Table 5.1 shows the composition of sheep and wolves classes in each of the five experiments performed.

Table 5.1 Composition of sheep and wolves classes in each of the five experiments performed.

	Sheep Class	Wolves Class
Case 1	[1,2,3,4,5,6,7,8,9,10,11]	[12,13,14,15,16,17,18,19,20,21,22]
Case 2	[1,3,5,7,9,11,13,15,17,19,21]	[2,4,6,8,10,12,14,16,18,20,22]
Case 3	[1,4,7,10,13,16,17,18,19,20,21]	[2,3,5,6,8,9,11,12,14,15,22]
Case 4	[1,2,3,4,5,6,18,19,20,21,22]	[7,8,9,10,11,12,13,14,15,16,17]
Case 5	[1,5,10,14,15,16,17,18,19,20,21]	[2,3,4,6,7,8,9,11,12,13,22]

The classification task was performed for each of the five cases shown in Table 5.1 on the spectrograms of the four-vowel concatenated FFRs. Accordingly, a total of 88 spectrograms (22 subjects × 2 Test/Retest records × 2 number of records) were divided into 50%-50% train-test splits, yielding the size of train and test sets to be 44. To train the SVM model, once again the two Test/Retest (train on Test recordings and test on the Retest recordings) and Retest/Test (train on Retest recordings and test on the Test recordings) scenarios were used. Classification for each of these two scenarios was performed using the three linear, polynomial and RBF kernels for different computational resolutions of spectrograms, to find the best possible accuracy.

For example, Figure 5.4 shows the results for classification of case 1 using spectrograms with a window size of 512 samples and overlap size of 511 samples, with RBF kernel for Retest/Test scenario. It can be seen that among the 22 negative and positive cases, 20 and 16 were correctly predicted as true negatives and true positives, respectively, yielding a classification accuracy of 81.82%. Also, only 2 cases were misclassified as a negative and 6 cases were misclassified to be positive, yielding a precision and recall of 0.89 and 0.73 for the positive class and 0.77 and 0.91 for the negative class, respectively.

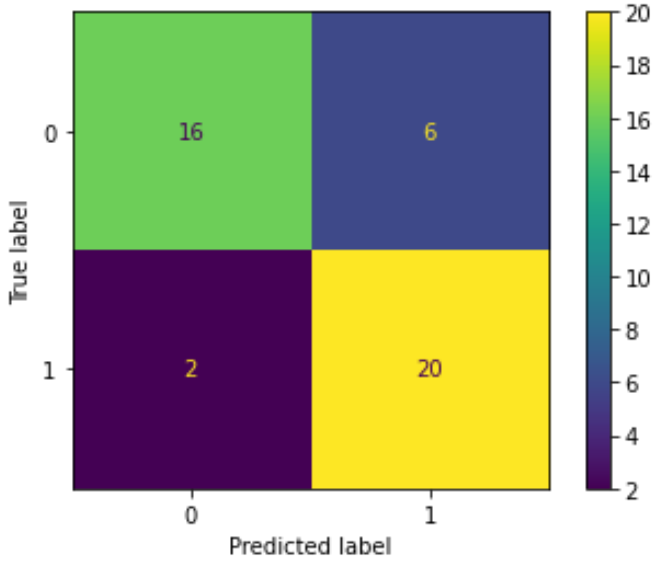

Figure 5.4 Confusion matrix for classification of sheep vs. wolves in a Retest/Test scenario with RBF kernel using spectrograms with window size of 512 samples and overlap size of 511 samples.

Tables 5.2 to 5.16 show the results of the classification accuracies for cases 1 to 5 using the three SVM kernels for different spectrogram computational resolutions. In this regard, Tables 5.2, 5.3 and 5.4 show the accuracies for classification of case 1 in which the sheep class composition included subjects [1,2,3,4,5,6,7,8,9,10,11] and the wolves class composition included [12,13,14,15,16,17,18,19,20,21,22] in the two scenarios of Test/Retest and Retest/Test. The results are shown for the RBF, linear and polynomial kernels, respectively. It can be seen that the best accuracy obtained for this composition of sheep and wolves is 84.09% which is for the Test/Retest scenario with the highest computational resolution of spectrograms (window size of 512 and overlap size of 511) and using linear kernel. Another good classification accuracy of 81.81% in this case was also recorded using the RBF kernel for the Retest/Test scenario and the same resolution of spectrograms. It is also noteworthy and in agreement with expectations that the accuracies increase with the computational resolution of spectrograms in the majority of experiments. Also, by comparing the accuracies obtained using RBF and linear kernels in Tables 5.2 and 5.3 with polynomial kernel in Table 5.4, it can be observed that the polynomial kernel had the poorest performance among the three. This is also in agreement with observations in Chapter 4.

Table 5.2 Classification accuracies for case 1 using SVM model with RBF kernel function and spectrograms with different resolutions.

(Window, Overlap, Computational Resolution)	Test/Retest	Retest/Test
256, 8, 16 samples	75%	75%
256, 128, 31 samples	70.45%	79.54%
256, 250, 641 samples	79.54%	77.27%
512, 511, 3585 samples	75%	81.81%

Table 5.3 Classification accuracies for case 1 using SVM model with linear kernel function and spectrograms with different resolutions.

(Window, Overlap, Computational Resolution)	Test/Retest	Retest/Test
256, 8, 16 samples	77.27%	72.72%
256, 128, 31 samples	72.72%	68.18%
256, 250, 641 samples	79.54%	70.45%
512, 511, 3585 samples	84.09%	75%

Table 5.4 Classification accuracies for case 1 using SVM model with polynomial kernel function and spectrograms with different resolutions.

(Window, Overlap, Computational Resolution)	Test/Retest	Retest/Test
256, 8, 16 samples	63.63%	59.09%
256, 128, 31 samples	63.63%	59.09%
256, 250, 641 samples	52.27%	59.09%
512, 511, 3585 samples	63.63%	52.27%

Table 5.5 Classification accuracies for case 2 using SVM model with RBF kernel function and spectrograms with different resolutions.

(Window, Overlap, Computational Resolution)	Test/Retest	Retest/Test
256, 8, 16 samples	77.27%	81.82%
256, 128, 31 samples	77.27%	84.09%
256, 250, 641 samples	79.54%	88.63%
512, 511, 3585 samples	88.63%	90.91%

Classification results for case 2 in which the sheep class composition included subjects [1,3,5,7,9,11,13,15,17,19,21] and wolves class composition included subjects [2,4,6,8,10,12,14,16,18,20,22] are shown in Tables 5.5, 5.6 and 5.7. The best accuracy for this case was obtained using the RBF kernel (90.91%) which was for the Retest/Test scenario of the highest computational resolution spectrogram (window size of 512 samples and overlap size of 511 samples). Also, it was observed that using RBF kernel provided the best accuracies among the three while the polynomial kernel still did not seem to be the appropriate choice of kernel for this classification task.

Table 5.6 Classification accuracies for case 2 using SVM model with linear function and spectrograms with different resolutions.

(Window, Overlap, Computational Resolution)	Test/Retest	Retest/Test
256, 8, 16 samples	70.45%	63.63%
256, 128, 31 samples	70.45%	65.91%
256, 250, 641 samples	70.45%	68.18%
512, 511, 3585 samples	68.18%	75%

Table 5.7 Classification accuracies for case 2 using SVM model with polynomial kernel function and spectrograms with different resolutions.

(Window, Overlap, Computational Resolution)	Test/Retest	Retest/Test
256, 8, 16 samples	59.09%	63.63%
256, 128, 31 samples	72.72%	59.09%
256, 250, 641 samples	56.81%	63.63%
512, 511, 3585 samples	72.72%	65.91%

Tables 5.8, 5.9 and 5.10 report the results of classification accuracies in case 3 with sheep class composition of [1,4,7,10,13,16,17,18,19,20,21] and wolves class composition of [2,3,5,6,8,9,11,12,14,15,22]. Once again, the best accuracy of 88.63% was obtained using the RBF kernel with highest resolution. However, linear kernel was also capable of offering a nearly as good result (86.36%).

Table 5.8 Classification accuracies for case 3 using SVM model with RBF kernel function and spectrograms with different resolutions.

(Window, Overlap, Computational Resolution)	Test/Retest	Retest/Test
256, 8, 16 samples	79.54%	81.82%
256, 128, 31 samples	81.82%	77.27%
256, 250, 641 samples	81.82%	81.82%
512, 511, 3585 samples	88.63%	81.82%

Table 5.9 Classification accuracies for case 3 using SVM model with linear function and spectrograms with different resolutions.

(Window, Overlap, Computational Resolution)	Test/Retest	Retest/Test
256, 8, 16 samples	79.54%	79.54%
256, 128, 31 samples	81.82%	86.36%
256, 250, 641 samples	81.82%	84.09%
512, 511, 3585 samples	81.82%	86.36%

Table 5.10 Classification accuracies for case 3 using SVM model with polynomial kernel function and spectrograms with different resolutions.

(Window, Overlap, Computational Resolution)	Test/Retest	Retest/Test
256, 8, 16 samples	59.09%	61.36%
256, 128, 31 samples	52.27%	63.63%
256, 250, 641 samples	52.27%	77.27%
512, 511, 3585 samples	52.27%	77.27%

The classification results for case 4 with the sheep and wolves composition of [1,2,3,4,5,6,18,19,20,21,22] and [7,8,9,10,11,12,13,14,15,16,17], respectively, are shown in Tables 5.11, 5.12 and 5.13. Here, an accuracy of 90.91% was the best result achieved with the RBF kernel for the Test/Retest scenario. As expected, the best accuracy once again occurred with the highest resolution spectrograms (window size of 512 samples an overlap size of 511 samples).

Table 5.11 Classification accuracies for case 4 using SVM model with RBF kernel function and spectrograms with different resolutions.

(Window, Overlap, Computational Resolution)	Test/Retest	Retest/Test
256, 8, 16 samples	86.36%	84.09%
256, 128, 31 samples	72.72%	72.72%
256, 250, 641 samples	88.63%	84.09%
512, 511, 3585 samples	90.91%	81.82%

Table 5.12 Classification accuracies for case 4 using SVM model with linear function and spectrograms with different resolutions.

(Window, Overlap, Computational Resolution)	Test/Retest	Retest/Test
256, 8, 16 samples	77.27%	70.45%
256, 128, 31 samples	72.72%	72.72%
256, 250, 641 samples	72.72%	77.27%
512, 511, 3585 samples	79.54%	70.45%

Table 5.13 Classification accuracies for case 4 using SVM model with polynomial kernel function and spectrograms with different resolutions.

(Window, Overlap, Computational Resolution)	Test/Retest	Retest/Test
256, 8, 16 samples	52.27%	61.36%
256, 128, 31 samples	70.45%	65.91%
256, 250, 641 samples	70.45%	72.72%
512, 511, 3585 samples	70.45%	70.45%

Finally, Tables 5.14, 5.15 and 5.16 report the results for case 5 with the sheep and wolves classes composition of [1,5,10,14,15,16,17,18,19,20,21] and [2,3,4,6,7,8,9,11,12,13,22], respectively. The best results here are once again obtained with the RBF kernel as 88.63% and 84.09% for the Test/Retest and Retest/Test scenarios with the highest spectrogram resolutions.

Table 5.14 Classification accuracies for case 5 using SVM model with RBF kernel function and spectrograms with different resolutions.

(Window, Overlap, Computational Resolution)	Test/Retest	Retest/Test
256, 8, 16 samples	84.09%	79.54%
256, 128, 31 samples	77.27%	77.27%
256, 250, 641 samples	79.54%	88.63%
512, 511, 3585 samples	88.63%	84.09%

Table 5.15 Classification accuracies for case 5 using SVM model with linear function and spectrograms with different resolutions.

(Window, Overlap, Computational Resolution)	Test/Retest	Retest/Test
256, 8, 16 samples	65.91%	77.27%
256, 128, 31 samples	65.91%	70.45%
256, 250, 641 samples	70.45%	72.72%
512, 511, 3585 samples	72.72%	81.82%

Table 5.16 Classification accuracies for case 5 using SVM model with polynomial kernel function and spectrograms with different resolutions.

(Window, Overlap, Computational Resolution)	Test/Retest	Retest/Test
256, 8, 16 samples	59.09%	61.36%
256, 128, 31 samples	77.27%	68.18%
256, 250, 641 samples	77.27%	75%
512, 511, 3585 samples	77.27%	77.27%

The results presented in Tables 5.2 to 5.16 show that the classification accuracies could be either composition or scenario (Test/Retest or Retest/Test) dependent and change even up to 10% among scenarios (since the data was recorded at different recording sessions) or different compositions. Accordingly, as discussed earlier, to report classification accuracies that are less class composition dependent, in the next step of the analysis, the values obtained for experiments with the five different cases are averaged. Tables 5.17, 5.18 and 5.19 report the minimum, maximum and average of the obtained accuracies at different spectrogram resolutions for each kernel. It can be seen that the RBF kernel offers the best average accuracy (86.36%) in a

Test/Retest scenario using the high-resolution spectrograms. Also, the increasing trend of the average accuracies with the increase in spectrogram resolution is observed in the majority of experiments. Comparing the results presented in Tables 5.17, 5.18 and 5.19 also shows that using RBF kernel has resulted in least minimum to maximum variation of accuracies when repeating the experiment for the five cases. Considering the limitations of the adopted dataset, the maximum average accuracy of 86.36% obtained with these experiments can be deemed as a promising performance for subject authentication using FFRs. This implies that in specific real applications where ample amount of data exists to train the model, better performance may also be achievable.

Table 5.17 Minimum, maximum and average classification accuracies for the 5 cases using SVM model with RBF kernel function and spectrograms with different resolutions.

(Window, Overlap, Computational Resolution)	Test/Retest (min-Max), Average	Retest/Test (min-Max), Average
256, 8, 16 samples	(75% - 86.36%), 80.45%	(75% - 84.09%), 80.45%
256, 128, 31 samples	(70.45% - 82.72%), 75.91%	(72.72% - 84.09%), 78.18%
256, 250, 641 samples	(79.54% - 88.63%), 81.81%	(77.27% - 88.63%), 84.09%
512, 511, 3585 samples	(75% - 90.91%), 86.36%	(81.81% - 90.91%), 84.09%

Table 5.18 Minimum, maximum and average classification accuracies for the 5 cases using SVM model with linear kernel function and spectrograms with different resolutions.

(Window, Overlap, Computational Resolution)	Test/Retest (min-Max), Average	Retest/Test (min-Max), Average
256, 8, 16 samples	(65.91% - 79.54%), 74.09%	(63.63% - 79.54%), 72.72%
256, 128, 31 samples	(65.91% - 81.82), 72.72%	(65.91% - 86.36%), 72.72%
256, 250, 641 samples	(70.45% - 81.82), 75.00%	(68.18% - 84.09%), 74.54%
512, 511, 3585 samples	(68.18% - 84.09%), 77.27%	(75% - 86.36%), 77.72%

Table 5.19 Minimum, maximum and average classification accuracies for the 5 cases using SVM model with polynomial kernel function and spectrograms with different resolutions.

(Window, Overlap, Computational Resolution)	Test/Retest (min-Max), Average	Retest/Test (min-Max), Average
256, 8, 16 samples	(52.27% - 63.63%), 58.73%	(59.09% - 63.63%), 61.36%
256, 128, 31 samples	(52.27% - 77.27%), 67.27%	(59.09% - 68.18%), 63.18%
256, 250, 641 samples	(52.27% - 77.27%), 61.81%	(59.09% - 77.27%), 69.54%
512, 511, 3585 samples	(52.27% - 77.27%), 67.27%	(52.27% - 77.27%), 68.63%

Chapter 6 Conclusions

6.1 Overview

This chapter discusses the major findings of the thesis in comparison to previous studies, the limitations of the work and recommendations for future work.

6.2 Major Findings

The major findings of this thesis can be discussed in two areas:

1. Application of FFRs in listener identification
2. Application of FFRs in listener authentication

In both areas, the results of this study provide a better understanding of the accuracy with which subjects could be automatically identified or authenticated using frequency following responses. Such a deeper understanding helps to evaluate the potential of performing real-life subject identification or authentication systems using these biometric measurements.

Accordingly, the findings of this thesis follow three proposed general questions:

1. Can discrimination between FFRs from different subjects be performed with better accuracies compared to what has been reported in past similar studies?
2. Can new features (not used in past similar studies) be extracted from FFRs and fed as inputs to machine learning algorithms to improve the accuracy of listener identification?
3. How accurately can the subject authentication task be performed by applying machine learning techniques on FFRs?

To answer the first question, we refer to a very recent study in the field performed by Sun [25], [26] that employed a variety of classification algorithms for listener identification on an FFR database recorded by Heffernan [24]. The best accuracy reported was 84.09% obtained by an SVM model with RBF kernel. In order to find if this reported result could be further improved, the same dataset, analogous scenarios (Test/Retest and Retest/Test) and the same input features (spectrograms) as in [25], [26] were adopted for this study. Also, the SVM classifier was selected as the machine learning model of this study, being the algorithm that provided the best result in in [25], [26]. By performing a systematic hyperparameter tuning for the SVM model through a

grid search loop and performing the classification task on different spectrogram computational resolutions, better accuracies were obtained.

The hyperparameter tuning helped to increase the accuracies in almost all computational resolutions that have been evaluated in [25], [26]. This improvement of accuracies was discussed in detail in Table 4.5 of this thesis. For example, using spectrograms with a window size of 256 samples and an overlap size of 128 samples (the highest computational resolution attempted in in [25], [26]), the accuracy using RBF kernel in a Retest/Test scenario was 81.82% in [25], [26]. In this study, the hyperparameter tuning of the SVM model increased the accuracy of the same scenario to 84.09%. The effect of hyperparameter tuning was fundamental when using polynomial kernel, where the accuracies of 9.09% and 13.64% reported in [25], [26] were increased to 72.73% and 79.54% for a Test/Retest and Retest/Test scenario, respectively.

Finally, attempting higher spectrogram resolutions in this thesis helped in increasing the best accuracy of classification to 93.18%, which is a 9.09% improvement compared to [25], [26].

To answer the second question, gammatonegrams were selected as another FFR feature to be used as the input of machine learning model for subject identification. However, the classification accuracies obtained using gammatonegrams were substantially lower compared to the case where spectrograms were used (71.59% compared to 93.18% in the best case scenario). This would suggest exploring other features to search for a possible performance improvement of the classification using FFRs.

To answer the third question, the sheep vs. wolves scenario was adopted to classify the listeners into two groups of trusted and untrusted subjects. In this case for a Test/Retest scenario the best average accuracy of 86.36% was obtained again using the SVM classifier with RBF kernel on high computational resolution spectrograms. However, due to the limitations of the adopted dataset size, implementation of other authentication scenarios such as one vs. all was not possible, but still the accuracies obtained for sheep vs. wolves experiments was promising. This can also shed more light on the future application of machine learning models for subject authentication tasks using evoked FFRs.

6.3 Limitations and Future Work

Some of the limitations of the current study are related to the adopted dataset. For instance, the dataset is limited to healthy adults with normal hearing and may not practically generalize to other populations. Accordingly, the study did not consider other physical or sociocultural factors, such as age, hearing loss, and multilingualism, which have been examined in other studies related to FFR. If a more diverse population group with factors like subjects with normal hearing and subjects with hearing-loss, different age groups and native and non-native language speakers were included in the 22 subjects of the adopted dataset, the higher diversity of the features would possibly allow discrimination between the subjects more easily, which would help to improve the classifier's performance.

Another limitation was the size of the adopted dataset. Further expansion of the dataset in the future would be helpful to improve the performance and confidence in the machine learning models. Furthermore, a larger dataset would allow more flexibility in partitioning the data into training and testing sets, which was one of the main reasons that the one vs. all authentication scenario was not attempted in this thesis.

There are also some other limitations related to the features used as the input of the machine learning models in this study that can be addressed in future studies. For instance, this study only considered the amplitude of the spectra of the evoked responses. However, there are other aspects such as the phase of the FFR that could be used as additional features to improve the classification performance.

Finally, the current study only focused on the SVM model as the classifier algorithm. However, there are a variety of other machine learning based classifiers that can be used for the task. Specifically, increasing the size of the dataset would make it possible to use deep learning algorithms that need more data than what was available in this study.

References

[1] J. Vacca, Biometric Technologies and Verification Systems, 1st ed., Burlington, MA: Elsevier Science, 2007.

[2] F. L. Podio and J. S. Dunn, "Biometric Authentication Technology: From the Movies to Your Desktop," *ITL Bulletin*, pp. 1-8, May 2001.

[3] L. Hong and A. K. Jain, "Integrating faces and fingerprints for personal identification," *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* vol. 1351(12), pp. 16-23, 1997.

[4] A. K. Jain, L. Hong, S. Pankanti and R. Bolle, "An identity-authentication system using fingerprints," *Proceedings of the IEEE,* vol. 85(9), pp. 1365-1388, 1997.

[5] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother and A. Mah, "Pushing the frontiers of unconstrained face detection and recognition," *IARPA Janus Benchmark A. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* vol. 1351(12), pp. 1931-1939, 2015.

[6] Y. Taigman, M. Yang, M. Ranzato and L. Wolf, "DeepFace: Closing the gap to humanlevel performance in face verification," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* vol. 1351(12), p. 1701–1708, 2014.

[7] Y. Chen, S. C. Dass and A. Jain, "Fingerprint Quality Indices for Predicting Authentication Performance," in *Audio- and Video-Based Biometric Person Authentication*, Berlin, Germany, pringer Berlin Heidelberg, 2005, pp. 160-170.

[8] M. Leghari, S. Memon, . L. D. Dhomeja and A. Jalbani, "Deep Feature Fusion of Fingerprint and Online Signature for Multimodal Biometrics," *Computers,* vol. 10(2), no. 21, 2021.

[9] K. Meena and N. Malarvizhi, "An Efficient Human Identification through MultiModal Biometric System.," *Braz. Arch. Biol. Technol.,* vol. 59, 2016.

[10] K. Kamal, P. Tiwari and . A. K. Singh, "Fingerprint recognition: A review," *Journal of Forensic Research,* vol. 7(5), pp. 1-9, 2016.

[11] S. Kamkar, "The quest for the master key: Extracting a secret AES-128 encryption key from a fingerprint sensor," in *DEF CON 24 Hacking Conference*, Las Vegas, 2016.

[12] Y. Dong, "DeepMasterPrints: Generating MasterPrints for Dictionary Attacks via Deep Learning," *Proceedings of the 2016 ACM on Asia Conference on Computer and Communications Security,* pp. 3-14, 2016.

[13] C. Shan, "Vulnerabilities of facial recognition technology and strategies for protection," *Journal of Computer Science,* vol. 15(10), pp. 856-868, 2019.

[14] A. Athalye, L. Engstrom, A. Ilyas and K.-W. Kwok, "Synthesizing Robust Adversarial Examples," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 1765-1774, 2018.

[15] S. A. Elvy, "Commodifying Consumer Data in the Era of the Internet of Things," *UC Irvine Law Reviews,* vol. 423, pp. 435-37, 2018.

[16] M. B. Kugler, "From Identification to Identity Theft: Public Perceptions of Biometric Privacy Harms," *UC Irvine Law Reviews,* vol. 10(1), pp. 107-152, 2019.

[17] D. Harwel, "Microsoft Calls for Regulation of Facial Recognition, Saying It's Too Risky to Leave to Tech Industry Alone," *Washingtonpost. com,* 13 July 2018.

[18] S. Grimes and E. Shinabarger, "Biometric Privacy Iitigation: The Next Class Action Battleground," *BIOOMBERG LAW: BIG LAW BUs.,* 2018.

[19] J. Valentino-DeVries, "Facebook Is Accused of Spying on Users Who Aren't on the Site," *The New York Times,* 29 April 2015.

[20] J. Johnson, "The Security of Iris Recognition Databases: A Critical Review," *Journal of Computer Science,* vol. 14(5), pp. 724-732, 2018.

[21] Y. Wu, "On the Security of Iris Recognition Systems," *Proceedings of the ACM Conference on Computer and Communications Security,* pp. 1173-1182, 2017.

[22] J. Smith, "Speaker Recognition: A Critical Review of Vulnerabilities and Limitations," *Journal of Computer Science,* vol. 15(7), pp. 587-596, 2019.

[23] T. Nguyen, "On the Vulnerabilities of Speaker Recognition Systems," *Proceedings of the ACM Conference on Computer and Communications Security,* pp. 1183-1192, 2016.

[24] B. Heffernan, "Characterization and Classification of the Frequency Following Response to Vowels at Different Sound Levels in Normal Hearing Adults," *PhD. thesis, University of Ottawa,* 2019.

[25] R. Sun, "Classification of Frequency Following Responses to English Vowels in a Biometric Application," *MSc. thesis, University of Ottawa,* 2020.

[26] R. Sun, M. Bouchard and H. R. Dajani, "Biometric Classification of Frequency Following Responses to English Vowels," in *IEEE International Instrumentation and Measurement Technology Conference (I2MTC),* Ottawa, ON, Canada, 2022.

[27] G. M. Bidelman and L. Powers, "Response properties of the human frequency following response (FFR) to speech and non-speech sounds: level dependence, adaptation and phase-locking limits," *International Journal of Audiology,* vol. 57(9), p. 665–672, 2018.

[28] B. Chandrasekaran and . N. Kraus, "The scalp-recorded brainstem response to speech: Neural origins and plasticity," *Psychophysiology",,* vol. 47(2), p. 236–246, 2010.

[29] H. G. Yi, Z. Xie, R. Reetzke, A. G. Dimakis and B. Chandrasekaran, "Vowel decoding from single-trial speech-evoked electrophysiological responses: A feature-based machine learning approach," *Brain and Behavior,* vol. 7(6), p. 1–8, 2017.

[30] T. W. Picton, Human Auditory Evoked Potentials, Plural Publishing INC., 2010.

[31] S. J. Aiken and T. W. Picton, "Envelope and spectral frequency-following responses to vowel sounds," *Hearing Research,* vol. 245(1–2), p. 35–47, 2008.

[32] E. Skoe and N. Kraus, "Auditory brainstem response to complex sounds: a tutorial Erika," *Ear Hear,* vol. 31(3), p. 302–324, 2010.

[33] W. F. Dolphin and D. C. Mountain, "The envelope following response: Scalp potentials elicited in the mongolian gerbil using sinusoidally AM acoustic signals," *Hearing Research,* vol. 58(1), pp. 70-78, 1992.

[34] F. Huis, P. Osterhammel and K. Terkildsen, "The frequency selectivity of the 500 hz frequency following response," *Scandinavian Audiology,* vol. 6(1), p. 35–42, 1977.

[35] T. C. Chimento, "Selectively eliminating cochlear microphonic contamination from the frequency-following response," *Electroencephalograph and Clinical Neurophysiology,* p. 88–96, 1990 .

[36] A. Krishnan, "Human frequency-following responses to two-tone approximations of steady-state vowels," *Audiology and Neuro-Otology,* vol. 4(2), p. 95–103, 1999.

[37] A. Krishnan, "Human frequency-following responses: Representation of steady-state synthetic vowels," *Hearing Research,* vol. 166(1–2), p. 192–201, 2002.

[38] . A. Yellamsetty and G. M. Bidelman, "Brainstem correlates of concurrent speech identification in adverse listening conditions," *Brain Research,* vol. 1714, p. 182–192, 2019.

[39] . I. Akhoun, S. Gallégo, A. Moulin and M. V. Ménard, "The temporal relationship between speech auditory brainstem responses and the acoustic pattern of the phoneme /ba/ in normal-hearing adults," *Clinical Neurophysiology,* vol. 119(4), p. 922–933, 2008.

[40] C. M. Bishop, "Pattern recognition and machine learning," vol. (Vol. 1), Springer, 2006.

[41] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research,* vol. 3, pp. 1157-1182, 2003.

[42] T. Mitchell, "Machine Learning," McGraw-Hill Education, 1997.

[43] A. V. Oppenheim, "Speech spectrograms using the fast Fourier transform," *IEEE spectrum,* vol. 8 (7), p. 57–62, 1970.

[44] M. Mulimani and S. G. Koolagudi, "Acoustic Event Classification Using Spectrogram," in *TENCON 2018 - 2018 IEEE Region 10 Conference*, Jeju, Korea, 2018.

[45] P. Virtanen, R. Gommers, T. Oliphant, , M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, B. Weckesser, J. Brigh, S. van der Walt, M. Brett,, J. Wilson, and K. J. Millman, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods,* vol. 17, pp. 261-272, 2020.

[46] J. B. Allen and L. R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proc. IEEE,* vol. 65, pp. 1558-1564, 1977.

[47] D. O'Shaughnessy, Speech Communication, Reading, MA: Addison-Wesley, , 1987.

[48] R. D. Patterson, "Auditory filter shapes," *Journal of the Acoustical Society of America,* vol. 55, p. 802–809, 1974.

[49] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang and M. H. Allerhand, "Complex sounds and auditory images," *Auditory Physiology and Perception, (Eds.),* pp. 429-446, 1992.

[50] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20(3), p. 273–279, 1995.

[51] B. E. Boser, . I. M. Guyon and V. N. Vapnik, "Training algorithm for optimal margin classifiers.," *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, p. 144–152, 2015.

[52] S. Shalev-Shwartz and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press., 2014.

[53] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *Journal of the Acoustical Society of America* , vol. 74.3, p. 750–753, 1983.

[54] J. Benesty, Springer handbook of speech processing, Springer Science &Business Media, 2008.

[55] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing research,* vol. 47.1, pp. 103-138, 1990.

[56] J. P. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE,* vol. 85(9), pp. 1437-1462, 1997.

[57] D. Zhang, Biometric Solutions, Springer US, 2012.

[58] A. Razaque, USER BIOMETRICS AUTHENTICATION (Comprehensive Analysis), Lulu.com, 2019.

[59] T. Chen, W. Liu, Y. Sun and Y. Yang, "A sheep versus wolves classification method based on biometrics behavior for blocking unauthorized access to online services," *IEEE Access,* vol. 7, pp. 5773-5783, 2019.

[60] Y. Zhang, W. Peng, P. Jiang and J. Zhang, ". (2020). A sheep versus wolves classification system based on machine learning for intrusion detection in IoT networks," *Journal of Ambient Intelligence and Humanized Computing,* vol. 11(2), pp. 537-548.

[61] A. Sadeghian, H. R. Dajani and A. D. C. Chan, "Classification of speech-evoked brainstem responses to English vowels," *Speech Communication,* vol. 68, p. 69–84, 2015.

[62] A. Sadeghian, H. R. Dajani and A. D. C. Chan, "Classification of English vowels using speech evoked potentials," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society,* p. 5000–5003, 2011.

[63] F. Llanos, Z. Xie and B. Chandrasekaran, "Biometric identification of listener identity from frequency following responses to speech," *Journal of Neural Engineering*, vol. 16(5), 2019.

[64] M. Laroche, H. R. Dajani, F. Prévost and A. M. Marcoux, "Brainstem auditory responses to resolved and unresolved harmonics of a synthetic vowel in quiet and noise," *Ear and Hearing*, vol. 34(1), p. 63–74, 2013.

[65] I. Kavati and M. V. N. K. Prasad, Search Space Reduction in Biometric Databases: A Review.Computer Vision: Concepts, Methodologies, Tools, and Applications, Chicago, Illinois: Information Resources Management Association, IGI Global, 2018.

[66] S. Greenberg, J. T. Marsh and W. S. Brown, "Neural temporal coding of low pitch. I. Human frequency-following responses to complex tones," *Hearing Research,* vol. 25(2–3), p. 91–114, 1987.

[67] G. M. Bidelman, "Sonification of scalp-recorded frequency-following responses (FFRs) offers improved response detection over conventional statistical metrics," *Journal of Neuroscience Methods*, vol. 293, p. 59–66, 2018.

[68] G. M. Bidelman and L. Powers, "Response properties of the human frequencyfollowing response (FFR) to speech and non-speech sounds: level dependence, adaptation and phase-locking limits," *International Journal of Audiology*, vol. 57(9), p. 665–672, 2018.

[69] S. Ananthakrishnan, A. Krishnan and E. Bartlett, "Human Frequency Following Response: Neural Representation of Envelope and Temporal Fine Structure in Listeners with Normal Hearing and Sensorineural Hearing Loss," *Ear and Hearing,* vol. 37(2), pp. 91-103, 2016.

[70] L. Salhi and A. Cherif, "Robustness of Auditory Teager Energy Cepstrum Coefficients for Classification of Pathological and Normal Voices in Noisy Environments," *The Scientific World Journal,* vol. 3(2), 2013.

[71] Y. Taigman, M. Yang, M. Ranzato and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," in *IEEE Conference on Computer Vision and Pattern Recognition*, Colombus, Ohio, 2014.