Classification of Frequency Following Responses to English Vowels in a Biometric Application

Rui Sun

Thesis submitted to the University of Ottawa

in partial fulfillment of the requirements for the degree of

Master of Applied Science

in Electrical and Computer Engineering

Ottawa-Carleton Institute for

Electrical and Computer Engineering

School of Electrical Engineering and Computer Science

Faculty of Engineering

University of Ottawa

Abstract

The objective of this thesis is to characterize and identify the representation of four short English vowels in the frequency following response (FFR) of 22 normal-hearing adult subjects. The results of two studies are presented, with some analysis.

The result of the first study indicates how the FFR signal of four short vowels can be used to identity different subjects. Meanwhile, a rigorous test was conducted to test and verify the quality and consistency of responses from each subject between test and retest, in order to provide strong and representative features for subject identification.

The second study utilized machine learning and deep learning classification algorithms to exploit features extracted from the FFRs, in both time and frequency domains, to accurately identify subjects from their responses. We used three kinds of classifiers with respect to three aspects of the features, yielding a highest classification accuracy of 86.36%.

The results of the studies provide positive and important implications for establishing a biometric authentication system using speech-evoked FFRs.

Acknowledgements

I would like to thank my supervisors, Dr. Martin Bouchard and Dr. Hilmi Dajani, for their guidance, suggestions, and great patience during my studies and research. Their help and support consistently encouraged me to finish this work. It is my honor to have them as my supervisors.

I am thankful to my parents for their support and motivation during my studies. Their understanding and love are invaluable.

I am grateful to my colleagues and friends Rajshekhar, Hitham, Hala and Hilda, for their kind help and sharing. They have shared advice on both my research and my life. It is a great pleasure to work with them — also, many thanks to my friends who helped me during my life and study.

Table of Contents

List of Tables

List of Figures

List of Acronyms

FFR	Frequency Following Response
AEPs	Auditory Evoked Potentials
CAP	Compound Action Potential
ECochG	Electrocochleogram
ABR	Auditory Brainstem Response
MLE	Middle-Latency Response
ASSR	Auditory Steady-State Response
cABR	Complex Auditory Brainstem Response
HMM	Hidden Markov Model
SNR	Signal to Noise Ratio
EEG	Electroencephalogram
eFFR	Envelope Frequency Following Response
sFFR	Spectral Frequency Following Response
PCC	Pearson Correlation Coefficient
FFT	Fast Fourier Transform
SVM	Support Vector Machine
CART	Classification And Regression Tree
CNN	Convolutional Neural Network
ReLU	Rectified Linear Unit
RBF	Radial Basis Function

1 Introduction

1.1 Background

A series of biometric technologies have been proposed, researched and evaluated in recent years with the improvement of understanding of human body measurement. Examples include physiological features like face recognition, DNA, fingerprint, iris recognition, hand geometry, ear, retina, and behavioral features like voice, signature, and gait. Biometric systems are logically pattern recognition systems as they involve detecting the unique biological or personal characteristics from the human being for verification purposes. However, the behavioral features in particular may slowly change during the course of one's life.

These biometric methodologies are expected to lead to a revolution of information security systems away from needing to remember numerous password and pin numbers to using unique biometric information, which can also prevent unauthorized users from getting access to secure area. It can be generally described as a more convenient identification method based on the physical and behavioral characteristics (Douglas et al., 2018). When comparing the biometric system with conventional knowledge-based or token-based systems, it links to the owners directly and cannot be lost or misused easily.

A biometric system is essentially a pattern recognition system. It can be logically divided into two distinct modes, namely, enrollment mode and recognition mode. During the phase of enrollment, the biometric trait of the individual is firstly scanned to acquire a digital copy of the trait. The digital copy is further processed through feature extraction in order to generate a 'template'. During the recognition phase, the biometric trait of the individual is

captured through the reader. It is further fed to the feature matcher that compares it against the stored data to determine the identity of the individual (Douglas et al., 2018).

The application of biometric information has evolved rapidly during the past decade and has been used for a great number of applications. These include but are not limited to biometric security, border control such as airports, financial systems, mobile biometrics and so on. These applications benefit a lot from biometric information when the identity or the information of humans can be recognized or recorded with efficiency. Meanwhile, each biometric technology has its own advantages and limitations. No single biometric approach is expected to meet the requirements of all the applications effectively. Which biometric approach should be used for a given application? The match between an application and a biometric approach depends upon the requirements and characteristics of the given application and the properties of the biometric approach.

A multitude of biometric techniques either used in practice or under investigation are presented to provide a clearer context. Facial images are the most commonly used biometric characteristic. Approaches for face identification are usually based on subject's skin color, gender, and shape of the facial attributes, such as the eyes, mouth, nose, forehead, and chin shape along with their spatial relationship and location and environment (indoor or outdoor) from both images and videos (Klare et al., 2015). Some state of art models for face recognition systems have been constructed in the past few years. For example, Taigman et al. developed a nine-layer deep neural network model for face recognition and achieved an accuracy of 97.35% based on the Labeled Faces in the Wild (LFW) dataset (Taigman et al., 2014). Fingerprints, on the other hand, have been used for personal identification for centuries and their validity has been clearly established by Jain et al. (Jain et al., 1997). The fingerprint is the pattern of local ridges and furrows on the surface of a fingertip, whose formation is determined in the fetal period (Hong & Jain, 1997). With regard to retinal biometrics, the pattern is formed by veins beneath the retinal surface of the eye, which is unique and stable, and is therefore a

feasible characteristic for identification (Hill, 1978). This pattern can be captured by projecting a low-intensity beam light into the eye and obtaining the image of the retina. However, imaging a retina needs a high degree of user cooperation in many actual applications, as the subject needs to closely gaze into an eye-piece and focus on a predetermined spot (Jain et al., 2000). As for the use of speech for identification, the speaker identity is dependent on the physiological and behavioral characteristics of the speaker, which exists in not only the short-time spectral content (vocal tract characteristics) but also in the supra-segmental characteristics of speech (Furui, 1997). However, the speech features are sensitive to a multitude of factors, for example, background noise, and the emotional and physical state of the speaker (Jain et al., 2000).

Several types of auditory evoked potentials (AEPs) have recently been suggested for potential use in biometric applications. The AEPs are a subclass of event-evoked potentials – brain neural responses that are time-locked to some event or to the omission of a stimulus (Heffernan, 2019). For AEPs, the stimuli are audio signals and the elicited response to the stimulus is a small electrical potential that can be recorded non-invasively from the scalp. The stimuli include but are not limited to speech sounds, clicks, tones and noise (Heffernan, 2019).

The frequency following response (FFR) to speech is a type of AEP. The FFR is a short-latency, scalp-recorded neuro response reflecting phase-locked activity from the human auditory system (Bidelman & Powers, 2018; Chandrasekaran & Kraus, 2010). The response components related to the stimulus envelope is referred to as the envelope FFR, which is distinguished from the components related to the stimulus spectrum, and is referred to as the spectral FFR. The envelope FFR is obtained by adding responses to complete opposite polarity stimuli, while the spectral FFR is obtained by subtracting these responses (Aiken & Picton, 2008a).

1.2 Motivation and Aims

There is increasing interest in using biometric technologies in daily life. Most popular systems rely on the audio-recorded and vision-recorded methods. It is natural to wonder whether or not an auditory evoked response, such as the Frequency Following Response, can be used for setting up an identification system. And if so, what is its performance?

Looking at the commonly used biometric methods nowadays, some of them such as fingerprint, speaker recognition, face recognition and so on can be easily recorded using a camera or audio recorder. In fact, a high-resolution camera or high-quality audio recorder is sufficient for taking the personal information by people with malicious intention. Moreover, with most biometric methods, once the physical characteristics are determined and stored, they become vulnerable to theft by determined hackers.

As a result, there is a high demand for biometric methods with higher security level to be developed, which means that the biometric data should ideally be both hard to copy and recorded using regular recording equipment. As such, the audio evoked frequency following response, as a subclass of auditory evoked potentials, potentially stands out compared with regular audio-recorded and vision-recorded biometric techniques. First, because it requires electrodes placed on the scalp, it cannot be easily recorded by a third party, and second, since new stimuli can be periodically used to refresh the stored responses, it is less vulnerable to theft than “fixed” biometric traits like fingerprints and retinal features. It should also be noted that currently we do not possess a mathematical model of the processes that generate this response in the brain. Therefore, a malicious party cannot generate the expected response to new stimuli, unlike biometric approaches based on speaker identification, where it is currently

possible to create a model of speech production for an individual based on samples of their recorded speech.

What is needed in order to verify that the Frequency Following Response or other kinds of auditory evoked response can be used for identification? Clearly, a thorough understanding of the normal response is necessary to characterize these signals. Therefore, as a starting point, the evoked response should be collected in a laboratory environment using specialized equipment, while later on, less controlled environments and more practical equipment can be evaluated.

Once the responses are recorded, a set of verified features should be extracted from the original neural potentials elicited by the stimuli. Here, a suitable feature selection methodology such as Pearson correlation coefficient or distance correlation or other methods can be used to find out the optimum features that can minimize the redundant and irrelevant information. Furthermore, a machine learning classifier which is compatible with the selected features could be constructed to maximize the classification accuracy.

Although there has been work done to characterize the frequency following response (Ananthakrishnan et al., 2016; Bidelman & Powers, 2018; Chandrasekaran & Kraus, 2010; Dolphin & Mountain, 1992; Greenberg et al., 1987; Huis et al., 1977; Krishnan, 2002; Laroche et al., 2013; Llanos et al., 2019; Russo et al., 2004; Sadeghian et al., 2011, 2015; Won et al., 2016; Yellamsetty & Bidelman, 2019; Yi et al., 2017), to the best of our knowledge, except for a very recent publication by (Llanos et al., 2019), there has been no other study to date that has focused on investigating automatic listener identification using the speech-evoked FFR. Therefore, in addition to the characterization and classification of speech-evoked FFR with different subjects, this work aims to explore if the extracted features from the speech-evoked

FFR are stable and representative, and thus effective for constructing a solid biometric identification system in laboratory environment.

In a previous study, the FFR in normal hearing subjects to 4 short synthetic vowels with different levels was analyzed with respect to the frequency peaks and spectral content of the envelope FFR and spectral FFR (Heffernan, 2019). A machine learning approach was used to classify the responses to the different vowels at different presentation levels. In addition, the effect of gender on the amplitude spectra and major waveform peaks of speech-evoked FFRs was investigated. This is the dataset that is used for our study.

1.3 Contributions

The following contributions resulted from this work:

1) The FFR data had previously been collected over two sessions on two separate days (referred to as Test and Retest). The stability of the responses between Test and Retest for each subject was investigated in detail in this work. It was found that the envelope FFR appears to provide more robust information related to subject identification than the spectral FFR.

2) Time and frequency domain features of responses to a single vowel and concatenation of responses to the four vowels were compared. A concatenation of time and frequency domain envelope FFR was chosen for robust listener identification. The spectrum of four vowel FFR concatenation was used to generate input features for machine learning.

3) A variety of classification algorithms were investigated for listener identification. Most were found capable of obtaining high classification accuracies when trained and tested on features from the FFRs. The evaluated machine learning algorithms include a support vector machine

with linear kernel, radial basis function kernel and polynomial function kernel, XGBoost, and a convolutional neural network.

To achieve the objectives of this thesis, an experiment was first conducted in Python version 3.6.3 environment. The Pandas and Scipy libraries were used for data manipulation and analysis, and signal feature extraction. Machine learning classification was conducted based on the Scikit-learn library and deep learning methodology in the Tensorflow library.

1.4 Structure

This thesis is composed of five chapters, and its structure flows naturally following the experiment design logic. The chapters are organized as follows:

1. Chapter 1 reviews the background of the auditory brainstem responses, motivations, aims and a brief introduction of the structure of the thesis.

2. Chapter 2 provides background information pertaining to the FFR and the literature relevant to the studies addressing the selectable features.

3. Chapter 3 provides detailed information of the methods, results and analyses for the FFR feature selection and the verification of the reliability of the features.

4. Chapter 4 describes and evaluates the machine learning and deep learning algorithms selected and the discussion of the performance of these algorithms.

5. Chapter 5 presents a general conclusion based on the results of the work and provides suggestions for future work.

2 FFR Feature Representation

This chapter commences with a general overview of auditory evoked potentials (AEPs) and subsequently introduces the characteristics of the frequency following response, which is a subset of the AEPs.

2.1 Overview of Auditory Evoked Potentials

The auditory evoked potential results from the electrical changes that are generated in the human auditory system in response to sounds, while the speech-evoked frequency following response is the AEP when the stimulus is speech. These are typically recorded using surface electrodes placed on the scalp.

Many different auditory potentials can be recorded when sounds are presented. Some way of categorizing them is essential for understanding. The first classification of these potentials was conducted by Hallowell (1976) based on the peak amplitudes and latencies, the time between the onset of stimulus and the onset of the response. Therefore, AEPs include first, fast, middle, slow or late responses (Picton, 2010). The first and fast responses are usually considered as early, and the slow and late are combined together as late. In this way, the compound action potential (CAP) of the auditory nerve is described as part of the electrocochleogram (ECochG), and the fast response is described as the auditory brainstem response (ABR). Meanwhile, the term “middle-latency response”, or MLR, and “late auditory evoked potential” are used to describe slow and late responses.

The latency time provides another way to classify the AEPs. A “transient” potential is evoked following a change of the stimulus, for example the onset and the offset, and is distinguished from “sustained” responses that occur continuously throughout the stimulation. If a stimulus changes in a regular way, a “steady-state” response is evoked. This type of response

is termed as the "auditory steady-state response" (ASSR). The following is a table adapted from "Human Auditory Evoked Potentials" by Terence W. Picton, presenting a detailed classification of the common AEPs that uses both onset-latency and time-course.

Table 2.1 : Classification of the Auditory Evoked Potentials. Adapted from "Human Auditory Evoked Potentials" (Picton, 2010)

Latency	Transient	Steady-State	Sustained
First (0-5 ms)	Cochlear Nerve Compound Action Potential (CAP: N1, N2)	Cochlear Microphonic (CM)	Summating Potential (SP)
Fast (1-15 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 (P3 or P300)		Contingent Negative Variation (CNV)

2.2 The Frequency Following Response

The frequency following response (FFR), typically recorded in response to a brief tone, is a scalp-recorded potential originating from the human auditory system. However, there has been a battle for terminological clarity declared in the last decade. In an authoritative tutorial involving FFR recording and analysis (Skoe & Kraus, 2010), the authors introduce it using

another acronym (cABR for the complex auditory brainstem response), so as to distinguish the ABR to complex sounds from the traditional clicked-evoked ABR. Meanwhile, this tutorial divides the cABR into two components, namely the envelope response and the spectral response. Many of the recent papers seems to simply refer to the 'FFR' and some more specific terms like 'human FFR' or the 'scalp-recorded FFR'. The issue here is that both envelope FFR and spectral FFR are often ambiguously referred to as FFR. On the other hand, Aiken and Picton distinguished and clarified the term "envelope FFR" and "spectral FFR" (Aiken & Picton, 2008a).

The speech sound can evoke both transient and sustained responses in the human brainstem and cortex. For example, a consonant-vowel diphone or a pure vowel stimulus can elicit the transient brainstem responses (Aiken & Picton, 2008a). In research led by N. Kraus, the speech-evoked auditory brainstem response evoked by /da/ was used to evaluate the auditory brainstem pathway encoding of speech in children with learning problems (Cunningham et al., 2001; King et al., 2002) and children with pronounced speech perception difficulties (Johnson et al., 2007). Meanwhile, the transient cortical responses have been widely used with regard to evaluating hearing impairment and with hearing aids (Billings et al., 2007; Golding et al., 2007; Korczak et al., 2005; Rance et al., 2002). However, these responses are less related to the content of the speech stimulus but rather to sudden changes in magnitude or frequency.

2.2.1 Envelope and spectral FFR

The nature of the FFR was contentious as lacking of a clear and standardized definition. The brainstem response to sustained speech and speech-related stimuli have been called "envelope-following responses" (Dolphin & Mountain, 1992), "frequency-following response" (Moushegian et al., 1973) and "auditory steady-state responses" (Stapells et al., 1987). In certain cases, the term 'frequency-following responses' was used for describing responses to speech

formants by (Plyler & Ananthanarayan, 2001), two-tone synchrony suppression (Krishnan, 1999), the speech harmonics (Krishnan, 2002) and the fundamental frequency (Skoe & Kraus, 2010). All of these parameters are related to the speech envelope and therefore the term "frequency-following response" is used in a general meaning, which is a response that is elicited by either the spectral content of the stimulation or the components of the envelope (Aiken & Picton, 2008a).

A clear description was given by (Picton, 2008a) in order to distinguish between the "envelope FFR" and the "spectral FFR". The envelope FFR is mainly insensitive to the stimulus polarity. Therefore, spectral FFR can be obtained by saving the responses to stimuli in alternate polarities and then averaging the difference between the two responses recorded (Huis et al., 1977). When averaging the sum of responses to stimuli with alternate polarities, this manipulation would eliminate the spectral FFR and the envelope FFR will be preserved (Chimento, 1990). The following table reproduces a table from (Picton, 2008a) that describes the effects of different combinations of responses to stimuli of positive and negative polarities.

Table 2.2 : Average response nomenclature. A summary of possible combinations of the evoked responses to the original stimulus and its opposite polarity, as well as the components contained within the response signal. Adapted from (Aiken & Picton, 2008a)

Response	Derivation	Components
++	Average together all responses to the original stimulus	Envelope FFR
		Spectral FFR
		Cochlear microphonic
		Stimulus artifact
+-	Average together an equal number of the responses to the original stimulus and responses to the inverted stimulus	Envelope FFR
- -	Subtract responses to the inverted stimulus from an equal number of responses to the original stimulus and divide by the total number of responses	Spectral FFR
		Cochlear microphonic
		Stimulus artifact

2.2.2 FFR study-related parameters

A summary of stimulus parameters for the frequency following response is given here. It includes stimulus type, duration, stimuli intensity and polarity, as well as the stimuli presentation-related and recording-related parameters such as materials, sampling rate, filter used, signal averaging, artifact minimization etc. It is adapted from (Skoe & Kraus, 2010).

Table 2.3 : Recommended stimulus and recording parameters. Adapted from "Auditory Brain Stem Response to Complex Sounds: A Tutorial" (Skoe & Kraus, 2010)

Parameter	Recommendation	Rationale/Comments
Stimulus		
Type	Speech, music, non-speech vocal sounds, environmental sounds, etc.	Examine how behaviorally relevant sounds are turned into neural code
Characteristics		
Transient	Well-defined temporal features such as strong attacks and amplitude bursts	Maximize transient responses
Sustained	$F_0 < 500Hz$	Maximize sustained responses
Creation	Natural, synthetic, or hybrid	cABR stimuli can be created with many different software packages
Duration	Short: 40 ms to 100 ms Long: 100 ms to 500 ms	Minimizes recording time Maximizes naturalness
Stimulus Presention		
Intensity	Well above threshold: 60 – 80 dB SPL	Stimuli should be precisely calibrated before each test session using a sound level meter
Monaural Stimulation	Separate norms should be collected for each ear	Monaural is preferred for children
Binaural Stimulation	Maximizes response characteristics	Binaural is more realistic than monaural
Transducer	Magnetically shield ear inserts	Minimizes stimulus artifact
Rate and ISI	Rate: dependent on stimulus duration ISI: >=30% of stimulus duration	See table 3 for recording-based issues that impact rate and ISI decisions
Presentation Software	Perform thorough testing to ensure precise, non-jittered stimulus presentation	Because of the temporal sensitivity of the cABR, a small amount of jitter will spoil the response
Electrode placement	Vertical montage (active Channel: Cz; reference: earlobe(s); ground: forehead)	For rostral brainstem recordings; a horizontal montage is used for recording more peripheral structures

Sampling Rate	6000 – 20000 Hz	Better temporal precision with higher sampling rates
Filtering	Low pass cutoff: 2000-3000 Hz High pass cutoff: 30-100 Hz If possible, collect cABR with open filters(1-3000 Hz) and bandpass filter offline using digital filters	More defined transient peaks. Depends on spectral characteristics of stimulus. Digital filters minimize temporal phase shifts.
Signal Averaging	2 or more sub-averages of 2000-3000 sweeps	Determine response replicability. Spectral-domain averaging will increase spectral estimates and require fewer sweeps
Averaging Window	Begin 10-50 ms before stimulus onset Extend 10-50 ms after onset	An adequate sample of the baseline is needed to determine whether a particular response peak is above the noise floor For running window analyses: the pre-stimulus tie window should be > or = to the duration of the analysis window Neural activity should return to baseline
Simultaneous cABR-cortical response recording Minimizing artifacts	Recommended only if large files can be accommodated and longer sessions are appropriate Passive collection protocol Electromagnetically shield insert ear phones Alternating stimulus polarity Artifact rejection criterion: >20 μV	 Minimizes myogenic artifacts Minimizes stimulus artifact Enables adding of response to minimize both stimulus artifact and cochlear microphonic Exclude trials exceeding typical neural responses size

Here, we will discuss the major parameters from the table given above, which includes the type of the response, namely transient response and sustained response, as well as the commonly discussed stimuli for speech FFR.

The type of the response is divided into two parts: the transient response and the sustained response. The transient features are usually evoked by non-sustained, brief stimuli. For audio stimuli, it usually refers to the onset of the sound and the offset of the sound. For example, with regard to the consonants, the transient response labels the onset portion of the consonant characterized by unvoiced, broadband frication (Chandrasekaran & Kraus, 2010). The morphology of speech FFR onset is described by the characteristics of the attack, which means how quickly the sound reaches full volume. The stimuli with a sharper rise, such as abrupt onset or clicks, are more broadband (less frequency specific) and result in broader and more simultaneous reaction of the cochlear nerve fibres (Skoe & Kraus, 2010).

The character of the attack is important for speech sound quality and helps in the identification of specific speech sounds (Rosen, 1992). By definition, the obstruent stop consonants (e.g., /d/, /p/, /k/) have sharper stimulus onset than nasals and glides (e.g., /m/ and /y/). Moreover, abrupt changes in the amplitude of the sound can also evoke envelope-like transient responses. For example, Strait et al. recorded brainstem responses to the sound of a baby's cry and included a series of transient responses, in addition to enhanced pitch and timbre amplitudes to the most spectrally complex section of the sound (Strait et al., 2009).

As for the sustained response, it follows the periodic components of the speech stimulus such as sinusoidal tones, harmonically complex vowels which elicit sustained brainstem responses. The sustained response in human scalp-recorded brainstem responses was first recorded by Moushegian in 1973 (Moushegian et al., 1973). They demonstrated that each frequency evokes a certain response and the tones in the speech range from 250 Hz to 2kHz

evoke predictable neural responses which are related to the frequency of the tones. The evoked responses, for this reason, are often called frequency following responses. The scalp-recorded FFRs are usually recorded within an upper limit of 1500 Hz (Aiken & Picton, 2008b; Moushegian et al., 1973; Skoe & Kraus, 2010) because neural phase-locking in auditory fibres becomes weaker when frequency increases (Greenberg et al., 1987).

For the recording times of the experiments, the duration of the stimulus is often as brief as possible so as to accurately capture the responses, but many responses (often in the thousands) that are synchronized to the stimulus are usually recorded and coherently averaged in order to achieve an acceptable signal-to-noise ratio. For speech stimuli, the fundamental frequency (F0) usually ranges from 80-500 Hz. When selecting a speech-phoneme for FFR recording, it is necessary to keep in mind that the second formant of many vowels are over the range of neural phase-locking (Moushegian et al., 1973) and therefore may not be observed in the response (Johnson et al., 2008).

The diversity of different subjects would also affect the results in different aspects. With respect to the effect of age, in (Sanfins et al., 2012) a set of frequency following responses was presented and analysed from different age groups: infants and young children, children and adolescents, and adults and the elderly. Based on the analysis of amplitude and latency of the evoked response, it was found that the FFRs in different age groups exhibit specific characteristics,.

Ananthakrishnan et al. examined the effects of sensorineural hearing loss to the FFRs compared with normal hearing subjects (Ananthakrishnan et al., 2016). The FFRs were recorded from 10 normal hearing subjects and 9 mild-moderate sensorineural hearing loss subjects in response to a steady-state vowel /u/ at multiple sound intensity levels. It was found that both F0 and F1 at equal sound level stimuli were stronger in normal hearing listeners compared to listeners with sensorineural hearing loss.

The Stimuli for speech-evoked FFR

In order to understand the brain processing of complex speech stimuli, several research studies have been conducted recently using vowels as speech stimuli.

Two comprehensive studies on normal hearing subjects were conducted by Krishnan in 1999 and 2002 (Krishnan, 1999, 2002). Both of them are similar in design and 10 normal hearing adults were chosen as subjects with stimuli at four sound levels (55, 65, 75,85 dB nHL). The former used two-tone steady state approximations to three vowels (/a/, /ɔ/ and /u/) and the latter used a synthetic steady state version of the same. Both of the studies analyzed the FFR in the time and frequency domains. The results indicate a strong phase-locking activity among two distinct populations of neurons and a clear effect on the spectral amplitude at both first and second formants.

The speech ABR with variants of a synthetic vowel in quiet and background noise was investigated by Laroche et al. (Laroche et al., 2013). The speech ABRs of 18 normal-hearing subjects were recorded using variants of a 300 ms formant-synthesized /a/ vowel in both white noise and quiet environment. It was found that the evoked response at the fundamental frequency with the variant dominated by resolved harmonics was more robust to noise than that with stimulus variants dominated by unresolved harmonics, which supports the idea that the pitch of resolved harmonics and that of unresolved harmonics are processed in interacting pathways converging in the upper brainstem.

A study for vowel decoding from single-trial speech-evoked frequency following response was recorded by Yi et al. (Yi et al., 2017). It was found that the ability to extract interpretable features with single trials provides significant potential for the assessment of human auditory function. In this study, a novel, data-driven approach, namely a machine learning algorithm, was used to decode information related to speech signal from single trial

FFRs. Thirty-eight young adults participated as subjects and the FFR was evoked by two vowels (/a/ and /u/) produced by two native English speakers. Scalp-recorded responses were projected onto a low-dimension spectral feature space from the same vowels produced by 40 other speakers.

Bidelman provided a different idea from a study of the sonification of scalp-recorded FFRs, as it offers an improved response detection over statistical metrics (Bidelman, 2018). He mentioned that the sonification has two or three times the efficiency compared to statistical approaches. He therefore suggests that simple listening to FFR responses provides a rapid technique for real-time FFR recording, and a stopping rule is provided to terminate signal averaging that is better than current approaches. In order to examine whether sonification of the FFR could be used to assess the quality of the running recordings, sixteen young adult with normal hearing were chosen to be stimulated with a synthetic, steady -state vowel token/a/. It was found that their evaluation of the speech-FFRs is based on their perception of the presence/absence of a tonal quality solely.

A hidden Markov model was used to decode the listener identity from the FFR spectro-temporal features through multiple frequency bands by (Llanos et al., 2019). Their study provided an initial and systematic biometric characterization of the FFR and confirms the practical viability of the FFR as a biometric identification system. A total of 20 subjects, 10 native speakers of English and 10 of Mandarin Chinese, listened to Mandarin Chinese tones across three sessions on different days. The subject identity was recognized within the same auditory context (same tone and session) and across different stimuli and recording sessions using an hidden Markov model (HMM). The best result is derived from the HMM model which decoded subject identity with a mean AUROC of 0.93 within the same auditory context (with same tone and session). It was found from the result that the HMM decoded listeners for one single FFR, but model performance increased when more responses were averaged. Meanwhile, listeners unfamiliar with the evoking stimuli (native English speakers) were better

decoded than listeners familiar with the evoking stimuli (native Chinese speakers), which shows that the identification results can be affected by the sound familiarity.

A test was conducted by Yellamsetty and Bidelman to examine subcortical encoding in concurrent speech identification by human subjects through FFR (Yellamsetty & Bidelman, 2019). The results demonstrated that the pre-attentive subcortical encoding could predict perceptual speed but not accuracy. With regard to the stimuli, two steady-state vowels (/a/ and / ɛ/) whose F0 differed by zero or four semitones (0ST, 4ST) were presented diotically in both quiet and noise-degraded (+5dB SNR) conditions. Listeners also needed to finish a speeded double vowel identification task. The neurophysiological data demonstrated more robust FFR F0 amplitudes for single vowels compared to double vowels and a weaker response in noise. The FFR F0 amplitudes, however, failed to predict the listeners' identification performance. In contrast, FFR F1 was found to be related to faster reaction times when limited by noise.

2.3 Methods

This study aims to answer the following two questions:

1. Does the FFR representation of four short vowels in individual subjects remain stable across test and retest conditions?

2. What is the practical viability of the FFR as a biometric trait across different subjects and recording days?

Based on the dataset recorded in an earlier study, the response of normal hearing adults to a fairly commonly seen stimulus was used. Therefore, English vowel sounds were selected here as stimulus since they are primitive but meaningful speech units and they have been studied

quite extensively in the literature. Four different vowel stimuli (100ms /a/, /ɔ/, /U/, /u/ vowels) were selected for this study. In addition, determining whether or not the speech FFR is stable from the test session to retest session is significant. Within the context of the work, a high test-retest reliability of responses would certainly have a positive effect on the biometric identification task. The recording of these FFRs was conducted in an earlier study for a different purpose, namely related to the classification of the responses to the different vowels rather than the different listeners (Heffernan, 2019). An overview of the recording method used is given below in Sections 2.3.3, but further details can be found in that study.

2.3.1 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 over two sessions (a minimum of 3 days apart, and 23 days apart on average). During the process of recording, two additional female subjects were excluded as one failed the audiometric threshold testing and another one had been very recently been referred to an ENT by an audiologist although she passed the audiometric testing. Therefore, the number (index) of subjects ranges from 1 to 25 but misses the numbers 5, 10, and 24, as shown on the following figures.

Before recording, each subject underwent an audiometric test to confirm that they had hearing thresholds of 20dB HL or less at 250, 500, 750, 1000, 2000, 4000 Hz in both ears. Participants were seated comfortably in a reclining chair in a sound attenuating booth that passes the ANSI standards for background noise level, ANSI S3.1-1999 (R2013). They were encouraged to keep awake and still throughout the duration of the recording session. All subjects provided consent and all recordings were performed according to the requirements of University of Ottawa's Ethics Board.

2.3.2 Stimulus creation and calibration

The study made use of four 100 ms vowel sounds, namely /a/ as in 'father', /ɔ/ as in 'call', /U/ as in 'boot', /u/ as in 'who'. All four selected stimuli have a common fundamental frequency (F0) of 100 Hz. The 100 ms duration was selected in order to have an integer number of periods at F0.

A simplified Klatt formant synthesizer with a sampling rate of 48kHz at 16-bit resolution was used for generating the stimuli. The synthesized vowel parameters are indicated in the following table.

Table 2.4 : Parameters of stimulus for vowel duration, fundamental frequency, first, second and third formant frequencies, bandwidth and related levels.

Vowel	Dura. (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

All stimuli were presented to each subject at the level of 85dBA and the stimulus levels were calibrated with an insert earphone connecting 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.

2.3.3 Experiment setup

The FFRs were recorded using the Bio-logic BioMARK system with version v7.0.2 and non-disposable silver chloride electrodes were used for all of the recordings. The stimulus was transferred to the right ear by a Bio-logic foam-tip phone insert. The left ear was kept excluded through the duration of the experiment. The active electrode was placed at vertex (Cz). The

reference was placed on the right earlobe and the ground electrode on the left earlobe. In order to keep the impedance under 6 kΩ, a mild abrasive and a conductive electroencephalogram (EEG) paste were used on the skin of the participants. The EEG signals were amplified with a gain of 100,000 and bandpass filtered from 100 to 3000 Hz. In order to suppress the myogenic contamination, an artifact rejection criterion of 23.8 microvolts was applied. The stimulus onset synchronizes with the recording and the stimulus presentation rate was set at 8.4/s. As the recording system fixed the maximum 1024-points per epoch, the epoch time was set as 106.6 ms, giving a sampling rate of approximately 9606 Hz. A 3.4 ms onset recording delay was set up so as to maximize the capture of the sustained response.

Two recording sessions, test and retest, were conducted in order to enable an analysis of test-retest stability of the responses. For each vowel, the response for 85dB sound level conditions was recorded in two consecutive 1500 sweep blocks, giving a total of 3000 sweeps per session. Therefore, 6000 sweeps were collected for each subject. Sweeps were presented in alternating polarity and the envelope FFR (eFFR) was derived by averaging the summed responses in each of the two polarities. Meanwhile, the spectral FFR (sFFR) was similarly derived through averaging the difference between the signals from the two polarities (Aiken & Picton, 2008a).

2.3.4 Data handling and processing

All of the recorded responses were exported to ASCII text files using the Bio-logic "AEP to ASCII" software version1.2.1 and they were imported into a Pandas library DataFrame data structure using Python to prepare in the next step for offline analysis. The grand-mean eFFR and sFFR waveforms were generated by summing all the waveforms of subjects and dividing by the number of summed waveforms. Grand-mean of the amplitude spectra is the frequency domain representation of the grand-mean waveforms.

For the time domain signals, all the signals are detrended by removing the DC offset as it is a potential source of distortion (Akhoun et al., 2008). Then, a Hamming window was applied so as to suppress the spectral leakage after calculating the Fourier transform. Regarding the frequency domain signals, zero padding is helpful for obtaining a more densely interpolated frequency domain representation. Therefore, a series of 86454 zeros was added to the end of the time domain signals to facilitate correctly estimating the amplitude at a given frequency (for example, at the fundamental frequency of the FFR of 100 Hz).

Normalization was then conducted before comparing the different FFRs. One approach to normalization used the standard deviation as denominator of the function and another approach used the root mean square. These two approaches are detailed later.

2.4 Data Analysis and Results

Analysis of both envelope and spectral FFRs is focused on the frequency domain, especially on the amplitude of each harmonic. However, the amplitude features of the transient response (onset response) and steady state response are also analyzed in the time domain.

The grand-average envelope and spectral FFRs response for the 100 ms /a/, /ɔ/, /U/ and /u/ vowels are shown below (in Fig 2-1 and Fig. 2-2, respectively), with the left column showing the FFRs in time domain and the right column showing the amplitude spectra of FFRs in frequency domain. It should be noted that the character a, e, n, u represents the /a/, /ɔ/, /U/, /u/ vowels respectively. With the eFFR, analysis is done on the spectral amplitudes at F0, the second harmonic (namely, H2) to the sixth harmonics (H6). With the sFFR, the analysis mainly focuses on the response spectral amplitude of the first and second formants (F1 and F2) for each stimulus.

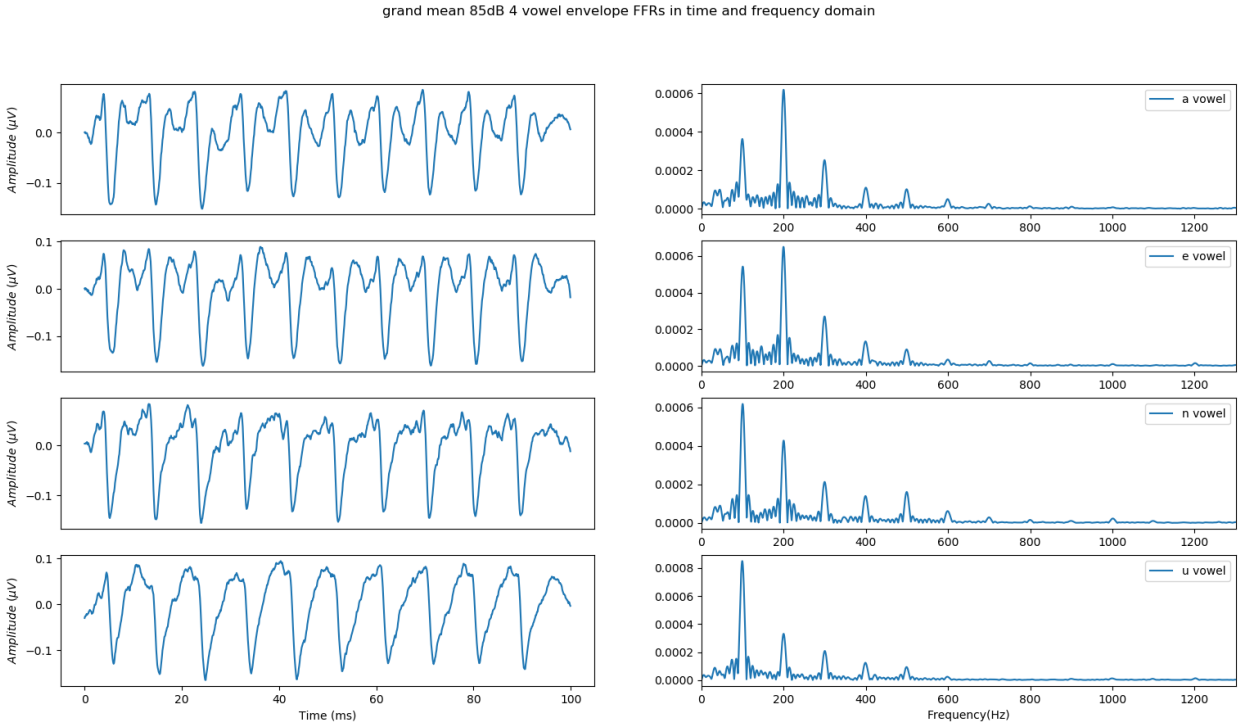

Figure 2-1 : Grand-average envelope FFRs for the 100ms 4 vowels at 85 dBA with F0=100 Hz in both time domain and frequency domain.

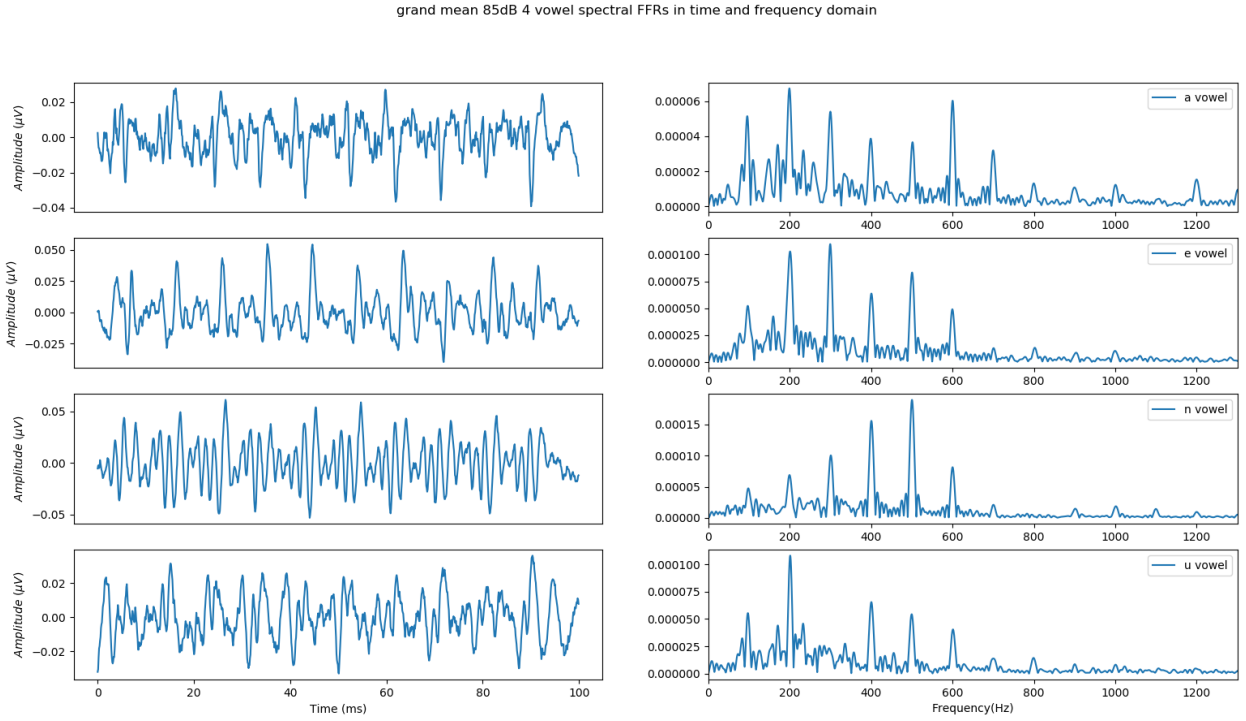

Figure 2-2 : Grand-average spectral FFRs for the 100ms 4 vowels at 85 dBA with F0=100 Hz in both time domain and frequency domain.

It should be noted that spectral amplitude of envelope FFRs at the fundamental frequency is typically high. It decays for higher harmonics, but sometimes the amplitude of the second harmonic is higher than that of the fundamental frequency, for example in /a/ and /ɔ/ for some subjects. On the contrary, the amplitude spectra of spectral FFR do not follow the same pattern, as often the component at F0 is absent or very weak. Meanwhile, the signal between the response harmonics is usually higher than that in the envelope FFRs, which may include personal biometric information but may also be mostly noise.

The grand-average of the time domain representations of both eFFRs and sFFRs are shown on the left side of the Fig. 2-1 and Fig. 2-2. It is noticed that, overall, envelope FFRs have obviously higher amounts of low frequency content than the spectral FFRs for all 4 vowels. Regarding the transient response, it is hard to find the general patterns with response signals.

The amplitude spectra of the eFFR and sFFR for each of the 100ms /a/, /ɔ/, /U/ and /u/ vowels for all 22 subjects are shown in Figs. 2-3, 2-4, 2-5, 2-6 and Figs. 2-8, 2-9, 2-10, 2-11 respectively, with the test and retest conditions (on different days). Here, each curve corresponds to the average of two sets of recordings (of 1500 trials each) on the same day. The blue and orange lines represent the amplitude spectra of the retest and test signals respectively, from 0 to 1300 Hz.

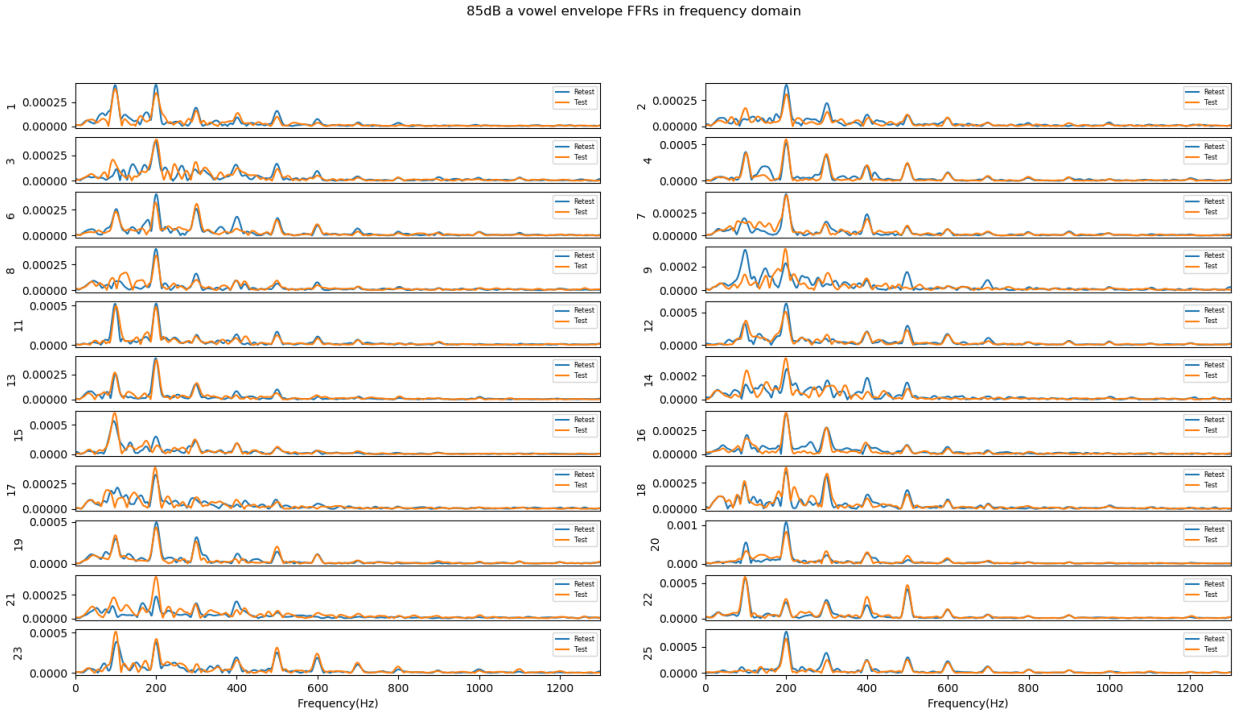

Figure 2-3 : Spectrum of frequency components of envelope FFR averaged over 3000 trials for a 100 ms /a/ vowel stimulus with F0=100 Hz presented at 85 dBA for all 22 subjects.

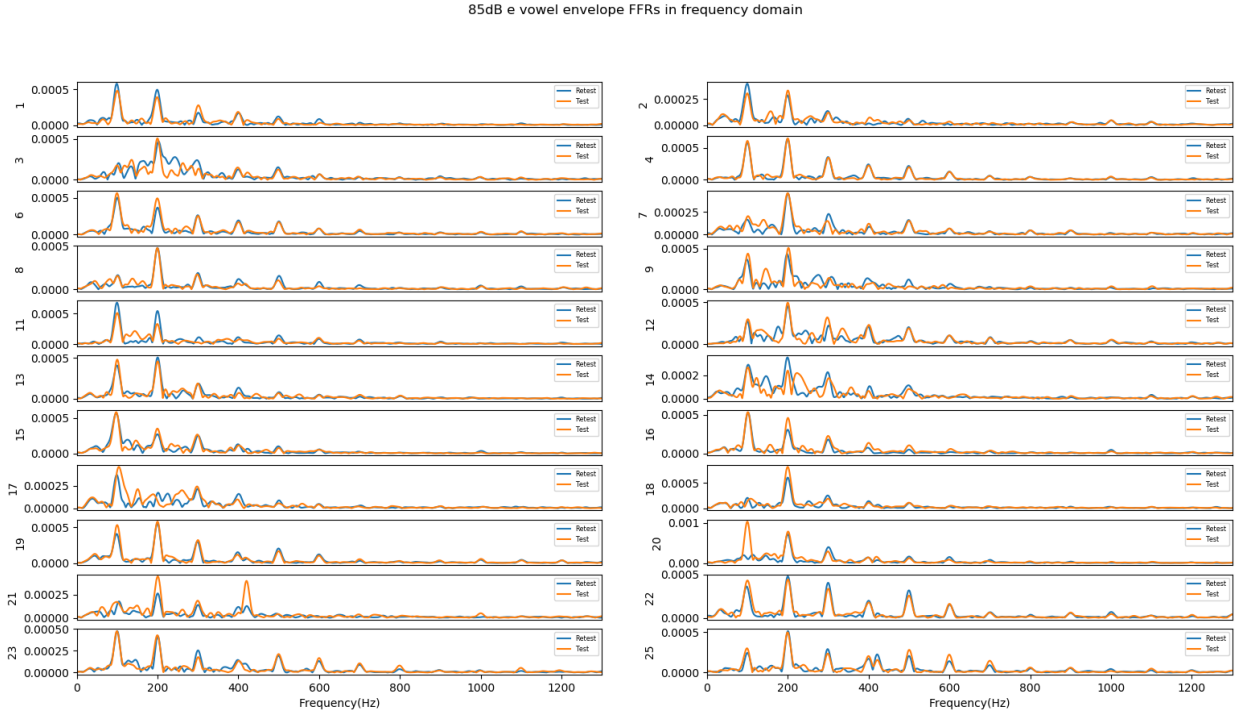

Figure 2-4 : Spectrum of frequency components of envelope FFR averaged over 3000 trials for a 100 ms /ɔ/ vowel stimulus with F0=100 Hz presented at 85 dBA for all 22 subjects

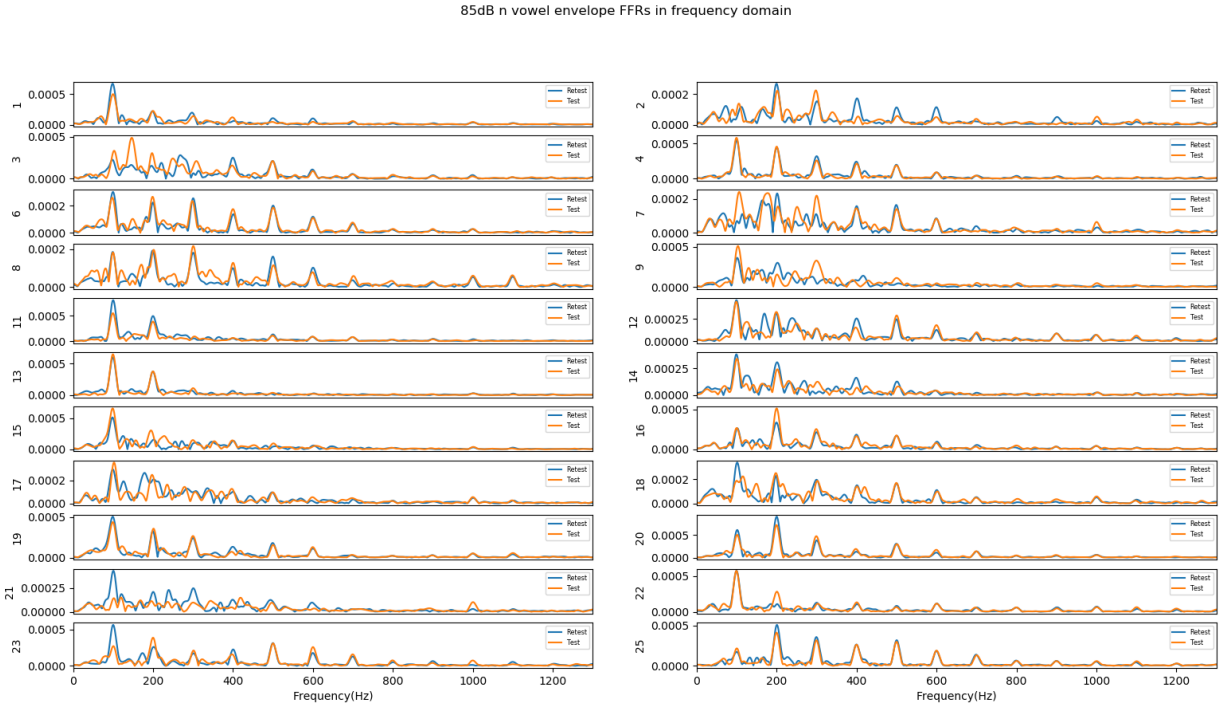

Figure 2-5 : Spectrum of frequency components of envelope FFR averaged over 3000 trials for a 100 ms /U/ vowel stimulus with F0=100 Hz presented at 85 dBA for all 22 subjects

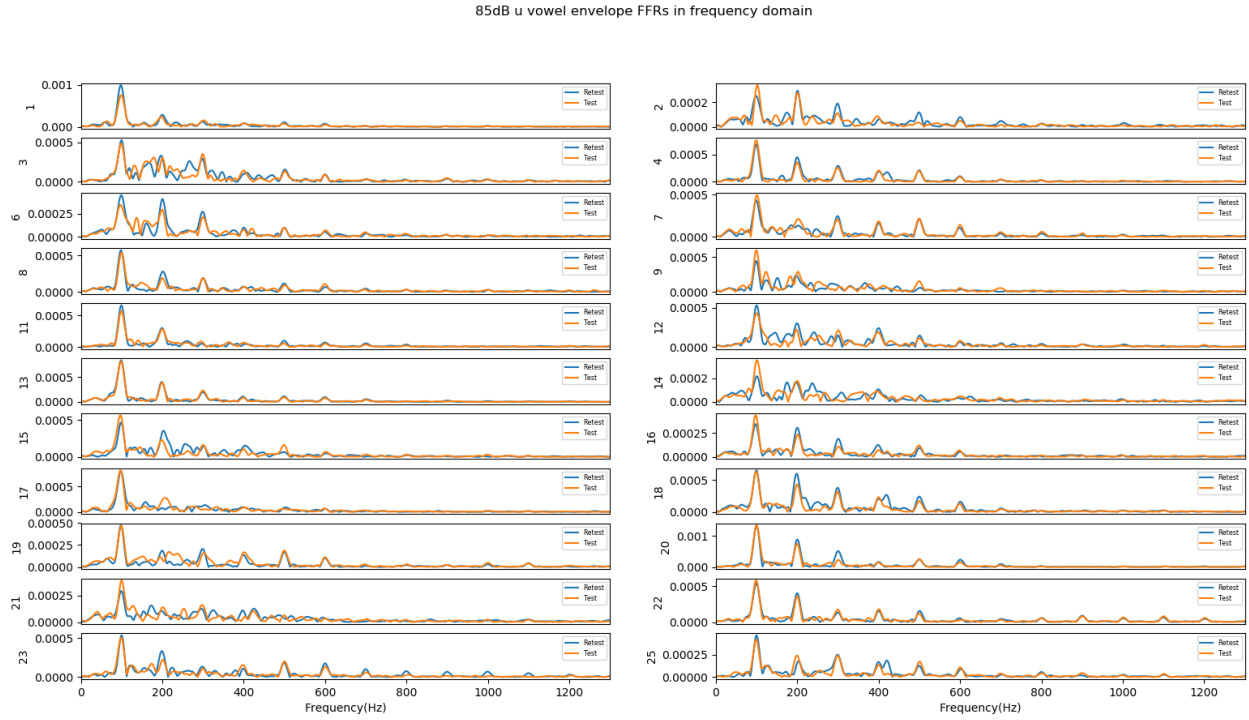

Figure 2-6 : Spectrum of frequency components of envelope FFR averaged over 3000 trials for a 100 ms /u/ vowel stimulus with F0=100 Hz presented at 85 dBA for all 22 subjects

The 100 ms /a/ vowel's averaged eFFR reveal a relatively similar tendency across subjects, with the amplitude of the peak at F0, and H2 to H6. The spectral amplitude of H2 is often higher than that of F0, with usually a decrease from H3 to H6. The spectra also seem to indicate that at certain harmonics for certain subjects there is a 'suppressed' peak, where the amplitude of the peak is obviously lower than others, such as at F0 for subject 25. Meanwhile, the spectral amplitude of H4 to H6 sometimes has the same amplitude as the signal between harmonics. Consequently, not all the subjects exhibit evident peaks at F0 and H2 to H6. Regarding intra-subject comparison, most subjects have a stable response between the test and retest conditions. However, some subjects indicate a much different waveform at certain harmonics. For example, the spectral amplitude at F0, H5 and H7 for subject 9 indicates obvious differences, with the amplitude of retest condition much higher than that of test condition.

The spectral FFRs averaged over 3000 trials for the 100 ms /a/ vowel shown in Fig. 2-8 reveals relatively clear harmonics at F1 (700 Hz) for most of the subjects. The majority of subjects seem to clearly show a peak at F2 (1299 Hz) as well, although the spectral amplitudes are relatively small in the high frequency region. This is due to the low-pass nature of the FFR. In contrast, it is difficult to identify response components at F0 and its low frequency harmonics.

The averaged envelope FFRs for the 100 ms /ɔ/ vowel shown in Fig. 2-4 reveals clear peaks at F0 and usually at H2 to H6. The spectral amplitude at F0 and H2 has similar value in most cases, except for certain subjects for whom the peak at F0 appears suppressed (e.g. subject 8 and subject 18). A good intra-subject comparison indicates that F0 and H2 to H6 are robust enough indicators of the /ɔ/ vowel, with the spectral amplitude of H2 mostly having the highest value compared with other harmonics.

The averaged spectral FFRs for the 100 ms /ɔ/ vowel shown in Fig. 2-9 reveal a generally similar pattern in that harmonics around F1 (600 Hz) mostly exhibit obvious peaks, although they are not stably represented for subjects 2, 3, 7 and 21. The F2 (900 Hz) is difficult to visually observe. It is noticed that the intra-subject comparison is difficult to make as the test and retest conditions are often not very similar, particularly for the lower harmonics.

For the averaged envelope FFRs for the 100 ms /U/ vowel stimulus shown in Fig. 2-5, it appears to show a different pattern compared with response to /a/ and /ɔ/. The F0 mostly exhibits a higher stability than harmonics from H2 to H6. It is noticed that the spectral amplitude of H5 (500 Hz, namely F1 for /U/ vowel) is generally slightly higher than the H4 (400). With respect to intra-subject comparison, the waveform from H4 to H6 usually matches well for most subjects except subject 2, which reveals obvious difference in the spectral amplitude of H4, H5, and H6. The F0 and H2 to H3 reveal an ideal matching in most subjects except subject 21 and 23, for whom the amplitudes of F0 in retest condition are higher than that of test condition.

The averaged spectral FFRs for the 100 ms /U/ vowel stimulus shown in Fig. 2-10 exhibit various patterns at F1 (500 Hz) and F2 (1200 Hz). However, they usually achieve a good matching between test and retest conditions in most subjects, which means that they are potentially significant with regard to identification between test and retest conditions.

The averaged envelope FFRs for the 100 ms /u/ vowel stimulus shown in Fig. 2-6 exhibit a clear pattern compared with the other three vowels observed. The spectral amplitude of F0 and H2 to H6 are clearly observed for most cases. Although some minor differences are visible, and the amplitude of H3 to H6 for subject 17 is difficult to visually observe, the intra-subject comparison between test and retest conditions generally reveals good matching.

The averaged spectral FFRs for the 100 ms /u/ vowel 85dB stimulus shown in Fig. 2-11 tend not to reveal a clear pattern for F1 (300 Hz) and F2 (900 Hz). At the same time, there is generally poor matching in intra-subject response, except for subject 22, who has strong representation of the peaks at F0 and H2 to H6 and good similarity between test and retest conditions.

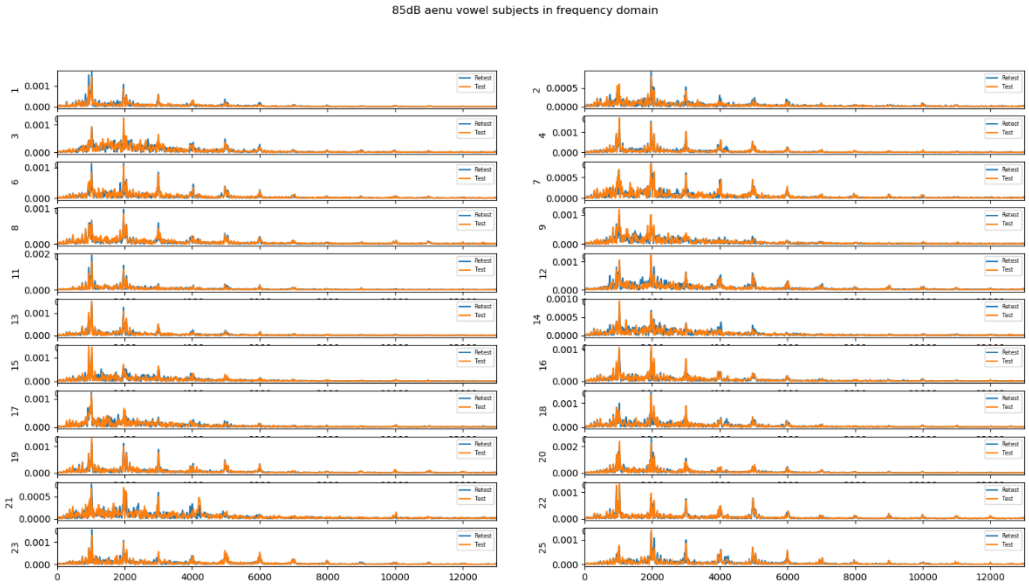

Figure 2-7 : Spectrum of frequency components of envelope FFRs averaged over 3000 trials for concatenated four vowel stimuli at 85 dBA

In Fig. 2-7, the spectral amplitude of four concatenated vowel envelope FFRs averaged over 3000 trials is shown. Here, the responses to the four vowels in the time domain are concatenated and then converted to amplitude spectra. It can be seen that the amplitude spectra contain features for all 4 vowels and may be more representative than individual vowels for differentiating between subjects.

Comparing with spectral amplitude envelope FFRs to single vowel stimuli, the waveform of vowel-concatenated stimuli covers all the frequency domain features but with much less effect from noise. Moreover, certain representative harmonics, such as H5, H6 and

H7, can be observed more clearly, which is evidence that they may be useful features for subject identification. At the same time, when observing waveform peaks in Fig. 2-7, harmonics for individual vowels are not always aligned when concatenating four vowels altogether. For example, two first formants can be clearly seen from concatenated vowel for subject 1 and subject 15. This may be because some of the responses to the fundamental and harmonics for a single vowel stimuli are left or right shifted. As a result, only picking one peak for each 100 Hz frequency bin as a representative feature is not appropriate for features of a concatenated responses signal, but instead the complete waveform of the amplitude spectrum may be more useful.

As the amplitude spectra of spectral FFRs for the four vowel stimuli do not appear to possess clear significance for intra-subject comparison, the concatenated 4 vowel waveform for spectral FFR will not be analyzed here in this work.

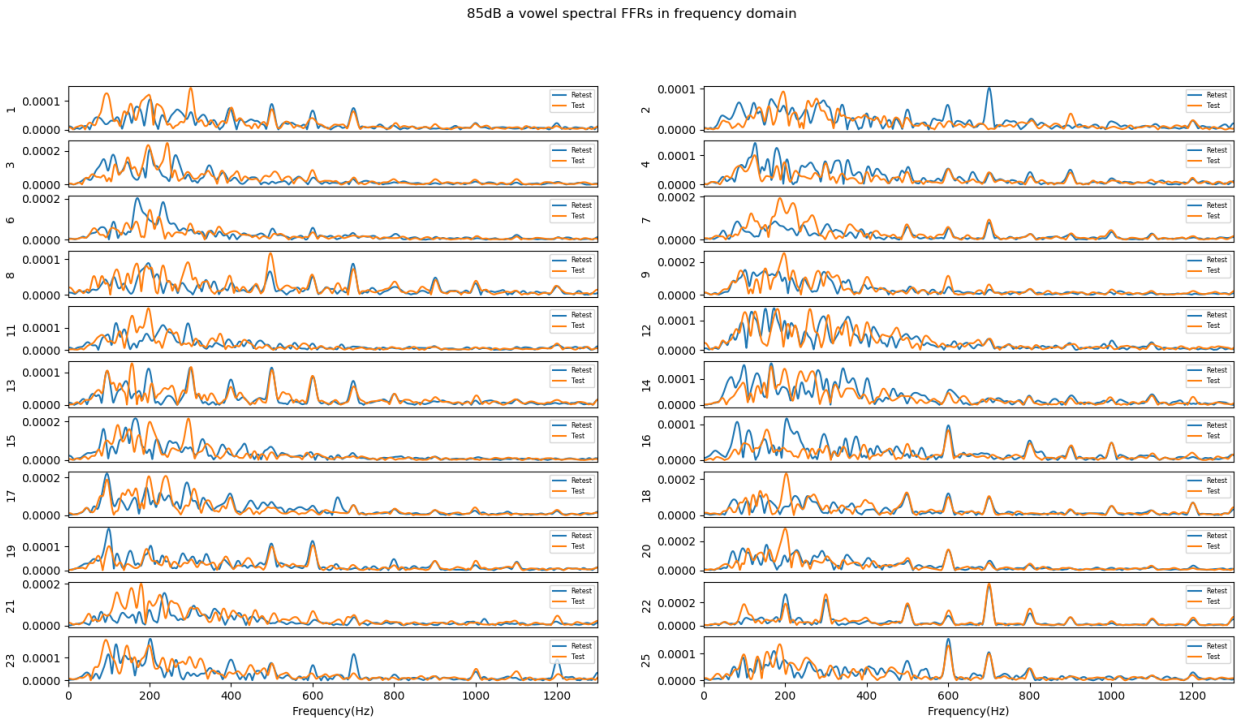

Figure 2-8 : Spectral FFR of test(orange) and retest(blue) conditions averaged over 3000 trials for a 100 ms /a/ vowel stimulus with F0=100 Hz presented at 85 dBA.

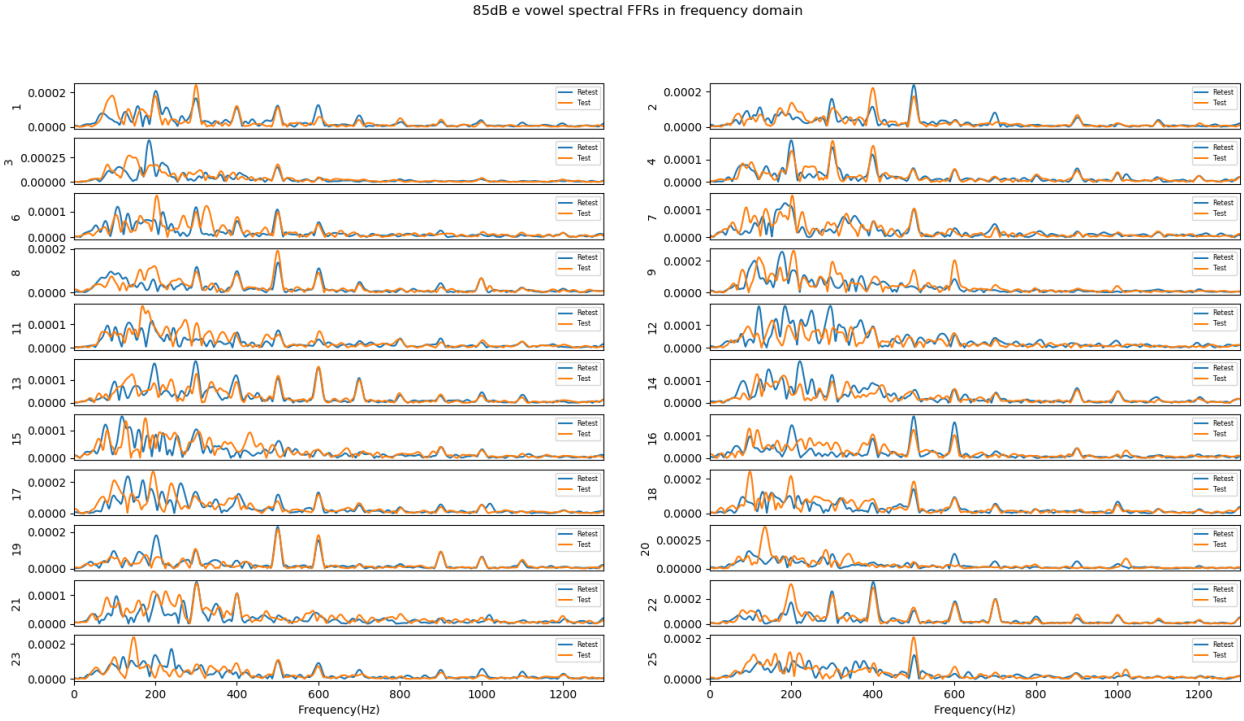

Figure 2-9 : Averaged spectral FFR of test(orange) and retest(blue) conditions averaged over 3000 trials for a 100 ms /ɔ/ vowel stimulus with F0=100 Hz presented at 85 dBA.

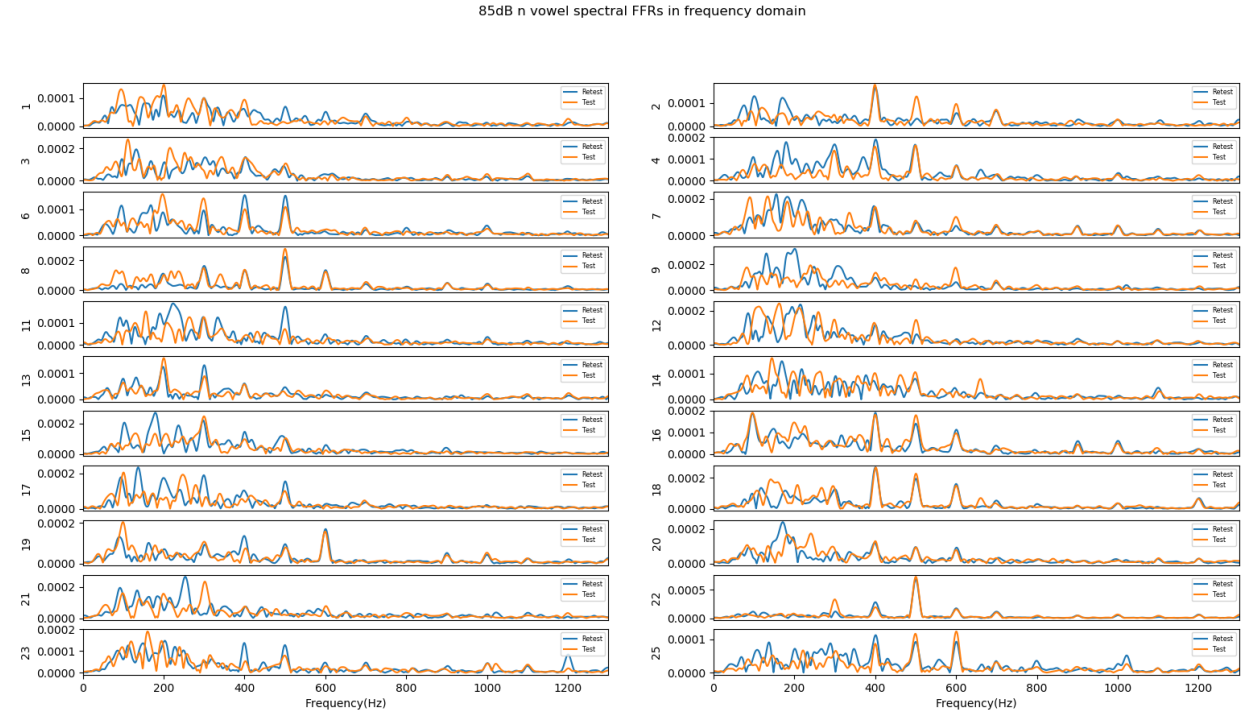

Figure 2-10 : Spectral FFR of test(orange) and retest(blue) conditions averaged over 3000 trials for a 100 ms / U / vowel stimulus with F0=100 Hz presented at 85 dBA.

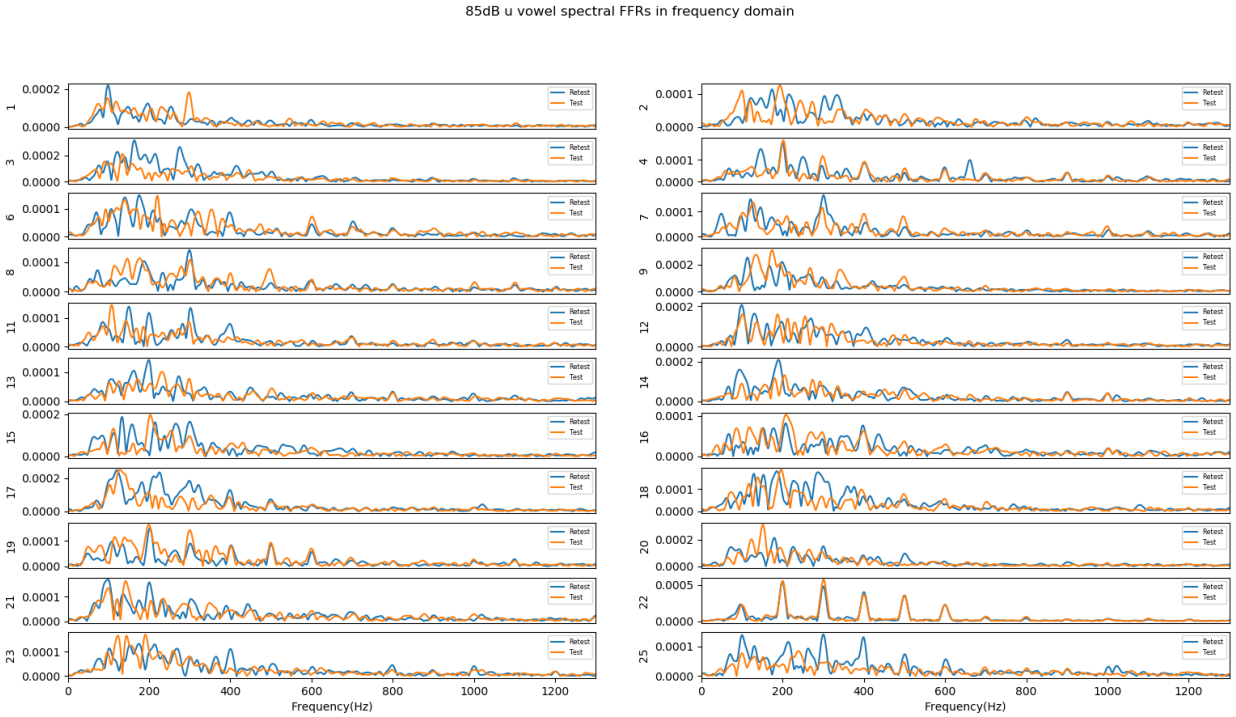

Figure 2-11 : Spectral FFR of test(orange) and retest(blue) conditions averaged over 3000 trials for a 100 ms /u/ vowel stimulus with F0=100 Hz presented at 85 dBA.

2.5 Conclusion

Comparison of the speech-evoked responses in the different subjects and across the four vowels showed that the differences between subjects can often be observed in both envelope and spectral FFRs.

The amplitude spectra of envelope FFRs reveal clear and robust encoding of F0 and H2 to H6 for all four vowels. Meanwhile, the amplitude spectra of spectral FFR reveal that the encoding of the F1 is more robust in comparison with that of F2 for /a/, /ɔ/ and /U/ vowel stimuli, although this is not clear for /u/ stimuli.

However, it appears that envelope FFRs could be expected to perform better than spectral FFR to produce features for subject identification, because they appear less noisy and less variable between the test and retest conditions.

The eFFR responses possessed visually obvious similarity within subjects, which will be investigated further in the following chapters. In general, the stability of eFFR indicates that it can be considered as a good candidate for biometric applications.

3 FFR Representation of Vowels in Different Subjects

3.1 Introduction

The third chapter aims to answer the following questions:

1. Does the FFR representation of four short vowels differ between normal hearing subjects?
2. Which features can maximize the differences between each normal hearing subject?
3. Are they stable across test and retest conditions?

Chapter 2 provided an overview of the relevant background information and a brief review of the auditory evoked potentials, along with relevant studies. A thorough characterization of the envelope and spectral FFR of normal hearing adults in both time and frequency domains was also conducted, which will be helpful for finding a set of suitable features in order to create a dataset for the machine learning study to be described later.

To verify the signal consistency and quality of the selected features, we make a comparison of the speech-evoked FFRs in normal hearing adults between the test and retest conditions. Two different methods were chosen for comparing the difference between subjects, the Pearson correlation coefficient and the Euclidean distance of amplitude spectra, which will be described in more detail below. The envelope FFRs in both time and frequency domain were selected based on the comparison results from chapter 2. Meanwhile, the features selected from the analysis in this chapter will be used for the machine learning study to be described later.

From Fig. 2-3 to Fig. 2-11, it is generally observed that differences in the four vowels between subjects can be identified in both time and frequency domain waveforms, which includes differences in the amplitude of the fundamental and harmonics, the waveform

between peaks, the latency etc. To further understand the differences and the suitability of these features, two comparison algorithms were chosen here. Therefore, this leads to a series of rigorous assessments of which components of the response would be selected to best identify and differentiate subjects.

In these assessments, both time and frequency features of envelope FFRs were considered. The spectral FFR was not considered as a candidate for subject identification due to the instability within and between subjects. The pairwise comparisons between test and retest conditions were performed among all twenty-two subjects with all four vowels.

Meanwhile, we summarized all the features into three aspects. The first is the evoked response to single vowel stimuli in either time or frequency domain. The second is the evoked response to multiple (all four) vowel stimulus in either time or frequency domain. The last aspect is a combination of both time and frequency domain features together. All three aspects will be tested so as to find the features that could maximize the difference between subjects.

In related studies (Sadeghian et al., 2015; Won et al., 2016; Yi et al., 2017), both the complete time domain series including transient evoked response along with the steady state response, and the complete amplitude spectra from 0 to 800 Hz along with the spectral features including fundamental frequency with H2 to H6, were selected for evaluation.

Prior to evaluating machine learning classification, in order to ensure the quality of the signal it is worthy to consider the results of the analysis of the Pearson correlation coefficients and Euclidean distances. In order to test and verify the quality of the preprocessed signal, the signal collected from the same day and same subject is tested in this step. Three validation methods are proposed here, namely, spectral flatness, Pearson correlation coefficient and the Peak Noise Ratio (PNR).

It is worth emphasizing here that it may be possible to further improve the performance of the identification system if a subject with poor signal quality can be detected early, and either excluded from the classification or asked to provide additional response samples for analysis.

3.2 Test – Retest Comparison

Pairwise comparisons were conducted between the test and retest conditions of not only the same subject but also different subjects with the same vowel stimuli. The experimental designs for the two methods are nearly identical, and the difference in logic of the two methods is discussed.

3.2.1 Comparisons based on the Pearson Correlation Coefficient

As the Pearson Correlation Coefficient (PCC) is considered as a criterion for evaluating the similarity of various signals, it is therefore considered for testing the similarity of speech evoked responses between test and retest conditions, which were collected from the same subject with the same vowel stimulus on two separate days (Lee Rodgers & Alan Nice Wander, 1988).

In the time domain, all the signals are detrended by removing the DC offset as it is a potential source of distortion when calculating the PCC. A Hamming window function was subsequently applied after DC removal. Two sub-average evoked responses collected on one day (each based on the coherent average of 1500 responses) are merged here through averaging of sums in order to reduce the effect of noise. The Pearson Correlation Coefficient is then used to calculate the similarity between the test and retest signals using the function shown below.

Here, σ_x and σ_y are the standard deviations of signal x and signal y. Meanwhile, μ_X is the mean of X and μ_Y is the mean of Y.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_x \sigma_y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_x \sigma_y} \tag{3.2.1}$$

The absolute value of the correlation coefficient is less than or equals to 1, which means that it can take the range of values from -1 to 1. A PCC equal to 1 corresponds to two perfectly positively correlated signals, while a PCC equal to -1 corresponds to perfectly negatively correlated signals.

Therefore, the PCC value between all the test and retest condition responses was calculated and further gathered for generating a correlation matrix in order to facilitate comparisons within same subjects and between different subjects. This correlation matrix for the time domain signal is generated for all four vowel stimuli.

For each test subject, the PCC value of this subject is chosen for making comparisons with that of all the other retest subjects. This allows the PCC values to be ranked and the top-1 can be picked out and highlighted in this matrix. Therefore, for each test subject, there is always a retest subject picked out and highlighted as the top-1 for that test subject. If top-1 retest subject is found to be the same as the test subject, then we can conclude that the PCC allowed identification of the subject in the retest condition. Otherwise, there is misidentification.

For the frequency domain signal, zero padding is applied as it is helpful for increasing the precision in the frequency domain. Therefore, a series of zeros is added to the end of the time domain signal, and then the Discrete Fourier transform is calculated. The Pearson Correlation Coefficient is then calculated based on the amplitude spectra series from 0 to 800

Hz. Another set of correlation matrices is generated by using only peaks from the fundamental frequency and H2 to H6, while ignoring the signal between the peaks.

An extra set of correlation matrices was generated when the evoked responses from all four-vowel stimulus in time or frequency domain are concatenated, so as to evaluate the assumption that it can help to improve the accuracy for subject identification in either time or frequency domain.

The last set of matrices was generated based on the correlation results which use both time and frequency domain features together. Two strategies were used for combining features from time and frequency domains. The first one is adding the time and frequency domain correlation matrices directly element-wise and the second strategy is to concatenate the signal in both time and frequency domain across the 4 vowels first and then the PCC is calculated using the new concatenated signal series. The various correlation matrices described in this section are shown below.

Improvement for PCC when concatenating the time and frequency domain signals

It is necessary to find solutions to deal with the difference between time and frequency domains when concatenating them together. Assume that both time and frequency domain series were treated as the same kind of parameter in order to use them together. The time and frequency domain signals should have the same total energy. Based on the Parseval's theorem, the sum of the square of a function is equal to the sum of the square of its transform. For the Discrete Fourier transform, Parseval's theorem is often written as:

$$\sum_{n=-\infty}^{\infty} |x[n]|^2 = \frac{1}{N}\sum_{k=0}^{N-1} |x[k]|^2 \qquad (3.2.2)$$

Here, $x[k]$ is the Discrete Fourier transform of $x[n]$, with both being of length N. Based on Parseval's theorem, it is found that the time and frequency domain series can therefore be normalized with respect to their own energy so as to balance their weight for the correlation. In such a case, time and frequency domain series can therefore be concatenated together with same weight.

Improved Pearson Correlation Coefficient with respect to amplitude spectra

Another improvement to the features of amplitude spectra is proposed here in order to further improve the PCC. Previously, the PCC score is used for measuring the linear relation of two input signal series. As the values of amplitude spectrum are all positive, which means that μ_X and μ_Y in the correlation coefficient function are always positive, this may not be helpful in certain cases. Therefore, it is worthwhile to consider removing the means as parameters in the correlation function, as indicated in eq. 3.2.3.

$$\rho_{X,Y} = \frac{E[XY]}{\sigma_x \sigma_y} \tag{3.2.3}$$

Here, σ_x and σ_y are the standard deviations of x and y respectively. In order to test this assumption, a series of example comparisons were performed in order to test the effect of not removing the mean value of the signal series. Fig. 3-1 shown below illustrates with a simple example why it may be clearly preferable to remove the mean before computing the normalized correlation between all-positive signals.

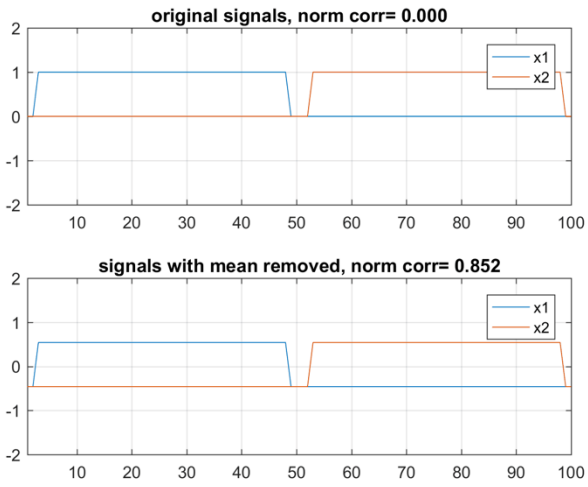

Figure 3-1 : Comparison of PCC obtained when means are not removed and when they are

It can be seen from the figure that the mean components in the original signals have a negative effect on the PCC. When the mean components are removed, as shown in the bottom figure, the correlation coefficient increased from 0 to 0.852.

Beyond this simple example, in the real case of our amplitude spectra, the peaks are narrower and have much less impact on the total mean of the signal. Below, real cases from our dataset are used to test the impact of removing the mean.

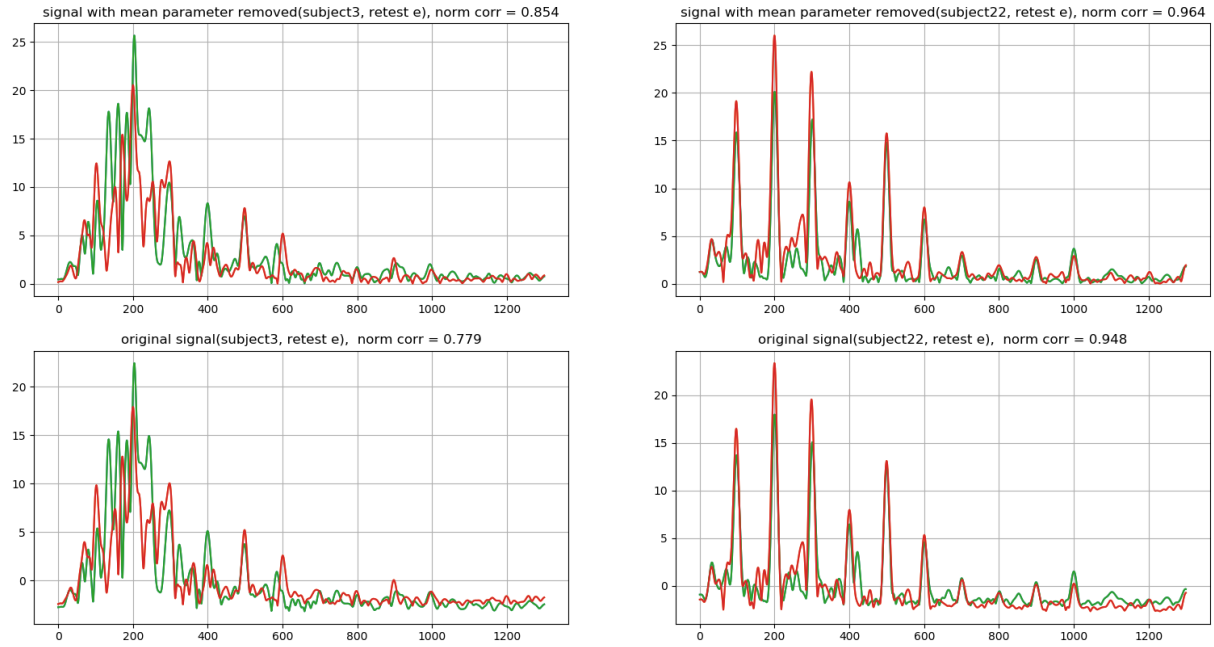

Figure 3-2 : Comparison of correlation coefficient for subject 3 and 22 in frequency domain with and without removing the mean prior to calculating the PCC

Those two examples, subject 3 and subject 22, are two representative signals with low and high peak to noise ratio. For subject 3, the original PCC is 0.854 and it decreases to 0.779 when the mean component is removed. For subject 22, the PCC changed slightly from 0.964 to 0.948 when the mean component is removed. It can be found that the more noisy (less peaky) the signal, the bigger the decrease in the PCC. Hence, it appears preferable to avoid subtracting the mean value.

The correlation results obtained with removing the mean for the frequency domain signals and for the concatenation of the time and frequency domain signals is shown in section 3.3.1.

3.2.2 Comparison based on the Euclidean distance

In mathematics, the Euclidean distance is the straight distance between two points in Euclidean space. As the frequency spectrum of the speech-evoked response is a series of complex numbers, the difference of two complex spectra can be considered as the Euclidean distance between two two-dimension parameters when taking the real part and imaginary parts of the complex spectrum as the first and second dimensions of the parameter. Hence, the distance between the two frequency spectra of the response was calculated as follows.

Firstly, the DC offset is removed from the time domain evoked response series in order to avoid distortion. Two sub-averages (each corresponding to the coherent average of 1500 responses) are combined together (averaged) in order to further reduce the effect of noise. The Hamming window is then applied so as to reduce spectral leakage after calculating the Discrete Fourier transform.

Normalizing the time domain signal is necessary for calculating the Euclidean distance when signals may differ in scale. Here, the root mean square of the series is used for the process of normalization.

$$\frac{x(t)}{\sigma = \sqrt{\frac{\sum |x(t)|^2}{n}}} \tag{3.2.4}$$

Zero padding is applied so as to smooth the frequency domain signal. Then, a series of complex values is obtained through the FFT of the time domain signal. Here, the complete spectrum series from 0 to 1300 Hz is selected for calculating the Euclidean distance using the eq. (3.2.5).

$$\overline{\boldsymbol{pq}} = d(\boldsymbol{p}, \boldsymbol{q}) = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2} \tag{3.2.5}$$

Here, $\mathbf{p} = (p_1, p_2, p_3, \dots, p_n)$ and $\mathbf{q} = (q_1, q_2, q_3, \dots, q_n)$ are two spectrum series with n elements respectively, with $p_i = R_{p_i} + \mathrm{j}\, I_{p_i}$ and $q_i = R_{q_i} + \mathrm{j}\, I_{q_i}$.

For each comparison between two sets of signals (two complex numbers series), a series of distance value is calculated based on the function provided above. All the values in this list are added together and the square root of the sum is used for representing the Euclidean distance between two signals. With each pair of comparisons generating a distance score, a distance matrix was then generated by using all the distance scores calculated.

A similar method is proposed here as the sum of absolute value of each spectrum series from 0 to 1300 Hz in equation 3.2.6.

$$\overline{\boldsymbol{pq}} = d(\boldsymbol{p}, \boldsymbol{q}) = \sum_{k=1}^{n} |p_k - q_k| \tag{3.2.6}$$

Here, the $\mathbf{p}$, $\mathbf{q}$, p_i, and q_i have the same definition with the definition of Euclidean distance parameter.

A distance matrix was therefore generated based on the modified Euclidean score for each vowel. In order to combine the information from 4 different vowels, a new matrix obtained by concatenating the responses to the 4 vowels is obtained. Four series of complex values, representing all four different vowel stimuli are concatenated together to generate a new series of complex values. Then, the process for calculating the distance score was repeated and a new distance matrix was generated through the new score.

3.3 Results

The overall results based on the PCC are presented below in Table 3.1. Results from specific cases are presented in the figures that follow, starting with the frequency domain analysis, including the analysis of several obvious outliers, and then followed by the time domain analysis. The results based on Euclidean distance are presented afterwards.

3.3.1 Results based on the Pearson Correlation Coefficient

The obtained accuracy based on the Pearson correlation coefficient matrix with the four vowel stimulus is shown in Table 3.1, with the four columns showing the feature, vowel stimulus, the obtained accuracy based on the corresponding correlation matrix, and other comments, respectively.

Table 3.1 : Comparison of obtained accuracy for envelope FFRs including time domain, frequency domain and the combination of the two. The term 'AS7' in the table corresponds to the sequence when only picking the peak values of F0 and H2 to H6. The term 'efr' refers to the eFFR waveform. The term 'aenu_as' refers to the amplitude spectra of the concatenated four-vowel eFFR waveform and the term 'as-aenu' refers to the concatenated four-vowel amplitude spectra of eFFR.

Feature	Vowel Stimuli	Accuracy	Comments
Time	/a/	63.63%	14/22
Time	/ɔ/	59.09%	13/22
Time	/U/	54.54%	12/22
Time	/u/	50.00%	11/22
Time	concatenate 4	72.72%	16/22(4096 in length)
Time	concatenate 4	72.72%	16/22, sum of 4 correlation matrices
Frequency	/a/	50.00%	11/22, amplitude spectrum
Frequency	/a/	59.09%	13/22, AS7 for fundamental and harmonics
Frequency	/ɔ/	40.91%	9/22, amplitude spectrum
Frequency	/ɔ/	36.36%	8/22, AS7 for fundamental and harmonics
Frequency	/U/	59.09%	13/22, amplitude spectrum
Frequency	/U/	31.82%	7/22, AS7 for fundamental and harmonics
Frequency	/u/	40.91%	9/22, amplitude spectrum
Frequency	/u/	36.36%	8/22, AS7 for fundamental and harmonics
Frequency	Concatenate 4	68.18%	15/22, sum of 4 correlation matrices
Frequency	Concatenate 4	68.18%	15/22, concatenate and convert to freq. domain (535 point response to 1300 Hz)
Frequency	Concatenate 4	63.63%	14/22, concatenate four amplitude spectra, 4*139=556 point, without zero padding
Frequency	Concatenate 4	36.36%	8/22, AS7, concatenate and convert to freq. domain (7 points)
Frequency	Concatenate 4	68.18%	15/22, AS7, concatenate four sets of peaks (28 points)
Time & Freq	Concatenate 4	81.82%	18/22, sum of two correlation matrices
Time & Freq	Concatenate 4	72.72%	16/22, concatenate efr and aenu_as
Time & Freq	Concatenate 4	72.72%	16/22, concatenate efr and as_aenu_

The table clearly shows that the comparison accuracy obtained using the PCC with concatenated vowel stimuli is always higher than the ones obtained with a single vowel stimuli in both time and frequency domains. The time domain signal includes all the information of transient response and steady state response whereas the frequency domain only includes the spectrum from 0 to 1300 Hz. Tests were conducted to compare the accuracy obtained with spectra between 0 to 700 Hz, which include F0, H2 to H6, from 0 to 1000 Hz, and from 0 to 1300 Hz. The results show that the performance with spectra from 0 to 1300 Hz performs best and hence it is included in the table. The best result of 81.82% obtained until now is from the correlation matrix when combining both time and frequency domain features of the concatenated vowels, with 18 subjects obtaining a match with their own test and retest condition signals. Furthermore, combining both time and frequency domain signals results in higher performance than single time or frequency domain results (18/22 versus 15/22 and 16/22).

Several representative correlation matrices with time and frequency domain signals are shown below along with analysis for some single subjects. The value of the PCC is reflected in the colour of the squares on the left side matrix and the number shown on each square is the rank of the PCC of the cell in a given row. These ranks are produced by comparing each retest subject signal (indicated on the horizontal axis) with all the test subject signals (indicated on the vertical axis). The best match (i.e. the square that is ranked number 1) is highlighted on the right-side figure. Note that in this
figure, 100% accuracy would correspond to only cells along the diagonal being ranked number 1.

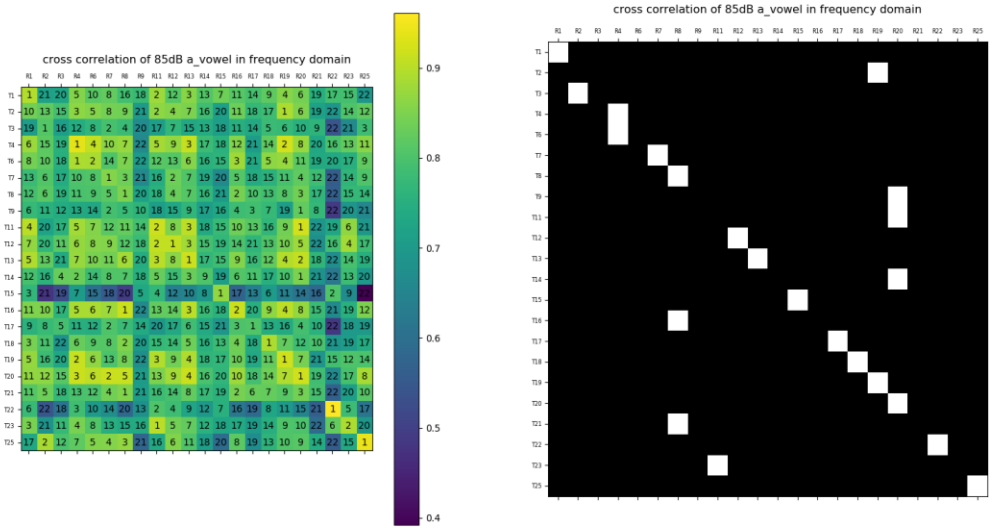

Figure 3-3 : Pearson correlation matrix for envelope FFR of 85dB /a/ vowel stimuli in frequency domain

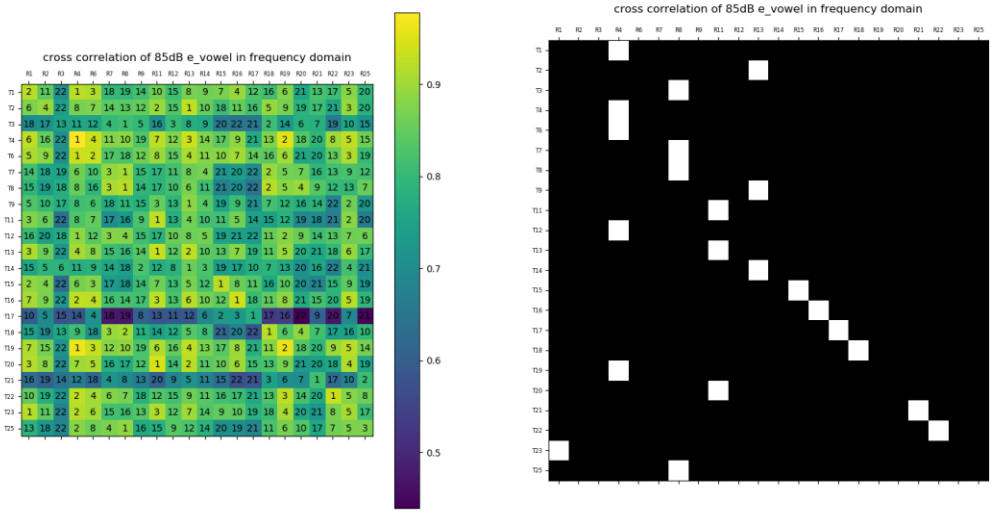

Figure 3-4: Pearson correlation matrix for envelope FFR of 85dB /ɔ/ vowel stimuli in frequency domain

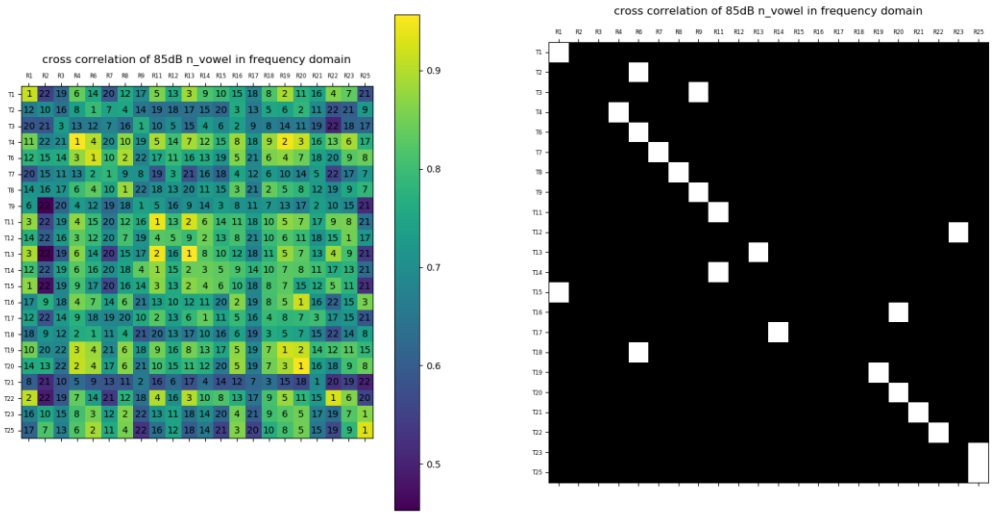

Figure 3-5 : Pearson correlation matrix for envelope FFR of 85dB /U/ vowel stimuli in frequency domain

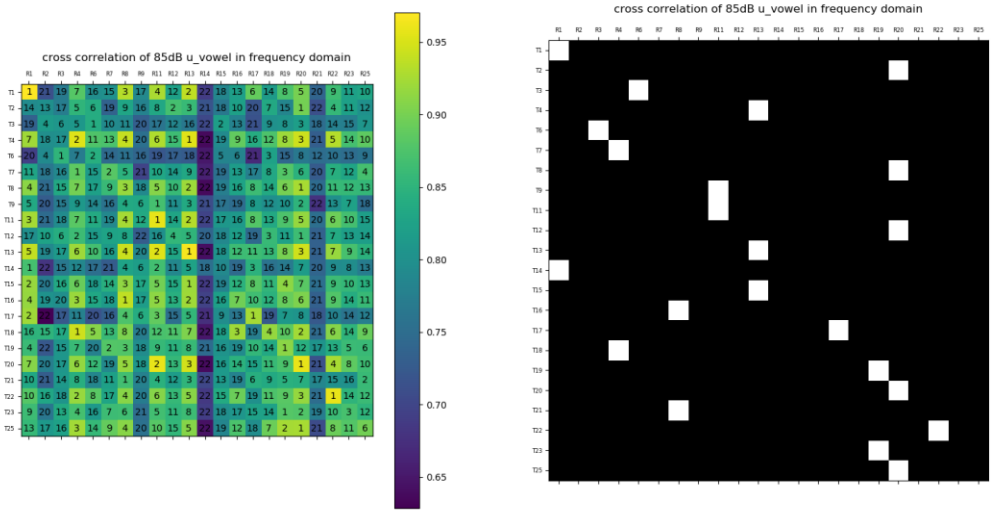

Figure 3-6 : Pearson correlation matrix for envelope FFR of 85dB /u/ vowel stimuli in frequency domain

With respect to the correlation matrix of 85dB 4 vowel signal in frequency domain, the best result shown above is derived from the 0 to 1300 Hz spectrum. For the best accuracy, it is shown that /a/ and /U/ vowels have a better result compared with /ɔ/ and /u/ vowels. The /U/ and /u/ vowels have a relative higher stability between test and retest compared with /a/ and /ɔ/ vowels, as the color of their top-1 matches on the diagonal are much brighter than the rest.

For the values of PCC, it is obvious that the diagonal for correlation matrix of /u/ vowel has the highest scores and obviously higher than the scores of squares not on the diagonal, which can be visualized through the color of the squares on the left side correlation matrix from Figs. 3-3 to 3-6. In contrast, the PCC value on the diagonal of /a/ vowel has the lowest difference compared to values not on the diagonal.

Concatenated four vowels in frequency domain

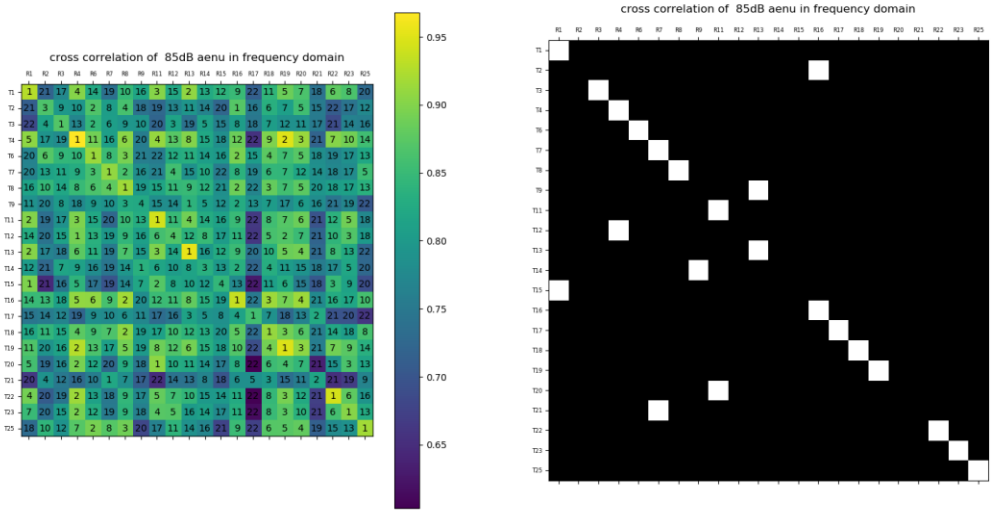

Figure 3-7 : Pearson correlation matrix for envelope FFR of concatenated responses to four vowel stimuli in frequency domain

For the concatenated signal, it is easy to see that the accuracy of the correlation matrix is higher than that of single vowels, as the stability of the top-1 on the diagonal is much better. Here, a 15 of 22 accuracy is achieved.

Analysis of the mistaken identification

The figures below show the frequency domain signals of the 7 subjects who were mistakenly identified (subject 2, 9, 12, 14, 15, 20 and 21). The plots on the left show their test spectrum and their own retest spectrum (which was not ranked #1 in terms of the PCC value), while plots on the right show their test spectrum and the retest spectrum which achieved a PCC value that was ranked #1.

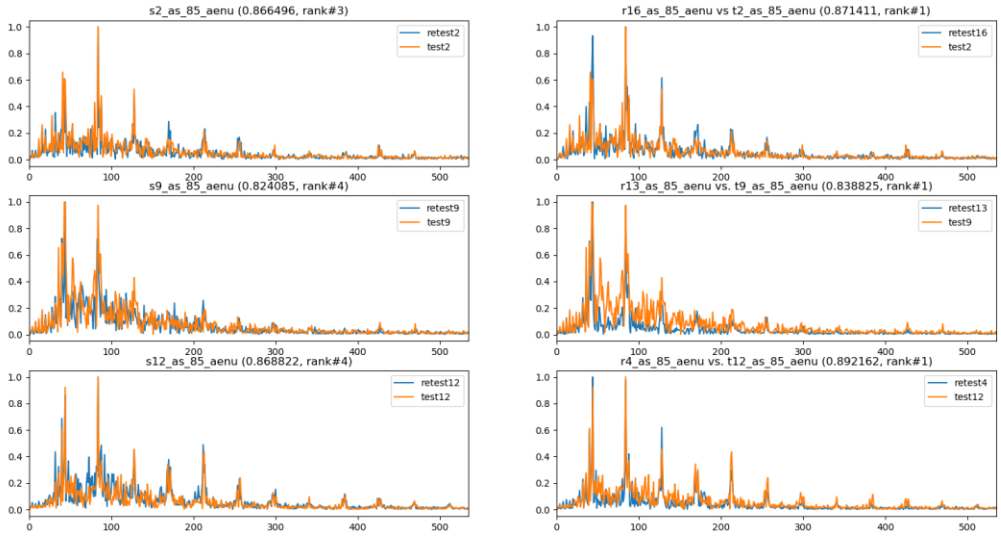

Figure 3-8 : Amplitude spectrum comparison between test and retest conditions for subject 2, subject 9, and subject 12

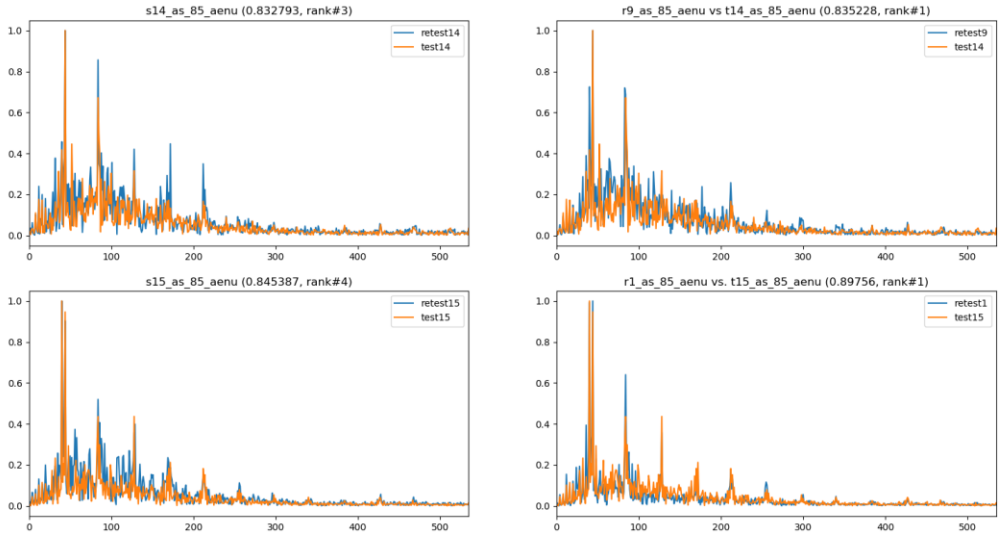

Figure 3-9: Amplitude spectrum comparison between test and retest conditions for subject 14 and subject 15

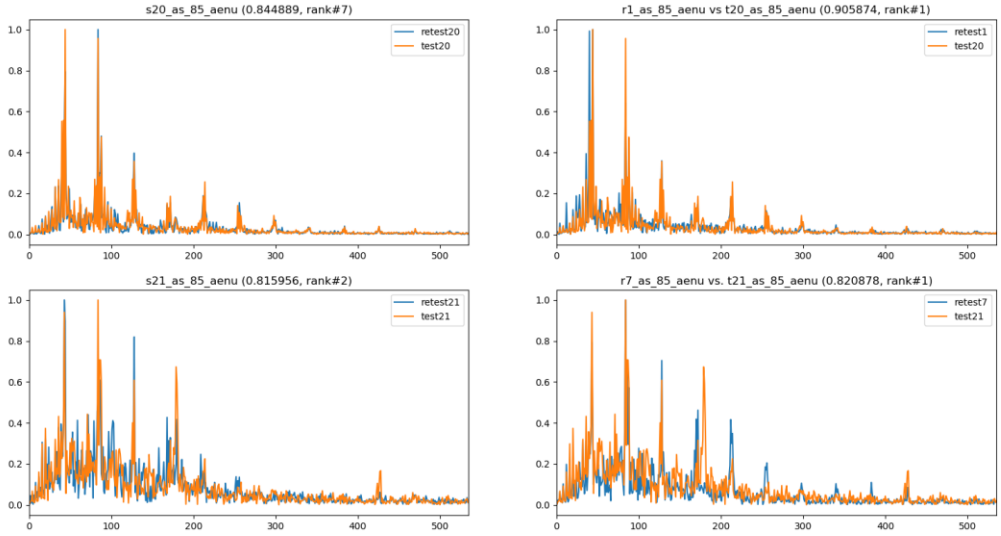

Figure 3-10: Amplitude spectrum comparison between test and retest conditions for subject 20 and subject 21

For subject 2, the most obvious difference between retest #2 and test #2 (PCC score 0.866, rank #3) is at 400 Hz peak, followed by the 100 to 200 Hz peaks. However, for retest #6 and test #2 (PCC score 0.871, rank #1), some difference is observed at 100, 300 and 400 Hz peaks but much less difference in the signal between peaks.

For subject 9, the harmonics between 100 to 300 Hz on retest #9 and test #9 (PCC score 0.824, rank #4) shows a great difference, as well as the peaks at 400 Hz and at 500 Hz. The same case is also shown on retest #13 and test #9 (PCC score 0.839, rank #1), when the signal is observed between 100 to 300 Hz.

The harmonics show a slight difference near the 100 Hz peaks between retest #12 and test #12 (PCC score 0.869, rank #4). In comparison, the frequency domain signal of retest #4 and test #12 (PCC score 0.892, rank #1) only has slight difference at 100 Hz and 300 Hz peaks.

For subject 14, peaks at 200, 300, 400 and 500 Hz do not match well between retest #14 and test #14 (PCC score 0.833, rank #3). The signal between peaks in the 0 to 400 Hz frequency range show a visible difference especially near the 100 Hz peak, and furthermore peaks at 600, 700, 800 Hz and higher are hard to be recognized. When making comparison between retest #9 and test #14 (PCC score 0.835, rank #1), an extra peak at 100 Hz is shown on retest #9 because of the peak shift with a single vowel. Meanwhile, the difference of value at the second harmonic between test #14 and retest #9 is smaller than that between retest #14 and test #14. Peaks at 300, 400 and 500 Hz and the signal between peaks from 100 to 300 Hz show obvious differences.

For subject 15, peaks at 100, 300 and 500 Hz match well between retest #15 and test #15 (PCC score 0.845, ranks #4). However, the signal between 100 to 300 Hz shows great differences. In contrast, when making comparison between retest #1 and test #15 (PCC score

0.898, rank #1), only a difference of peak value at 200 and 600 Hz can be clearly seen, as well as in the signal between 200 and 300 Hz.

For subject 20, when comparing retest #20 and test #20 (PCC score 0.845, rank #7), three peaks are seen at 100 Hz for test #20. In contrast, retest #20 only has one peak. For retest #1 and test #20 (PCC score 0.906, rank #1), a clear difference is seen at 100 Hz peak. Meanwhile, signals near 100 Hz also have low similarity between retest #1 and test #20. Harmonics on and above 600 Hz are relatively hard to recognize as they have values similar to the peaks near the harmonics.

The last matching is done between retest #21 and test #21 (PCC core 0.816, rank #2), which has a clear difference at 300 Hz peak, with the amplitude of retest #21 being higher than that of test #21. Meanwhile, the signal between peaks in the 100 to 400 Hz frequency range also shows clear differences. Peaks at 700, 800 and 900 Hz can hardly be seen for both of them. In comparison, retest #7 and test #21 (PCC score 0.821, rank #1) has more obvious differences, especially at the peaks. The fourth harmonic of test #21 is significantly right shifted to 420 Hz and the peak at 500 Hz of test #21 has lower value than for retest #7, which is also seen with peaks 600, 800 and 900 Hz. It is interesting to note that the 10^{th} harmonic at 1000 Hz for test #21 has a higher value than the 500 to 900 Hz peaks, which can be considered a significant pattern for test #21.

In conclusion, the amplitude of an harmonic at a certain frequency usually shows great significance for subject identification, especially from the first to the seventh harmonic, as the value of their peaks is almost always much higher than other signal components. Multiple peaks at signal harmonics should be considered as the signal is derived from the concatenation of several vowels and their harmonics may not always be aligned. Meanwhile, relying only on those peaks is not a good strategy, as the value of the harmonics does not always match between retest and test. It may be more meaningful to rely on 10 or 11 peaks in the range of 0

to 1300 Hz. However, at this stage, it is hard to predict which peaks are most useful for subject identification.

Correlation Matrix of concatenated responses to four vowels in the time domain

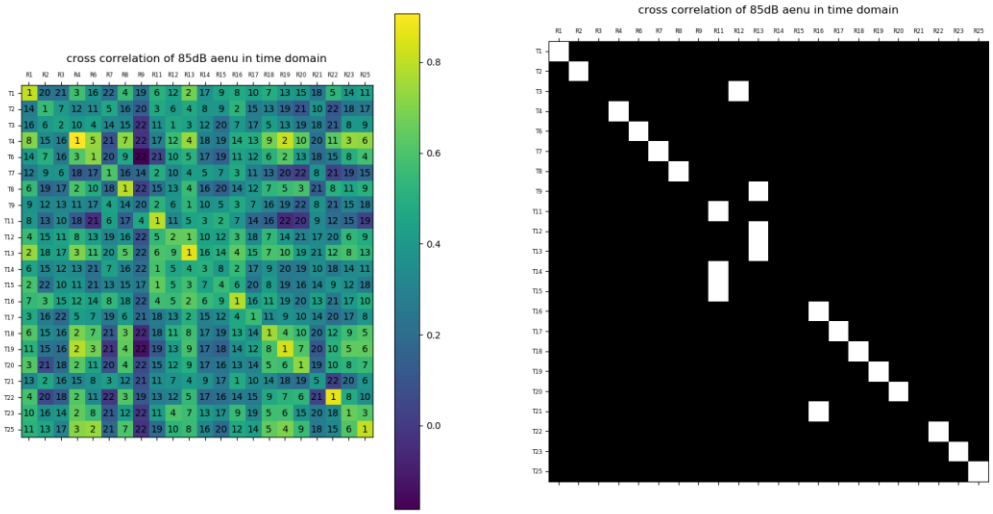

Figure 3-11 : Correlation matrix for concatenated time domain responses of the 4 vowels

Here, the complete time series signal for the concatenated responses to the four vowels, which includes transient-state responses and steady-state responses, is used for determining the Pearson correlation coefficient. It can be seen that with this signal, a relatively higher accurate result of 16 out of 22 is obtained (Fig. 3-11).

Correlation matrix of combinations of the concatenated time and frequency domain responses

After generating the correlation matrices for time and frequency domains separately, it is worthy to consider using both time and frequency domain information to generate a new matrix that may help further improve the accuracy. There are several options to make use of both time and frequency domain information when we consider generating a new matrix. The

first and simplest logic is to add two matrices together. A second approach is concatenating the two signals series together and then generating a new matrix.

For the first approach, the signal series with the responses to all four vowels in time and frequency domains are concatenated separately, generating two correlation matrices. Then, those matrices are added together element-wise in order to generate a new matrix, as the score in the same position represents the same subject comparison. The result is shown below. This combination gave an identification accuracy of 18 out of 22, which is one of the highest accuracy results achieved.

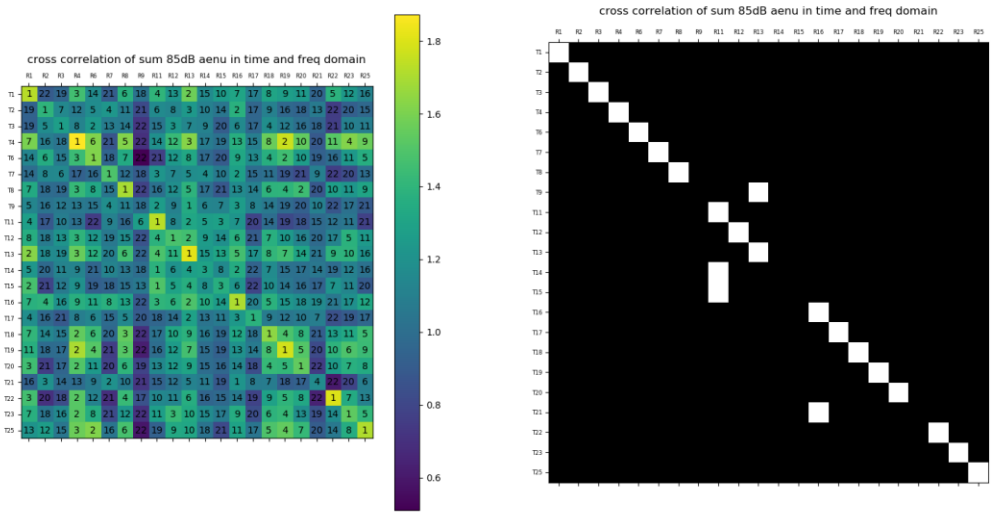

Figure 3-12 : Pearson correlation matrix obtained by adding time and frequency domain matrices for the concatenated responses of the 4 vowels

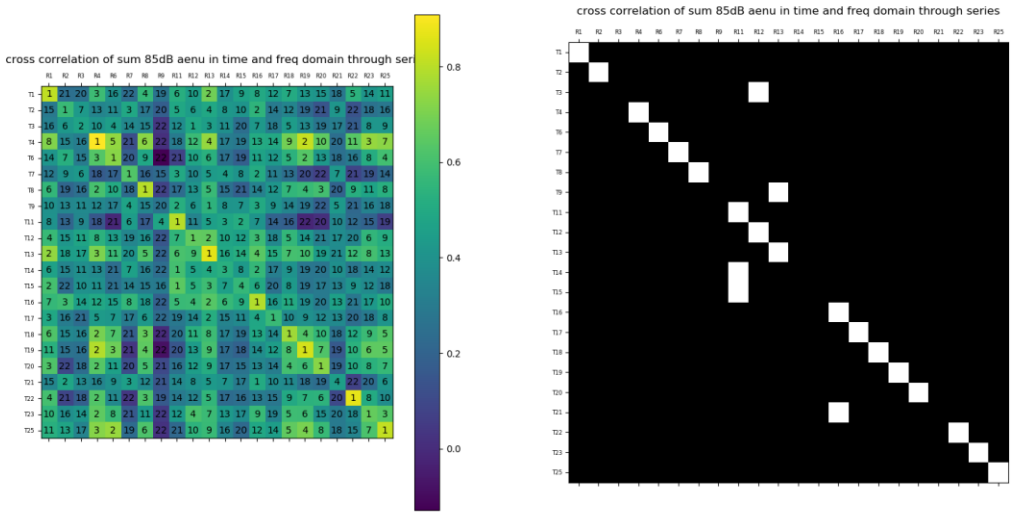

Figure 3-13 : Pearson correlation matrix obtained by concatenating time and frequency domain signals for the 4 vowels.

For the figure above, the matrix is generated based on the concatenation of time and frequency domain signals. The complete series in the time domain for 4 vowels and the 0 to 1300 Hz frequency domain series for 4 vowels is used. Here, the time domain signal was obtained by concatenating the responses to the four vowels, and the frequency domain signal was restricted to the range 0 to 1300 Hz (time domain signals concatenated first and then converted to frequency domain). Because the time and frequency domains have different scales, the signals were normalized first so that each of them accounts for 50% of the weight (i.e., power in the resulting signal). The outcome of this operation was an identification accuracy of 17 out of 22.

Improved Pearson Correlation Coefficients with removing mean component of the signals

The results of the improved Pearson correlation coefficients obtained with removed mean component are shown in Table 3.2, with feature type, vowel stimuli, comparison matrix accuracy, and comments displayed, respectively. Only the results of features related to frequency domain are affected by mean removal, and hence are shown here.

Table 3.2: Results with respect to the PCC with mean removed

Feature	Vowel Stimuli	Accuracy	Comments
Frequency	/a/	63.63%	14/22, amplitude spectrum (from 0 to 1300 Hz)
Frequency	/ɔ/	45.45%	10/22, amplitude spectrum (from 0 to 1300 Hz)
Frequency	/U/	59.09%	13/22, amplitude spectrum (from 0 to 1300 Hz)
Frequency	/u/	40.91%	9/22, amplitude spectrum (from 0 to 1300 Hz)
Frequency	concatenate 4	81.82%	18/22, concatenated four spectra and calculated PCC with mean removed
Time & Freq	concatenate 4	81.82%	18/22, concatenated time and frequency domain signals and calculated PCC with mean removed (time domain for PCC and frequency domain for PCC with mean removed)

As seen from Table 3.2, clear differences of accuracy between the previously obtained PCC and the PCC with mean removed obtained with mean component removed are observed. For single vowel stimuli, two out of four correlation matrices exhibit an obvious increase, with the comparison accuracy of /a/ vowel ascending from 50% (11/22) to 63.63% (14/22) and that of /ɔ/ vowel increasing from 40.91% (9/22) to 45.45% (10/22). The remaining two vowels, /U/ and /u/, do not experience a change in accuracy, with 13/22 and 9/22 respectively. For the result obtained with the concatenated four-vowel responses, the accuracy also climbed from 63.63% (14/22) to 81.81% (18/22), which equals the accuracy obtained by concatenating time and frequency domain responses. The concatenated vowel correlation matrices are shown below with some explanations.

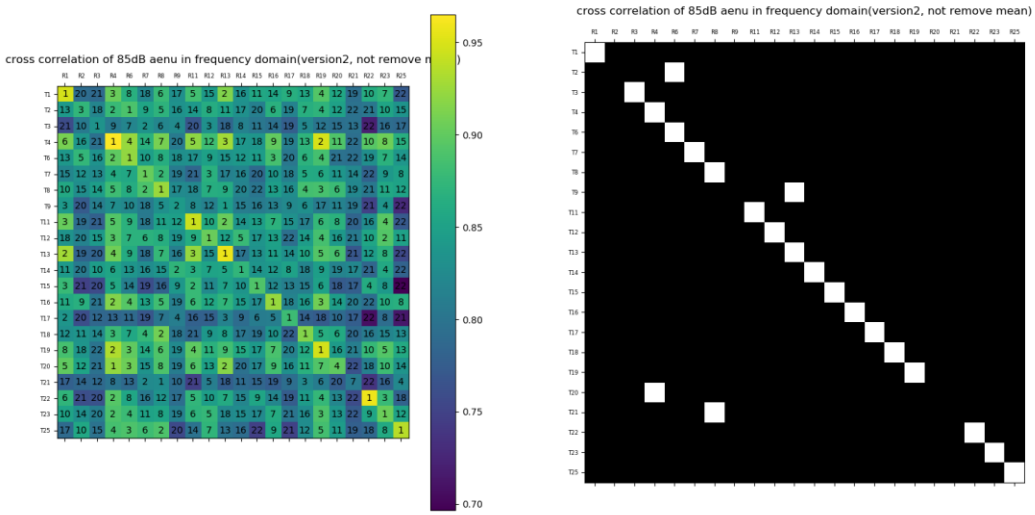

Figure 3-14: Pearson Correlation matrix for concatenated frequency domain responses of the 4 vowels (with mean removed)

The result with PCC with mean removed for concatenated four vowel responses in frequency domain is shown in Fig. 3-14. Here, only subjects 2, 9, 20 and 21 do not match between test and retest. In comparison, the subjects 12, 14 and 15, which were previously not on the diagonal, are now correctly identified.

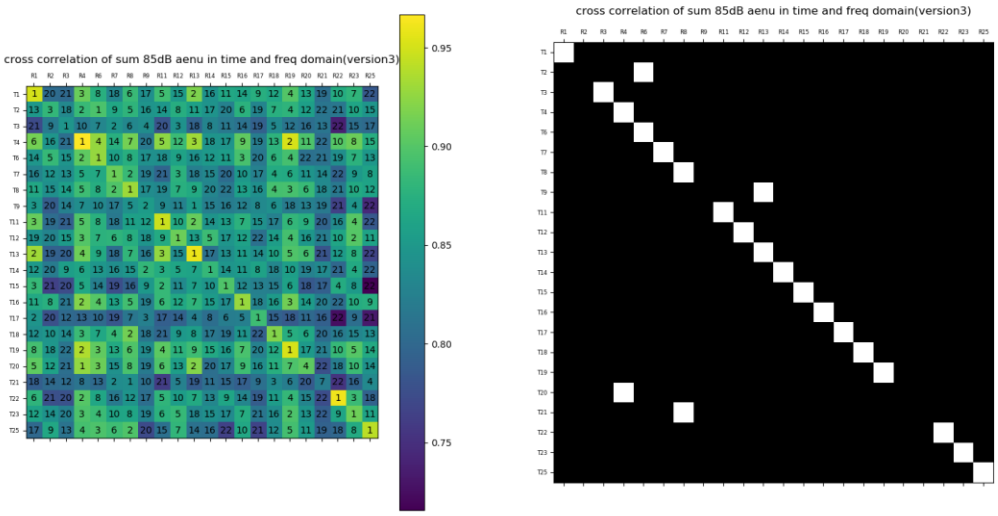

Figure 3-15: Correlation matrix for concatenate time and frequency domain signals (mean removed)

Fig. 3-15 shows the result of the PCC with mean removed correlation matrix with the combination of time and frequency domain signals with mean removed. As before, the time and frequency domain signal series are generated separately. In order to make sure that both series have the same weight, the energy of each signal is used for normalizing each of the two series. The final series is obtained by concatenating the two normalized signal series together, and then the correlation matrix is generated. Note that the mean component for the time domain series is kept (although it may be very small so it does not need to be removed) and it is removed only for the frequency domain series. Overall, the obtained accuracy was relatively high (81.82%) but was not an improvement over previously obtained accuracy.

3.3.2 Comparisons using Euclidean Distance

Figs. 3.13, 3.14, 3.15 and 3.16 show the distance matrices derived from Euclidean distance between test and retest frequency domain signals. The black cell in the matrices is the one which has the shortest distance between itself (test subject) and the retest subjects. This is the opposite of the previously shown correlation matrices, where the white cells corresponded to the highest correlations.

Table 3.3: **Results obtained using Euclidean distance matrices and modified distance matrices for evoked FFRs with frequency domain signals from responses to individual vowels and from concatenation of the four responses**

Feature	Vowel Stimuli	Accuracy	Comments
Frequency	/a/	59.09%	13/22, sum of absolute value element-wise
Frequency	/a/	63.64%	14/22, Euclidean distance
Frequency	/ɔ/	63.64%	14/22, sum of absolute value element-wise
Frequency	/ɔ/	59.09%	13/22, Euclidean distance
Frequency	/U/	54.55%	12/22, sum of absolute value element-wise
Frequency	/U/	59.09%	13/22, Euclidean distance
Frequency	/u/	50%	11/22, sum of absolute value element-wise
Frequency	/u/	54.55%	12/22, Euclidean distance
Frequency	concatenate 4	68.18%	15/22, sum of absolute value element-wise
Frequency	concatenate 4	68.18%	15/22, Euclidean distance

The results obtained using both Euclidean distance and modified Euclidean distance (sum of absolute value element-wise) for both the responses to single vowels in the frequency domain and the concatenated responses to the four vowels are shown in Table 3.3. In the following figures, the left matrix exhibits the sum of absolute values from the subtraction of each frequency point, and the right matrix shows the result of Euclidean distance from the frequency points. From the distance matrix with single vowel responses and concatenated vowel responses, it can be clearly seen that the accuracy of the latter is always slightly higher than any of the former. The results obtained with the sum of absolute values is mostly similar

to those obtained with Euclidean distance. When making a comparison with the PCC, the results obtained with Euclidean distance appear to vary with respect to the vowel stimuli. For /a/ vowel stimuli, the result of Euclidean distance (14/22) exhibits the same performance with PCC with mean removed (14/22). The /ɔ/ and /u/ vowels show a similar trend when making comparison among those three methods, with the accuracy of Euclidean distance higher than that of the two other methods, namely 14/22 versus 9/22 and 10/22 for Euclidean distance, regular PCC and PCC with mean removed, respectively, for /ɔ/ vowel stimuli, and 11/22 versus 9/22 and 9/22 for /u/ vowel stimuli. In contrast, the accuracy of Euclidean distance of /U/ vowel (12/22) is slightly lower than regular PCC (13/22) and PCC with mean removed (13/22).

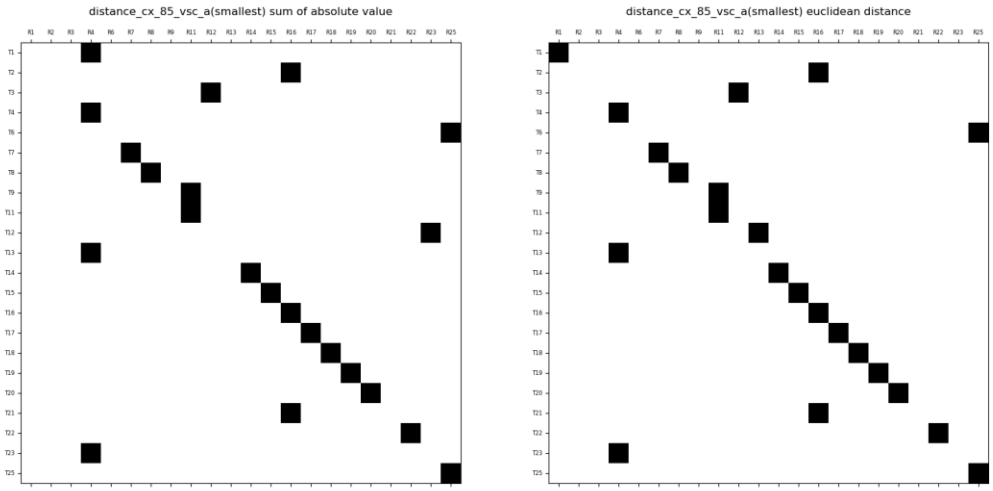

Figure 3-16: Euclidean and modified distance matrices for 85dB /a/ vowel signal in frequency domain

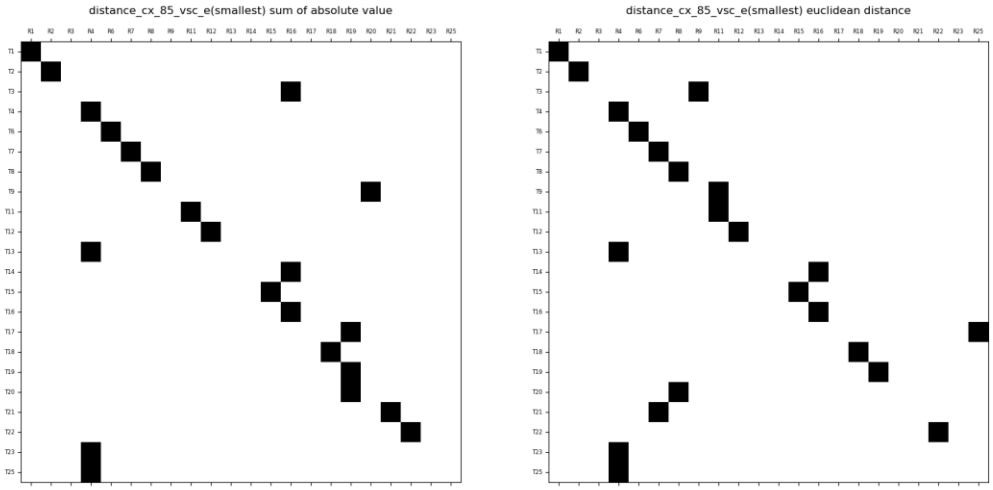

Figure 3-17: Euclidean and modified distance matrices for 85dB /ɔ/ vowel signal in frequency domain

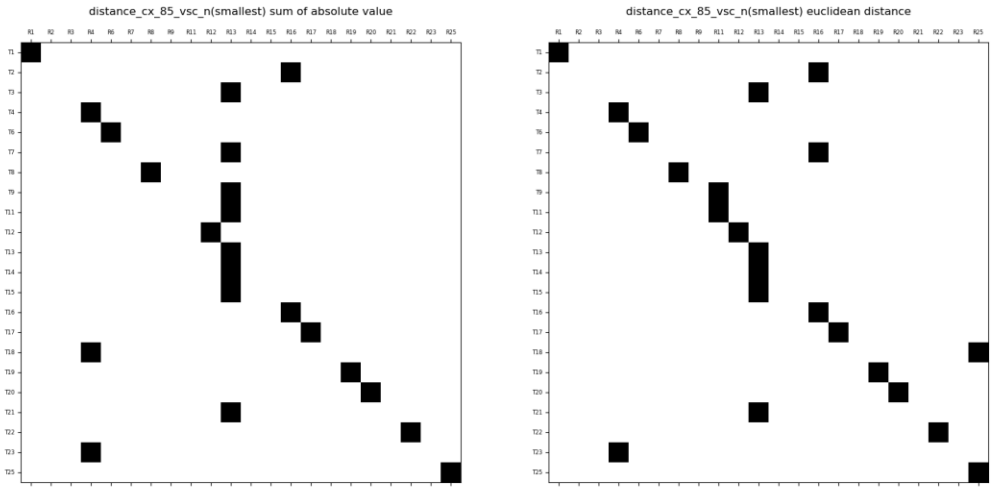

Figure 3-18: Euclidean and modified distance matrices for 85dB /U/ vowel signal in frequency domain

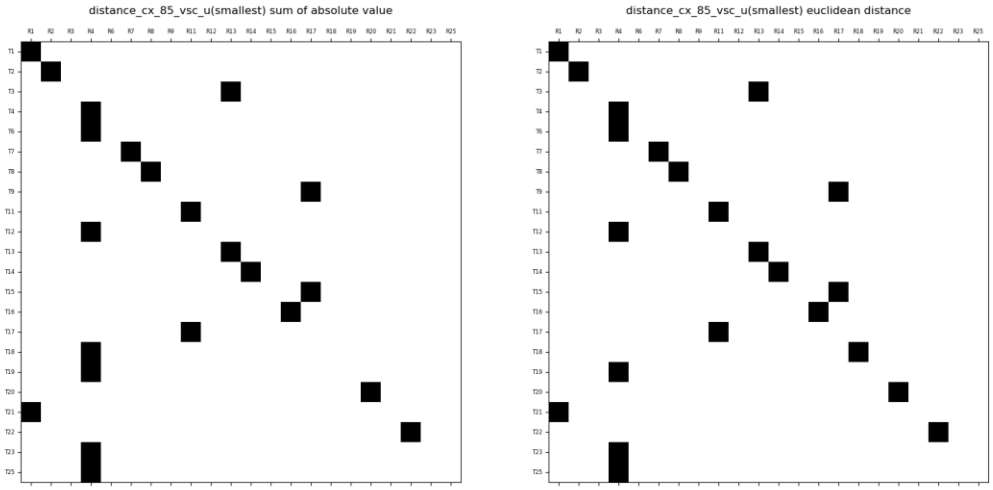

Figure 3-19 : Euclidean and modified distance matrices for 85dB /u/ vowel signal in frequency domain

In the distance matrices shown in Fig. 3-16 to Fig. 3-20, the black cells correspond to the minimum distance score between two frequency domain complex-valued series of the test and retest conditions, which contain both amplitude and phase information. When making comparisons between the overall accuracy obtained with Pearson correlation coefficient and Euclidean distance, it is found that the accuracy is similar overall. Meanwhile, the result of PCC method only relies on the amplitude information of the evoked response. In comparison, the Euclidean distance matrix uses both amplitude and phase-related information, which could theoretically provide more information for the classification.

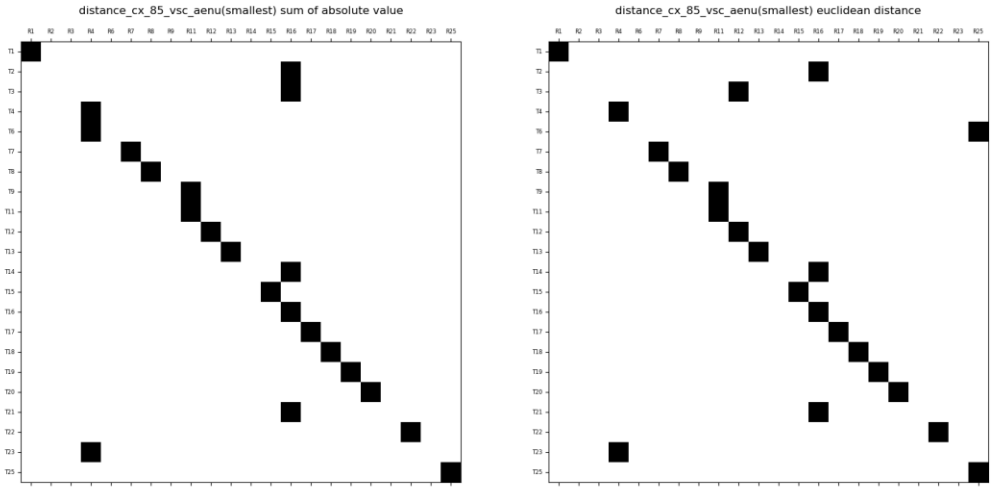

Figure 3-20 : Euclidean and modified distance matrices for 85dB concatenated four vowel signals in frequency domain

As shown in Fig. 3.17, the distance matrices obtained with the 85 dB concatenated four vowel responses demonstrate a slightly higher accuracy than the distance matrices obtained with single vowel responses. Although the distance matrix for concatenated vowel signals exhibits the same accuracy as with the previous PCC correlation matrix, the accuracy is still much lower than that with PCC with mean removed coefficients. In summary, using the Euclidean distance provides a similar performance to the regular PCC, but not as good as the PCC with mean removed obtained with mean component removed in the signals.

3.4 Verification of Signal Quality and Consistency

In order to assure the quality of the preprocessed signals before classification, the signal collected from the same day and same subject, namely the two sub-averages from the same

session, is compared in this section. Three aspects are validated here, namely, spectral flatness (Tonality coefficient), Pearson Correlation Coefficient (PCC) and the Peak Noise Ratio (PNR).

Here, these three quality metrics validate the quality in three different ways. Tonality coefficient is used for quantifying how tonal or noisy the signal is. The Pearson Correlation Coefficient is chosen for testing the consistency of the two signal sub-averages (each corresponding to 1500 trials) obtained on the same day and with the same subject. Regarding PNR, it is used for testing how "peaky" the magnitude frequency components are against the content between peaks (so it is related to the tonality coefficient).

These three quality metrics can allow rejection or inclusion of a signal for classification. The relationship between the accuracy of classifiers and the signal quality and consistency will be evaluated later in this chapter, and it will be also be used in the next chapter.

The tonality coefficient, namely spectral flatness, is an information-theoretic measure for the amount of randomness or stochasticity that exists in a signal, which has been used for characterizing the spectra of audio signals (Dubnov, 2004; Johnston, 1988). Typically, it provides a way to quantify how tonal-like or noise-like the signal is:

$$Flatness = \frac{\sqrt[N]{\prod_{n=0}^{N-1} x(n)}}{\frac{\sum_{n=0}^{N-1} x(n)}{N}} = \frac{\exp\left(\frac{1}{N}\sum_{n=0}^{N-1} ln\ x(n)\right)}{\frac{1}{N}\sum_{n=0}^{N-1} x(n)} \tag{3.4.1}$$

Here, spectral flatness is defined by dividing the geometric mean of the magnitude spectrum by the arithmetic mean of the magnitude spectrum. $x(n)$ represents the magnitude spectrum at the bin number n, and N is the number of bins considered. Note that the definition could also be applied to the square magnitude spectrum (related to power spectral density).

A high PCC score indicates that the subject has two similar response signal series from the same day and so the response signal has high consistency. The PCC is computed as:

$$\rho_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_x \sigma_y} \quad (3.4.2).$$

σ_x and σ_y in the above equation are the standard deviation of signal x and signal y from the same subject and same day.

For the PNR, it is designed based on the average auditory brainstem response in the frequency domain, with harmonics at 100, 200, 300 Hz until 800 Hz, which are the frequency bins used.

$$PNR = \frac{\sum_{i=0}^{N-1} P_i}{\sum_{j=0}^{N} I_j} \quad (3.4.3)$$

Here, P_i is the sum of amplitudes near the i^{th} harmonic (for example, P_0 represents the sum of 85-115 Hz amplitude spectrum, which corresponds to the spectrum around the 100 Hz peak). I_j is the sum of the amplitude spectrum of the remaining region between peaks. For example, I_0 represents the sum of spectral amplitudes from 0 to 85 Hz and I_1 represents the sum of spectral amplitudes from 115 to 185 Hz.

Three methods are adopted, each with a set threshold. The value of the threshold is determined for each method. The first method, namely logic 1, is proposed to accept the subjects whose 8 signal series (retest and test, /a/, /ɔ/, /U/, and /u/) are all above the threshold; otherwise, the subject for which any of these 8 scores is lower than the threshold will be rejected. The second method, namely logic 2, is defined by accepting the subject for which any of the 8 scores is higher than the threshold; otherwise, the subject for which all 8 scores are

lower than the threshold will be rejected. The third method, logic 3, is simplified by averaging the 8 scores and comparing this average score with a threshold. A subject whose average score is higher than the threshold is accepted.

The difference between the three methods is mainly related to the starting point of selecting the subjects. The first method (logic 1) is designed for selecting the subjects with most stability, as all of the 8 scores of the signal series should be above the threshold. In contrast, the second method (logic 2) only rejects the subjects who have low stability and keeps all the others. The behavior of the third method is expected to be somewhere in between the first two methods.

The threshold values are adjusted starting from the average value of all the retest and test subjects with 4 vowels and ending when the number of incorrectly predicted subjects becomes zero. Meanwhile, there is a slight difference between these three methods. For the Pearson Correlation Coefficient, the score is derived based on the two signal series from the same subject and same day. Therefore we get 8 PCC scores, corresponding to 2 conditions (test and retest) by 4 vowels, in total. For Peak Noise Ratio, as each signal series has its own PNR, we thus get 16 PNR scores, corresponding to 2 sub-averages with 2 conditions (test and test) by 4 vowels. Likewise, we get 16 tonality coefficients for each subject.

Figs. 3-21 to 3-32 below show the stacked bars of the three metrics (tonality, PCC, PNR) based on the three logics. The first four graphs from Fig. 3-21 to Fig. 3-24 are derived from the result of logic 1 with the three metrics. The number of subjects with correct match between test and retest sessions and over the threshold are labeled as orange, and the number of subjects with incorrect match between test and retest sessions and over the threshold are labeled as blue.

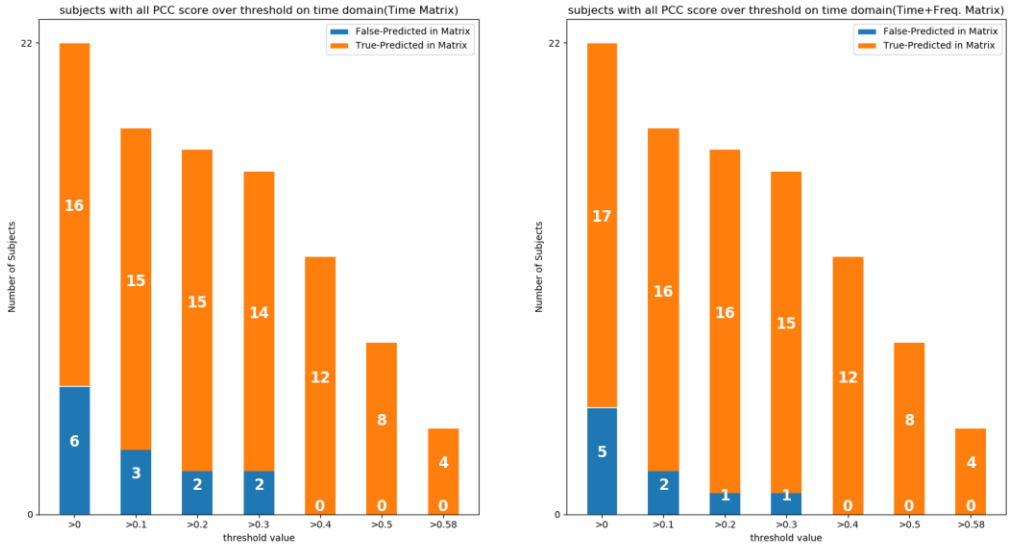

Figure 3-21 : Number of subjects with PCC quality score in time domain satisfying logic 1

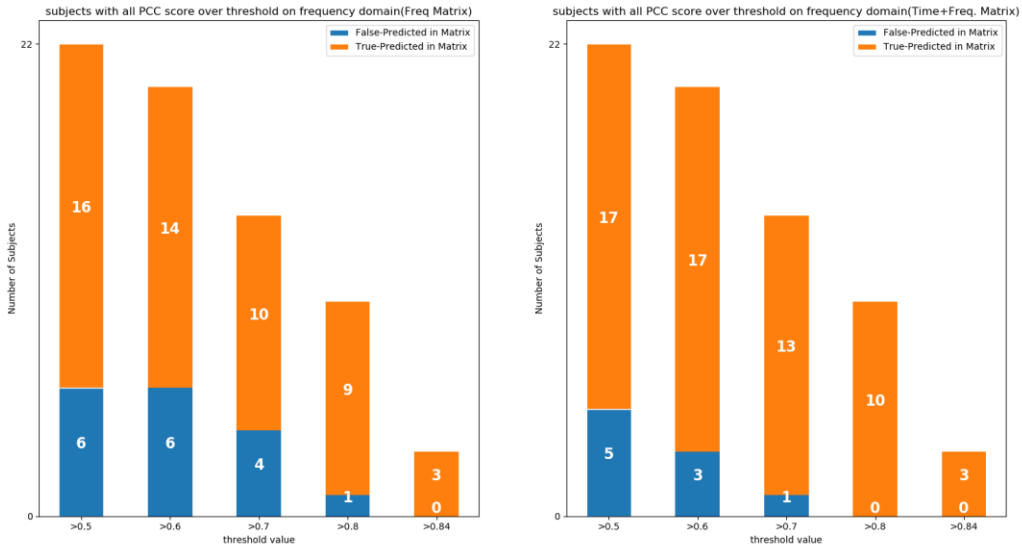

Figure 3-22 : Number of subjects with PCC quality score in frequency domain satisfying logic 1

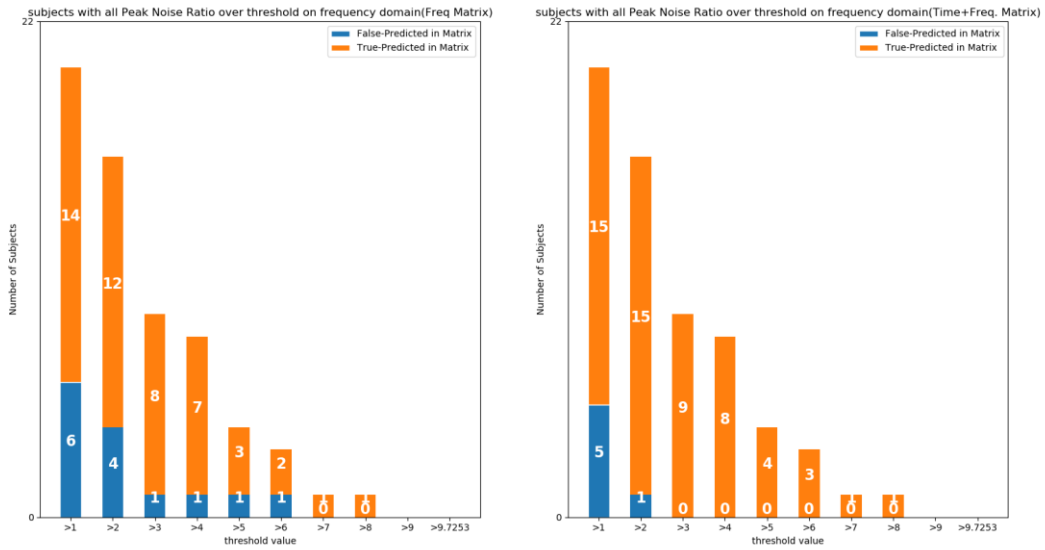

Figure 3-23 : Number of subjects with Peak Noise Ratio quality score satisfying logic 1

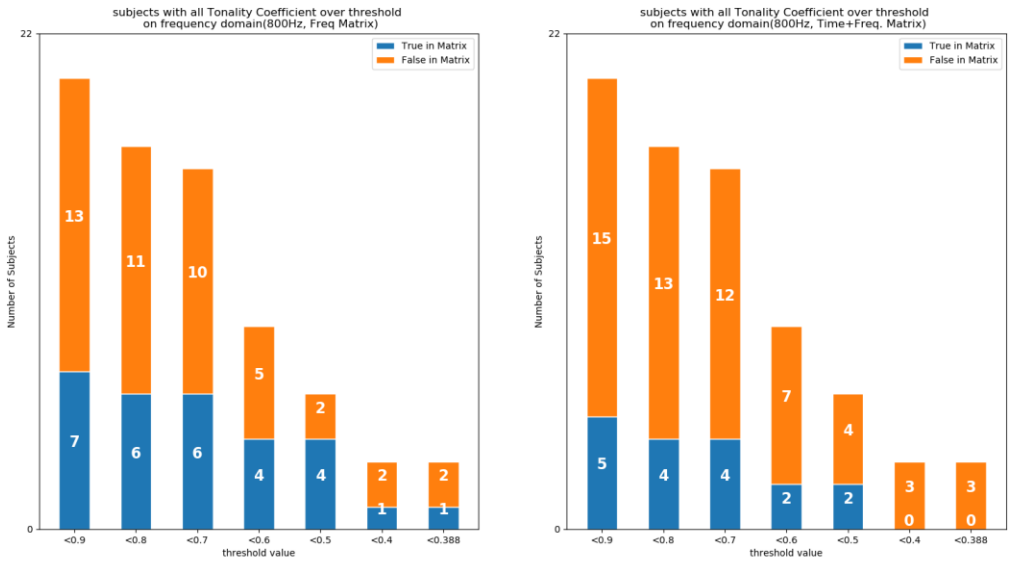

Figure 3-24 : Number of subjects with tonality coefficient quality score satisfying logic 1

With the first logic, it is obvious that the number of false-predicted subjects increases when the threshold is set higher. If the number of subjects is to be reduced until the quality of the remaining response signals is such that no classification error occurs, then the PCC quality score in time domain with threshold of 0.4 exhibits the best performance with 12 true-predicted subjects when using the time domain correlation matrix and the time and frequency domain correlation matrix for classification, which is better than the performance for the frequency domain correlation matrix. For PNR quality score, only 1 subject passed the testing with 100% accuracy based on the frequency correlation matrix, while with time and frequency matrix 9 subjects passed. Regarding the tonality coefficient quality score, there were always false predicted subjects selected with the increase of threshold, which means that the tonality coefficient cannot help in selecting the subjects with good quality response signals in this test.

With the logic 2, the change in the number of true-predicted vs. false-predicted subjects is shown in Figs. 3-25 to 3-28.

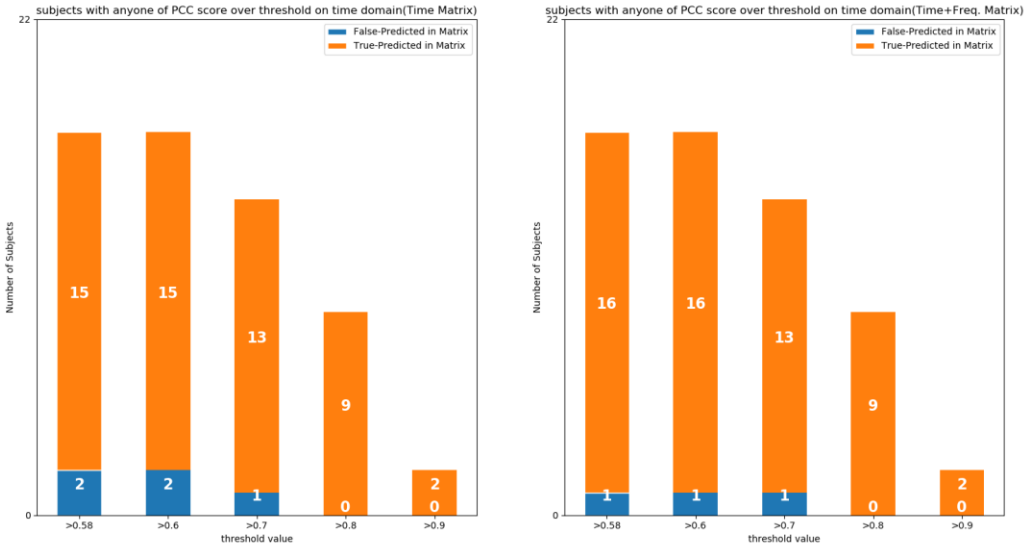

Figure 3-25: Number of subjects with PCC quality score in time domain satisfying logic 2

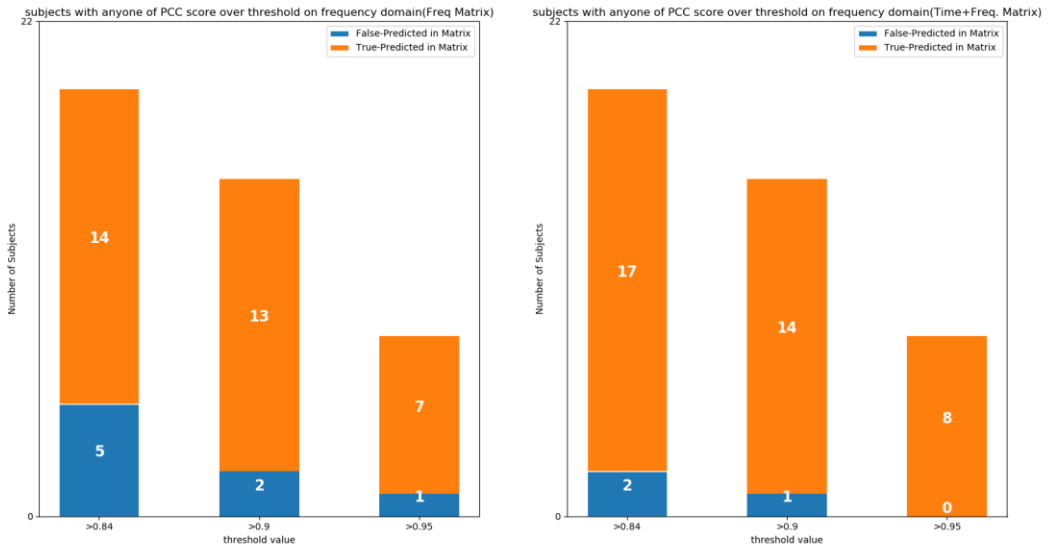

Figure 3-26 : Number of subjects with PCC quality score in frequency domain satisfying logic 2

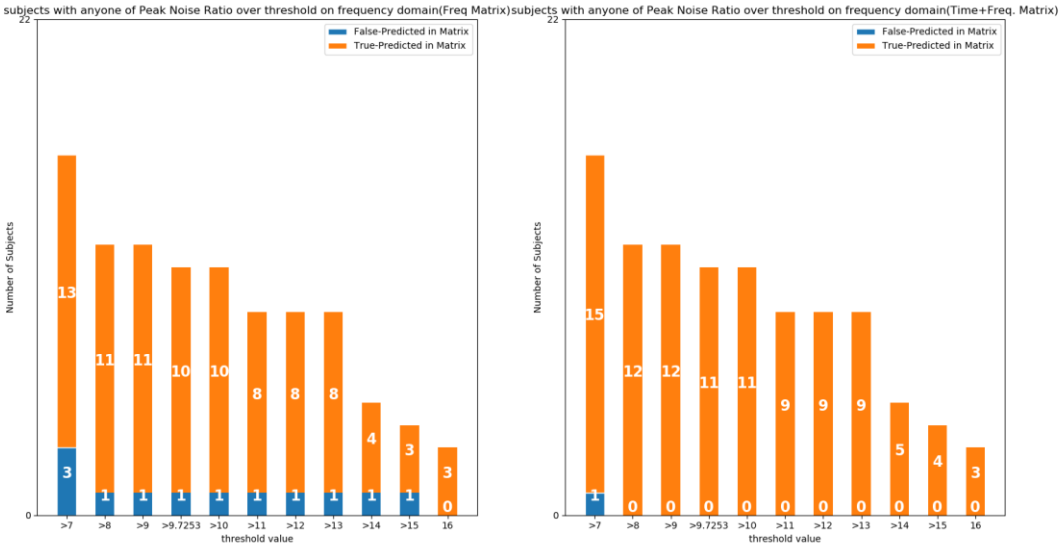

Figure 3-27 : Number of subjects with Peak Noise Ratio quality score satisfying logic 2

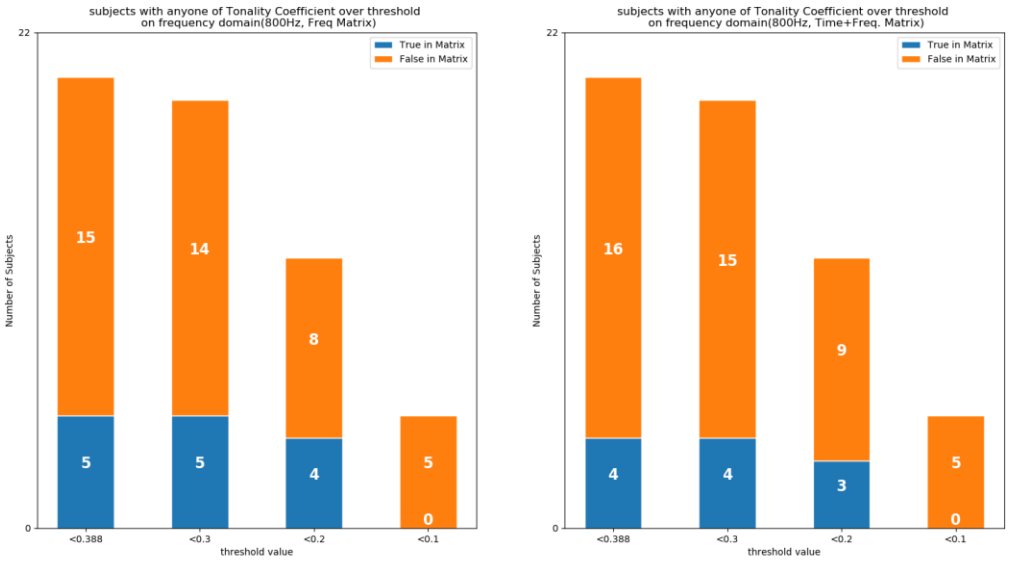

Figure 3-28 : Number of subjects with tonality coefficient quality score satisfying logic 2

With logic 2, the time domain PCC quality score provides a 100% accuracy with 9/9 subjects at threshold 0.8, which is a bit lower than the result from logic 1 (12/12). With the frequency domain PCC quality score, the results are weaker than for the time domain PCC quality score, and weaker than the previous case of logic 1. In contrast, the number of subjects with 100% accuracy with Peak Noise Ratio quality score gave a better result with logic 2, with 3/3 and 12/12 respectively, for the frequency domain correlation matrix and the time and frequency domain correlation matrix. Meanwhile, a 5/5 result is obtained with the tonality coefficient quality score based on either correlation matrix when the quality threshold was set as 0.1. Hence, so far, the result from time domain PCC quality metric based on logic 1 and from Peak Noise Ratio based quality metric on logic 2 exhibit the best result with 12/12 subjects.

For logic 3, the results are shown in Figs. 3-29 to 3-32.

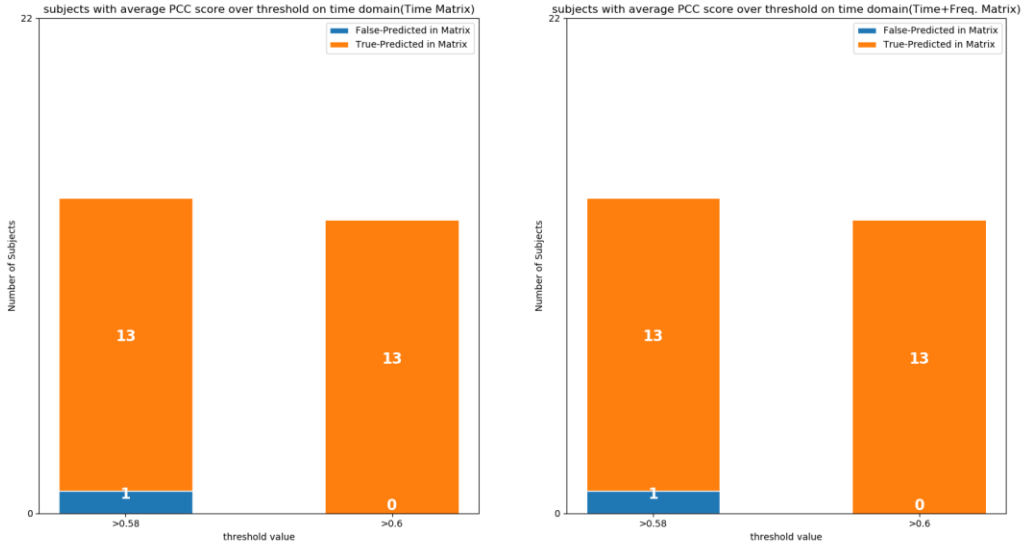

Figure 3-29 : **Number of subjects with average PCC quality score in time domain satisfying logic 3**

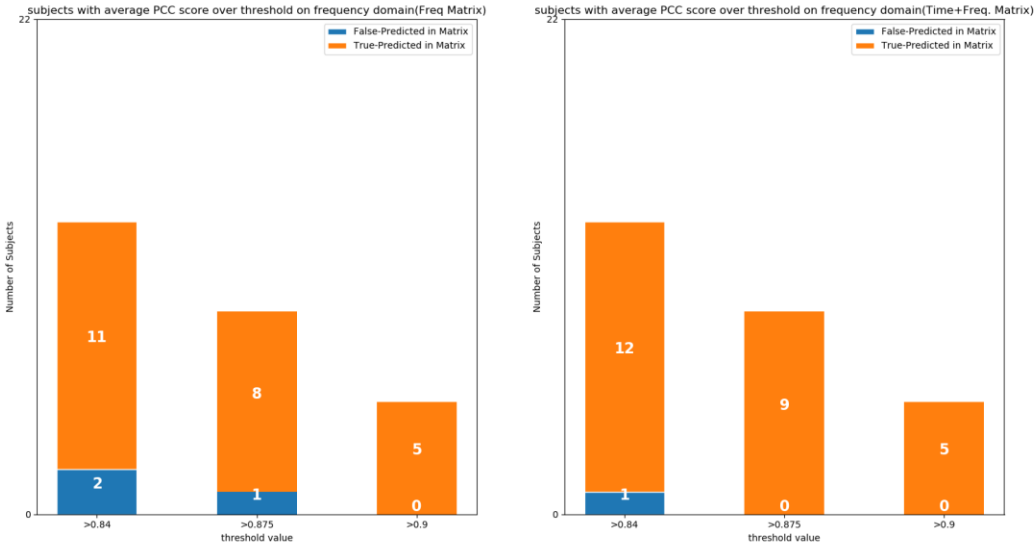

Figure 3-30 : Number of subjects with average PCC quality score in frequency domain satisfying logic 3

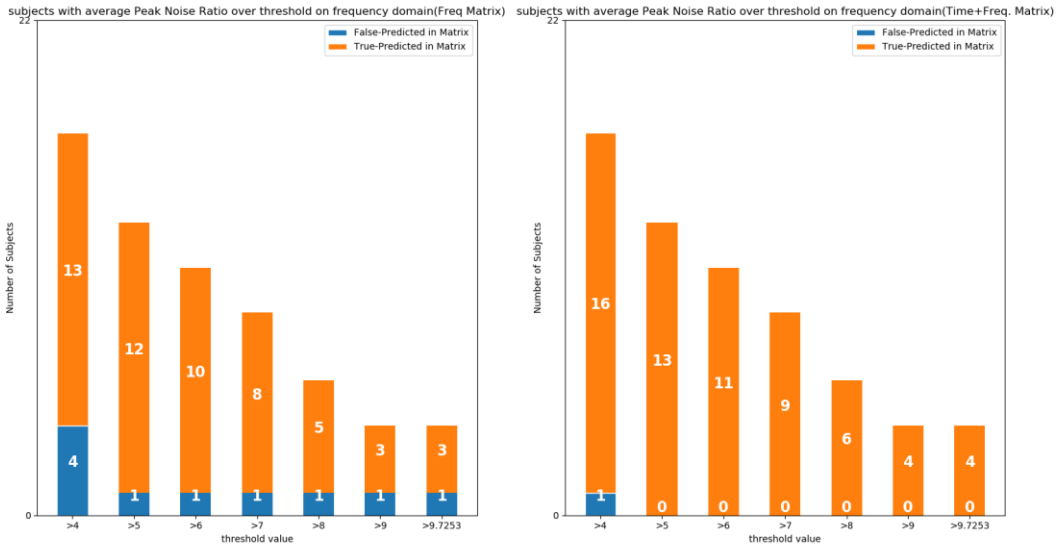

Figure 3-31: Number of subjects with average Peak Noise Ratio quality score satisfying logic 3

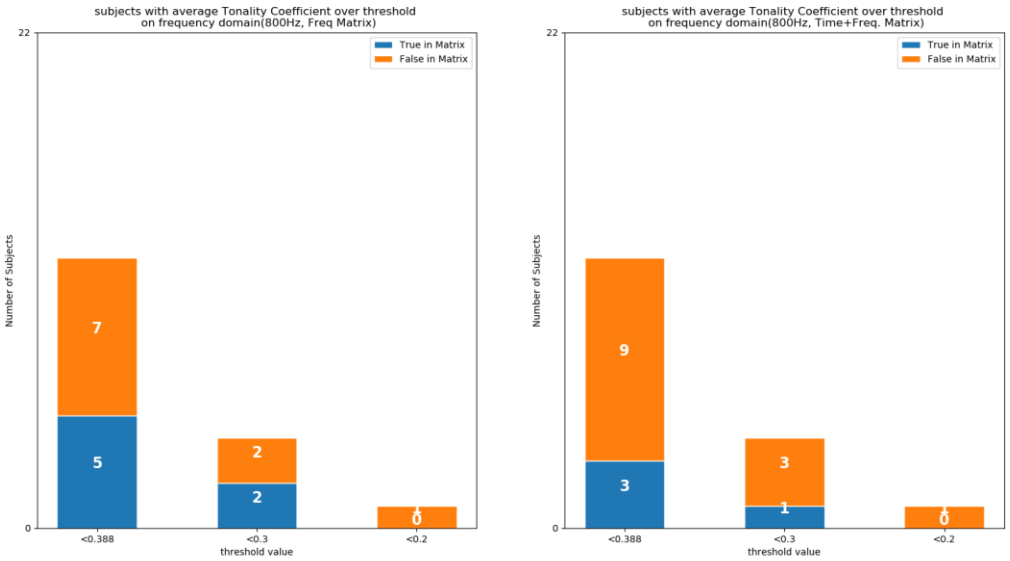

Figure 3-32 : Number of subjects with average tonality coefficient quality score satisfying logic 3

With logic 3, a result of 13/13 is found with PCC quality score in time domain and a threshold of 0.6, for either the time domain correlation matrix and the time and frequency domain correlation matrix, which is the best result that has been found. Results of 5/5 and 9/9 from frequency domain PCC quality score is obtained when the threshold is 0.9 and 0.975 respectively, for the frequency domain correlation matrix and the time and frequency domain correlation matrix.

At the same time, the Peak Noise Ratio quality score based on logic 3 gave a 13/13 performance with time and frequency correlation matrix, but no case without classification error when the frequency correlation matrix is used, which is not as stable as the results from time domain PCC quality score. Lastly, a 1/1 classification result is found when the threshold was set as 0.2 based on the tonality coefficient quality score, for either correlation matrix.

When making comparisons among the three methods, it is clear that the results from the PCC and PNR quality metrics are better than those from the tonality coefficient quality metric, which means that the tonality of the signal series is not a good indicator of signal stability and quality. On the contrary, the correlation (PCC) between the signals obtained with the same subject and on the same day, or the strength of certain peaks relative to the spectral noise (PNR) are better indicators of the quality and stability of the signals. Furthermore, with the three logics used, it was observed that using the average of the sub-scores (logic 3) gives a higher number of subjects identified with 100% accuracy number than the other two logics.

In conclusion, it can be seen from the results that the quality of the signals does affect the subject identification accuracy. The overall accuracy of the correlation matrix can be improved when the signal quality is improved.

3.5 Conclusion

The test and retest comparison and quality verification of signals in this chapter indicate that the evoked responses from four short vowels can be used to identify subjects through PCC or Euclidean distance. When the evoked responses are concatenated in both time and frequency domain and using the PCC with mean removed to generate the correlation matrix, the best test and retest comparison provided a subject identification accuracy of 81.82%. Moreover, the signal quality and consistency were investigated through the tonality coefficient, PCC and PNR quality metrics, with also three kinds of logic. The results showed that the test-retest comparison accuracy (subject identification) can be improved with increase of the signal quality.

4 Automatic Classification of FFR using Machine Learning and Deep Learning

4.1 Introduction

The study in this chapter focuses on the following research question: Can machine learning and deep learning methods discriminate between envelope FFR responses from different subjects with good accuracy? It also revisits the topic of envelope FFR signal quality and consistency between test and retest conditions, addressed in the previous chapter.

In order to test the proposed hypothesis, several well-known classifiers based on machine learning and deep learning are implemented and tested. Meanwhile, another motivation to classify the FFR with machine learning methods is to improve the experimental efficiency by reducing the number of trials required in order to obtain an acceptable subject identification performance.

4.2 Methods

Several machine learning and deep learning classifiers are used in this chapter including support vector machines (SVM), XGBoost and convolutional neural networks (CNN). These algorithms were implemented using the scikit-learn python library, the XGBoost package, and Keras from the Tensorflow library, respectively.

4.2.1 Classifier methods

Support Vector Machine

The SVM is a well-known binary classification method proposed by Cortes and Vapnik (Cortes & Vapnik, 1995), which was extended as a non-linear classifier through kernel functions to maximize the margin of hyperplanes (Boser et al., 1992). Consider that there is a

dataset consisting of d feature vectors as $(X_1, y_1), (X_2, y_2), \ldots, (X_i, y_i), \ldots (X_n, y_n)$, where $X_i \in R^d$, y_i is a scalar and $y_i \in +1, -1$. Assume that the data from the two classes are linearly separable, and therefore any hyperplane can be written as:

$$X^T W + b = 0 \tag{4.2.1}$$

where W is the normal vector for the hyperplane and the parameter b is a bias term. An hyperplane with a dataset is illustrated in Fig. 4.1, with two classes of training data and two hyperplanes.

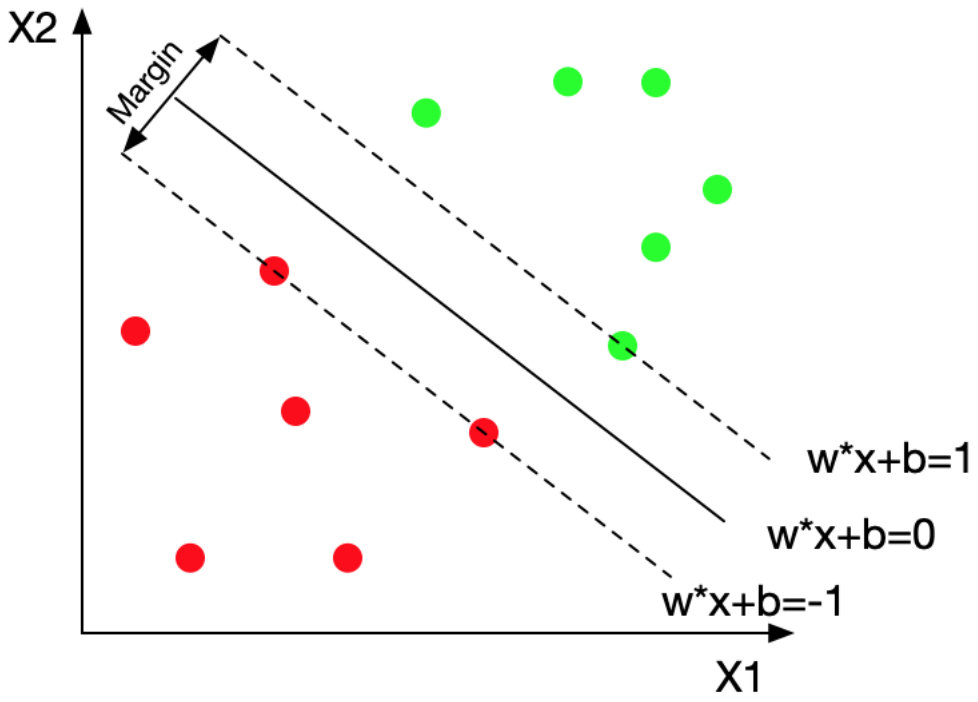

Figure 4-1: Linearly separable two-class example, with samples on the margin called the support vectors

As the training data is linear separable, we can plot two parallel hyperplanes that can separate the two labels, and the distance between the two hyperplanes should be as large as possible. Here, the term margin is defined as the room (space) that is bounded by the two

hyperplanes, and the maximum-margin hyperplane is the one that is equidistant between them. When normalizing the dataset, the two hyperplanes can be described as

$$X^T W + b = 1 \tag{4.2.2}$$

and

$$X^T W + b = -1 \tag{4.2.3}$$

Any data on and over the first boundary belongs to one class and is labeled as 1, while any data on or under the second boundary belongs to another class and is labeled as -1. The task is to find the "maximum-margin hyperplane" in order 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.

When 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 realized by a kernel function defined as:

$$K(x_i, x_j) = K(x_i^T x_j) = \phi(x_i)^T \phi(x_j) \tag{4.2.4}$$

where $\phi(x_i)$ is a mapping function. In this study, the linear kernel, polynomial kernel function, and Gaussian radial basis function are used and defined respectively follows.

For linear kernel:

$$K(x_i, x_j) = x_i^T x_j \tag{4.2.5}.$$

For polynomial function:

$$K(x_i, x_j) = (x_i^T x_j + b)^d \text{ , b} \geq 0 \tag{4.2.6}$$

where d is a specified degree of the polynomial (e.g. d=2 for quadratic) and b is a free parameter used to trade off the influence of higher-order versus lower-order terms in the polynomial.

For Gaussian radial basis function:

$$K(x_i, x_j) = \exp\left(-\gamma \left|\left|x_i - x_j\right|\right|^2\right), \gamma > 0 \tag{4.2.7}.$$

XGBoost

The XGBoost classifier is a decision-tree based ensemble machine learning algorithm proposed by Chen and Guestrin (Chen & Guestrin, 2016). It uses a gradient boosting framework and ensemble tree methods that apply the principle of boosting weak learners (Classification and Regression Tree, or CART). Consider that we have K trees, the model is

$$\tilde{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i) \tag{4.2.8}$$

where each f_k is the prediction from a decision tree, and x_i is the feature vector for the ith data point. To train the model, we need to optimize the loss function and we choose the root mean square error for regression here. Regularization is another significant part of the model. It controls the complexity of the model which prevents overfitting and is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{4.2.9}$$

where T is the number of leaves and w_j^2 is the score on the jth leaf. When putting the loss function and regularization together, we obtain the objective function of the model:

$$L(\phi) = \sum_i l(\tilde{y}_i, y_i) + \sum_k \Omega(f_k) \tag{4.2.10}.$$

Here the loss function controls the predictive power of the model and the regularization controls the simplicity of the model.

For the training objective, the gradient descent approach is used to optimize the objective function. The objective function can be redefined as:

$$L^{(t)} = \sum_{i=1}^{N} l\left(y_i, \tilde{y}_i^{(t)}\right) + \sum_{i=1}^{t} \Omega(f_i) = \sum_{i=1}^{N} l\left(y_i, \tilde{y}_i^{(t-1)} + f_t(x_i)\right) + \sum_{i=1}^{t} \Omega(f_i) \tag{4.2.11}.$$

We need to calculate the gradient to optimize the gradient descent. Therefore, the first and second order gradients can be considered. The second order Taylor series approximation of the objective function is calculated as:

$$L^{(t)} = \sum_{i=1}^{N} l\left(y_i, \tilde{y_i}^{(t-1)} + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right) + \sum_{i=1}^{t} \Omega(f_i) \tag{4.2.12}$$

where $g_i = \partial_{\tilde{y}^{t-1}} l(y_i, \tilde{y}^{(t-1)})$ and $h_i = \partial^2_{\tilde{y}^{t-1}} l(y_i, \tilde{y}^{(t-1)})$. Removing the constant terms gives:

$$L^{(t)} = \sum_{i=1}^{N} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) \tag{4.2.13}$$

This equation corresponds to the objective function at the t^{th} step. We define a tree as $f_t(x) = w_{q(x)}$, where $q(x)$ is a function which assigns every data point to the $q(x)^{\text{th}}$ leaf. The index set is defined as $I_j = \{i|q(x_i) = j\}$. This set contains the indices of data points that are assigned to the j^{th}. Then we can further rewrite the objective function as

$$\begin{aligned} L^{(t)} &= \sum_{i=1}^{N} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i \in I_j} h_i + \lambda\right) w_j^2\right] + \gamma T \end{aligned} \tag{4.2.14}$$

As all the data points from the same leaf share the same prediction, this form sums the prediction by the leaves. It becomes a quadratic problem of w_j and the best w_j to optimize the function is

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{4.2.15}.$$

The corresponding optimal value is:

$$L^{(t)} = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{4.2.16}.$$

Now, we define 'the best split' for the algorithm. Every time we make a split, we change a leaf into an internal node. Let I be the set of indices for data points assigned to the node.

Then I_L and I_R would be the sets of indices for data points assigned to two new leaves. When we recall the optimal value of the objective function on the j^{th} leaf, the gain of the split is:

$$\tilde{L}_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} + \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \qquad (4.2.17).$$

In order to build a tree, we find the best splitting point recursively until we reach the maximum depth, and prune out the nodes with a negative gain in a bottom-up order.

Convolutional Neural Network

CNNs are a class of biologically-inspired neural networks that use a mathematical operation called convolution in place of general matrix multiplications in one or more of the layers. It has been widely used in image recognition (Parkhi et al., 2015; Wei et al., 2016), audio classification (Hershey et al., 2017), natural language processing (Hershey et al., 2017), etc. However, little research has been performed on the use of CNNs for EEG signal classification. In this study, we deploy the CNN for identification of subjects based on their envelope FFRs.

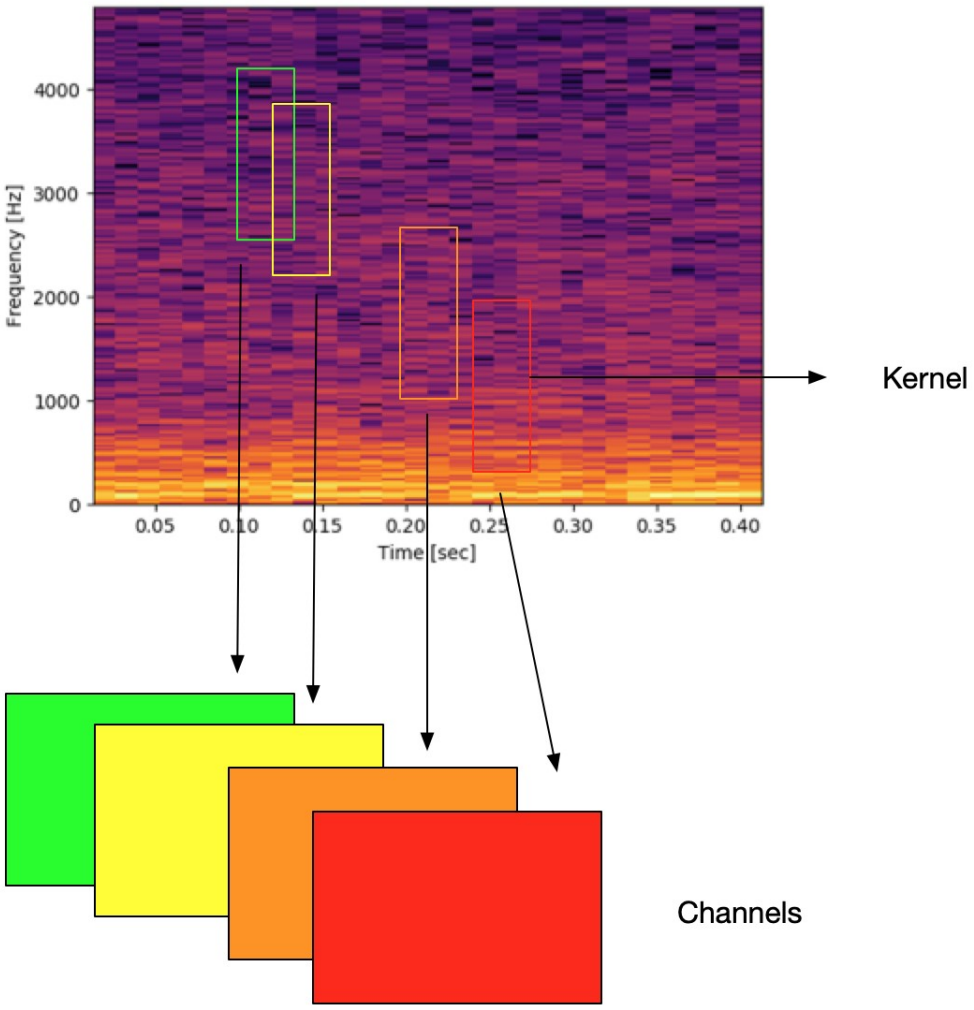

Figure 4-2 : Two-dimensional convolutional neural network with spectrogram as input

The CNN architecture used in our study consists of two different types of layers: convolutional layers and fully connected layers. A convolutional layer performs a convolution on input two-dimensional signals (coming from a previous convolutional layer or an input layer), with filters whose weights are learned by the CNN. The outputs of all the convolutions are known as a 2-D feature maps, and there can be multiple "channels" of feature maps (Fig 4-2). Each output feature map is produced by a single 3-D filter and its weights (the 3-D dimensions of the filter come from a 2-D filter dimension and the number of input channels).

Each output sample in each output feature map is the result of a convolution involving the input samples of all the input 2-D channels and located in a region around the location of the output sample in the 2-D feature map (e.g. rectangles in Fig. 4-2). The size of the 2-D region considered for each convolution is defining the 2-D filter dimension (e.g. 2 x 2, or 3 x 3). The convolution equation can be written as

$$y_{k',i',j'} = \sum_{k=0}^{K-1} \sum_{i=-I/2}^{\frac{I}{2}-1} \sum_{j=-J/2}^{\frac{J}{2}-1} x_{k,i'+i,j'+j} h_{k',k,i,j} + b_{k'} \tag{4.2.18}$$

where for each layer, x represents the input samples, k is the input channel index, i and j are the 2-D coordinates in an input feature map, y represents the output samples, k' is an output channel index, i' and j' are 2-D coordinates in an output feature map, K is the number of input channels, I,J (assumed even-valued in the equation above) are the 2-D dimensions of each filter, h are the filter weights, and b is a bias term (i.e., another filter weight).

Note that all the samples in a given output feature map are produced by the same filter weights, and equivalently common weights are applied to any 2-D region in an input feature map. There is therefore a lot of weights sharing. At the opposite, for a fully connected layer, every single input sample in a layer is connected with a unique weight to each output of that layer. Fig. 4.3 shows a combination of convolutional layers and fully connected layers.

The loss function is a method of evaluating how well a specific algorithm models the given data. During the training process, we aim to minimize the error for each training example and the error comes from the loss function. The categorical cross entropy shown in equation 4.2.19 is used in this study as a loss function:

$$C = -\frac{1}{n}\left(\sum_{j=1}^{K}\left[y_j ln(p_j) + (1-y_j)\ln(1-p_j)\right]\right) \tag{4.2.19}$$

where n is the size of the training set, y is a binary indicator (0 or 1), p is the predicted probability and K is the number of classes. It is a loss function suitable for single-label multi-

class classification tasks, where for each input feature there is always a single output label with a positive classification.

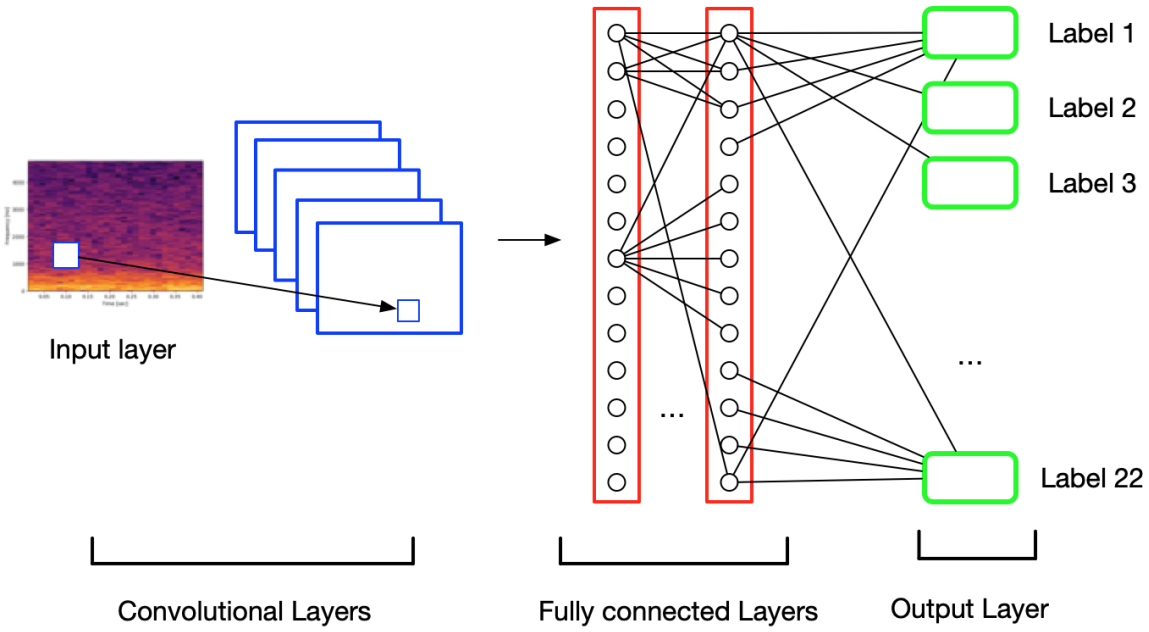

Figure 4-3 : A convolutional neural network structure example used in this study

Meanwhile, two types of activation functions are used in the study, the rectified linear activation unit (ReLU) and the softmax. The activation function is part of each neuron and it maps a set of inputs to an output in order to impart non-linearity to the network. The rectified linear unit is defined as the positive part of its argument:

$$f(\mathrm{x}) = x^{+} = \max(0, x) \tag{4.2.20}$$

where x is the input to a neuron.

The softmax function computes the probability distribution of the output classes. Hence, the output layer uses the softmax function outputs to predict which class the input signal belongs to (i.e., by selecting the class with the largest softmax value). The softmax function is defined by the following equation:

$$P_j = \frac{e^{x_j}}{\sum_1^k e^{x_k}} \text{ for } j = 1, \dots k \tag{4.2.21}$$

where x represents the inputs to the softmax function, and k is the number of classes (i.e., number of neurons in the last layer of the neural network). Each P_j is in the interval (0,1) and the sum of the P_j components in the output vector is 1.

Due to the small data set used in this study, overfitting easily happens during the training process. This happens when the model fits too well to the training set and then becomes difficult to generalize to new examples that are not in the training set. Two strategies are used to reduce overfitting, namely L2 regularization and dropout.

Regularization is the process of adding information so as to prevent overfitting in the study. The regression model using L2 regularization is called Ridge regression. It adds a squared magnitude of coefficients as a penalty term to the loss function, which is shown below:

$$C = -\frac{1}{n}\left(\sum_{x_j}\left[y_j ln(p_j) + (1 - y_j)\ln(1 - p_j)\right]\right) + \boldsymbol{\frac{\lambda}{2n}\sum_w w^2} \tag{4.2.22}$$

where the bolded part represents the L2 regularization element, n is the size of the training set, y is a binary indicator (0 or 1), and p is the predicted probability. The L2 regularization is scaled by the factor $\lambda/2n$, where $\lambda > 0$ is known as the regularization parameter.

Alternatively, dropout can be an easy and effective way to prevent overfitting (Srivastava et al., 2014). The key idea is to randomly select neurons to be dropped-out with a given probability during the training process. It prevents units from co-adapting too much.

Meanwhile, during the training process, an optimization algorithm needs to be used for updating the model parameters such as weights and bias values. The stochastic gradient descent and an adaptive subgradient method named Adagrad were used for this task.

All the neural network structures used in the study are described in Table 4.1, all with convolutional layers using rectified linear unit (ReLU) activation functions, with fully connected hidden layer also using ReLU activation functions, and with the output layer using softmax as the activation function.

Table 4.1: Detail of CNN structures used in this study

Model Name	Feature	Model Structure	Optimizer	Average Accuracy
CNN model 1	Spectrogram in the shape of [800, 31]	20*Conv(100*10) 50*Conv(50*5) Flatten Dense(22)	Adagrad	75%
CNN model 2	Spectrogram in the shape of [800, 31]	20*Conv(100*10, L2) 50*Conv(50*5, L2) Flatten Dense(22)	Adagrad	77.275%
CNN model 3	Spectrogram in the shape of [800, 31]	20*Conv(100*10, L2) 20*Conv(50*5, L2) Flatten Dense(22)	Adagrad	78.41%
CNN model 4	Spectrogram in the shape of [800,4]	64*Conv(100*4, L2) 64*Conv(10*1, L2) Flatten Dense(22)	Adagrad	65.91%
CNN model 5	Spectrogram in the shape of [800,4]	32*Conv(100*4, L2) 32*Conv(10*1, L2) Flatten Dense(22)	Adagrad	69.315%
CNN model 6	Spectrogram in the shape of [800,4]	32*Conv(100*1, L2) 32*Conv(10*4, L2) Flatten Dense(22)	Adagrad	65.91%
CNN model 7	Mel Spectrogram in the shape of [22,9]	32*Conv(2*2, L2) 32*Conv(2*2, L2) Flatten Dense(50, L2) Dense(22)	SGD	56.82%
CNN model 8	Mel Spectrogram in the shape of [22,9]	32*Conv(2*2, L2) 64*Conv(2*2, L2) Flatten Dense(50, L2) Dense(22)	SGD	61.36%
CNN model 9	Mel Spectrogram in the shape of [22,9]	32*Conv(2*2, L2) 64*Conv(2*2, L2) Flatten Dense(50, L2) Dropout(0.2) Dense(22)	SGD	64.775%

4.2.2 Model training, testing, and evaluation methods

The classification task performed in the study can be broken into:

1. The application of XGBoost on spectrogram and Mel spectrogram of the FFRs
2. The application of SVM with four different kernel functions to the spectrogram and Mel spectrogram of the FFRs
3. The application of the CNN to the spectrogram of the FFRs.

The dataset used for classification contains 88 sets of data with 22 labels corresponding to the 22 subjects, with each set corresponding to the coherent average of 1500 responses. Each dataset includes a concatenation of responses from all four vowels with the same subject. 44 sets of the data are from the test day and the remaining 44 sets are from the retest day. Due to the limitation of the dataset, it will be split into training and testing sets in the ratio of 50:50, with test and retest signals representing training and testing datasets, respectively. The time domain envelope FFRs from four vowels were preprocessed to extract both a Mel spectrogram and a spectrogram. All instances of the XGBoost classifiers in this section are default implementations (implemented using XGBoost library), with a maximum tree depth for base learners as 3, a number of trees to fit as 100, and a learning rate equal to 1.

The dataset was shared for both XGBoost and SVM classifiers. Three kernel functions (implemented using scikit-learn library) were used for training the SVM and making comparisons, namely the linear kernel function, the polynomial kernel function, and the radial basis function. Meanwhile, the kernels use a one versus the rest approach to process the multiclass classification.

Finally, in the third section, CNNs were trained with several kinds of features from the dataset (as indicated in Table 4.1). Due to the different sizes of the input feature sets from the Mel spectrogram and the spectrogram, the structure of the network was tuned so as to fit the

different input feature sizes, with regard to the number of hidden layers, optimizer, number of neurons in each layer, and the strategy to avoid overfitting.

The classification accuracy is computed with the aggregate of correctly classified testing subjects over the total number of testing subjects. Therefore, the chance accuracy of the classification task is 1/22 or 4.54%. In addition to classification accuracy scores, a confusion matrix was computed with the actual label versus predicted label, in order to give a more intuitive understanding of the performance of the classifier and of the classification errors made by each classifier.

4.3 Results

4.3.1 Using FFR time series data as features

The subject-level classification performance of the classifiers trained on the time domain eFFR with test or retest session as training set is displayed in Fig. 4.4. The SVM with the radial basis function as kernel produced a better performance, with an average classification accuracy of 82.95% (an average of 84.09% and 81.82% from the two different training sets).

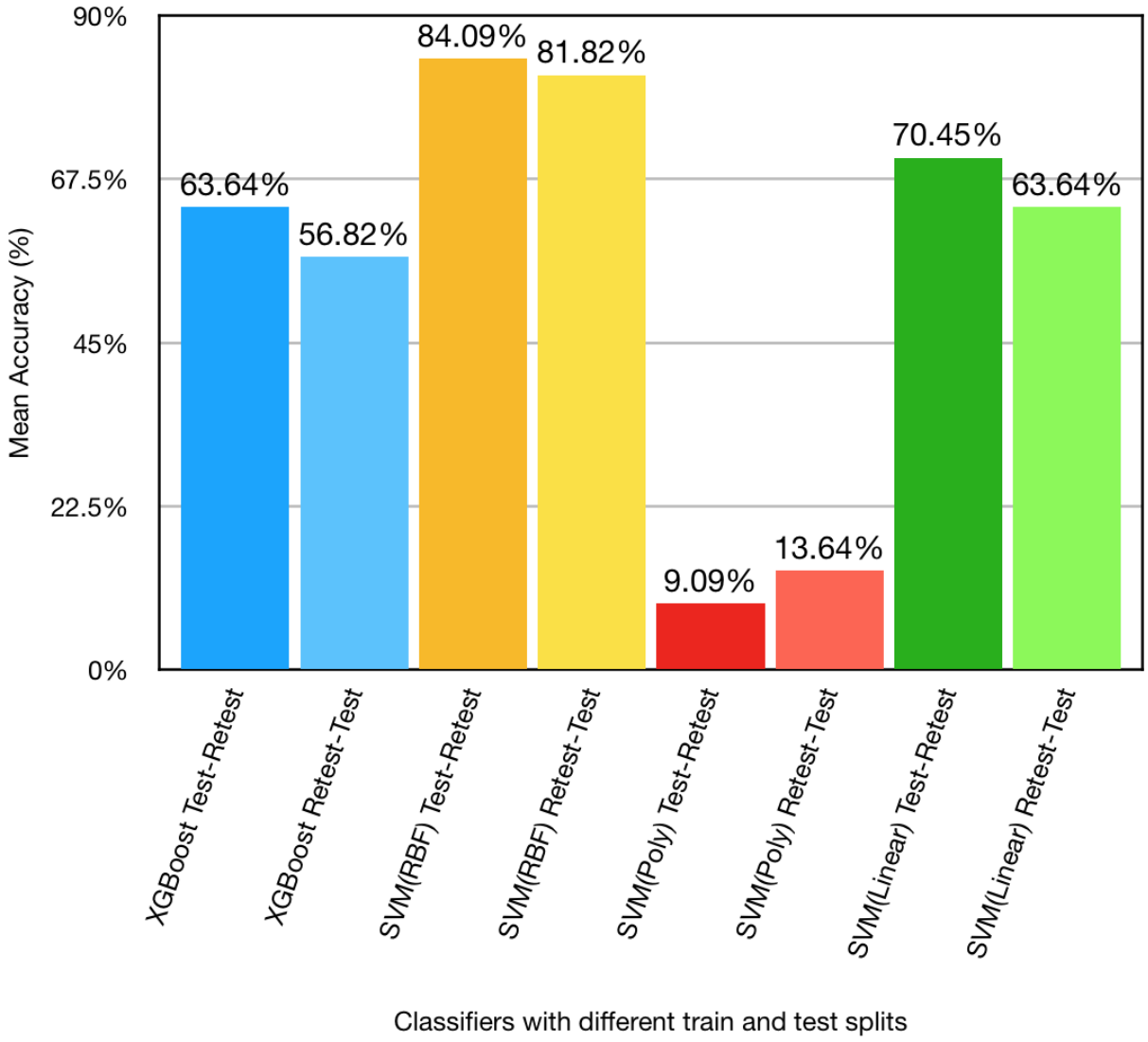

Figure 4-4 : Subject-level class predictive performance of each classifier trained and tested on either test or retest session, shown in terms of mean accuracy.

A heat map for the confusion matrix obtained using SVM with RBF kernel is shown in Fig. 4.5. A large majority of predicted matches appear on the diagonal. A deeper inspection indicates that a good proportion (94.7%, 18 out of 19) of matches on the diagonal correspond to cases where there is a good match between the test and retest data (good quality indicator score), whereas three of the subjects - namely subject #3, subject #8, and subject #19 - result in no match between test session and retest session (poor quality indicator score).

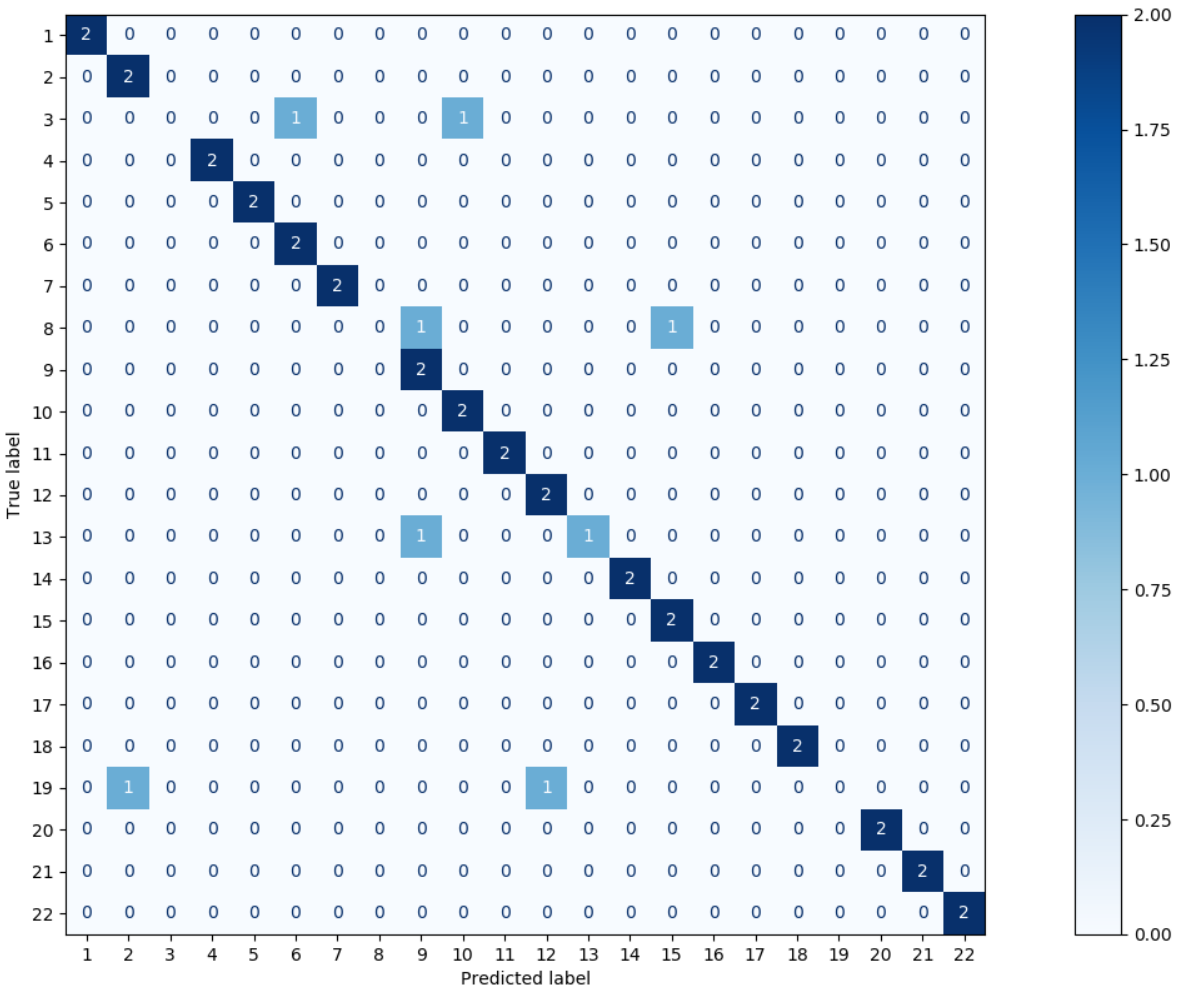

Figure 4-5 : Confusion matrix for SVM with radial basis function kernel on the subject-level classification task using time domain evoked response as features, with test session as training set and retest session as testing set.

The results across all classifiers for the subject-level task are summarized in Fig. 4.4. The SVM with RBF kernel outperforms all other models regardless of the features used for training.

4.3.2 Using the spectrograms as features

The relative performance of each classifier trained and tested on either test session or retest session data with the spectrogram features is displayed in Fig. 4.6. Here the spectrogram features were derived through Discrete Fourier transforms over consecutive 256 point segments, with 50% overlap, the FFT length as 9606 and a Tukey window with a shape parameter of 0.25, leading to a spectrogram with a shape of [4803, 31] for each data file with length 106.6 ms. The frequency frame from 800Hz to 4803Hz was discarded and keep the new shape as [800, 31] for classification. The 800 size is for frequencies ranging from 0 to 800 Hz and the 31 size is for the time frames.

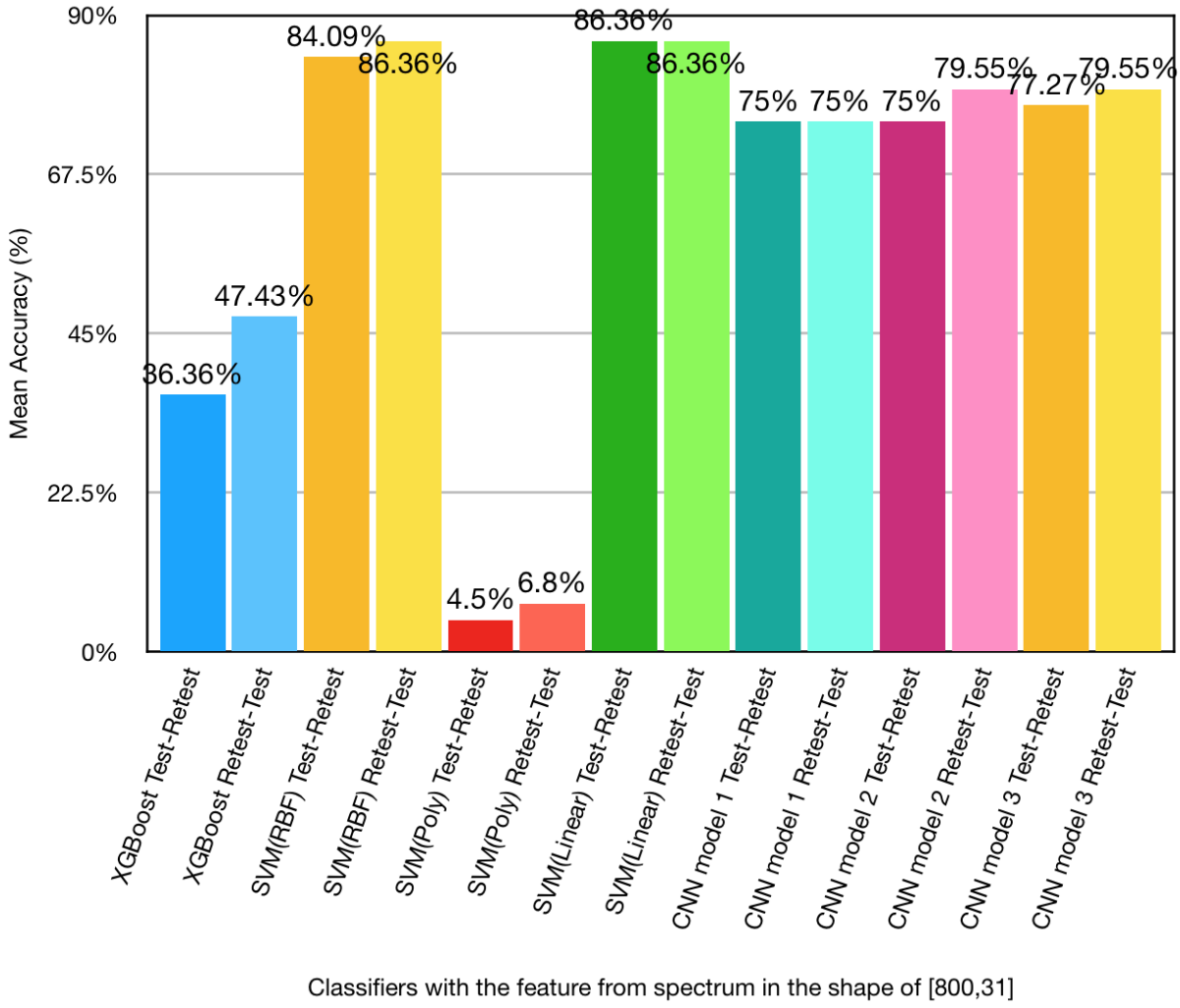

Figure 4-6 : Predictive performance of each classifier trained and tested on either test or retest session with the spectrogram feature set, shown in terms of the accuracy score.

It can be seen from the figure that all three CNN models performed with a similar performance. The SVM with linear kernel function achieved the best accuracy by correctly classifying 86.36% of the subject labels for the two sets of train-test-split sessions. The SVM with RBF kernel follows with the second highest average accuracy of 85.23% through averaging two sets of train-test-split session results. The SVM with polynomial kernel function achieved a low average accuracy of 5.65%, which is only slightly higher than the chance accuracy of 4.54%. This is in line with the result in Fig. 4.4, wherein the SVM with polynomial function achieved an average accuracy of 11.37%.

The confusion matrices for each of the SVM classifiers with linear kernel function for test session and retest session as training set are presented in Fig. 4.7. Although both classification results yield identical classification accuracies, there are some differences in the confusion matrices. For example, the classifier using test session as training set confuses subjects 1 and 3 when predicting using retest session, whereas the SVM using retest session as training set confuses subjects 2 and 9. Meanwhile, the prediction for subject 18 reveals a no-match but stable performance between the two cases, which also indicates the stability of the classifier. Overall, the SVM with linear kernel function classifier generally exhibits better performance than the other classifiers.

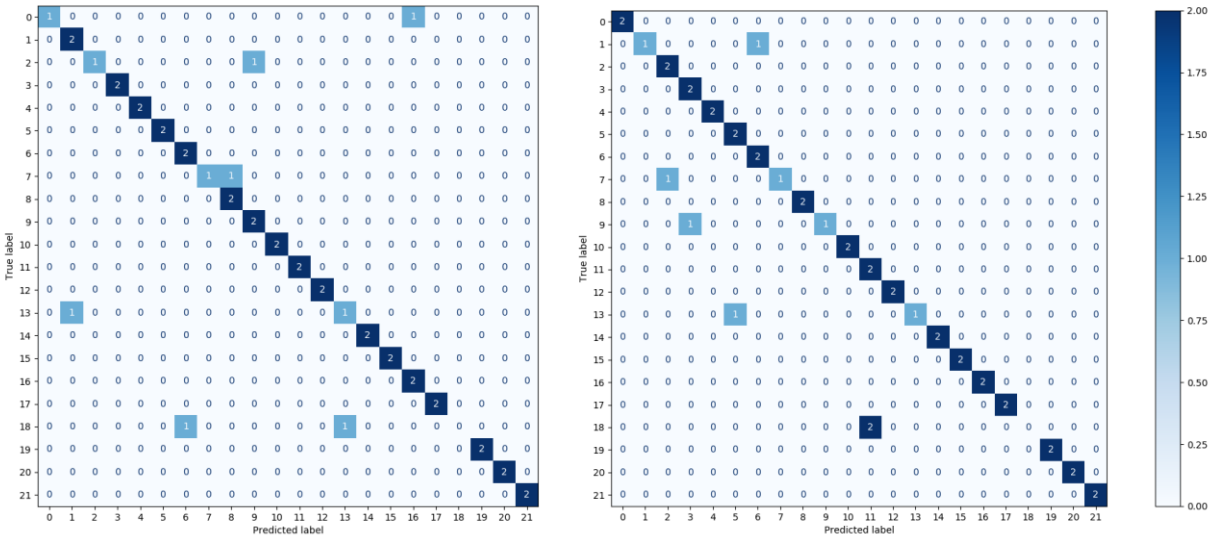

Figure 4-7 : Confusion matrices for the subject classification performance of the SVM classifier with linear function using test (left) and retest (right) session as training set with the spectrogram feature set.

Another set of classification performance results for classifiers trained on spectrograms with shape [800,4] is shown in Fig. 4.8. These spectrograms are calculated using 1024 point segments with no overlap between them, which ensures that the 4 columns correspond to the Discrete Fourier transforms of all four vowels. Again, only a range between 0 to 800 was

selected for the amplitude spectra, in order to avoid the effect of frequencies over 800 Hz which have less impact of classification.

The results of all classifiers are summarized in Fig. 4.8. The performance of SVM with linear kernel function outperforms all other models once again. However, the average accuracy (80.68%) is slightly lower than the result obtained using spectrogram with the shape of [800,31] which suggest that the frequency content over time provides slightly more useful information than the static Discrete Fourier transform.

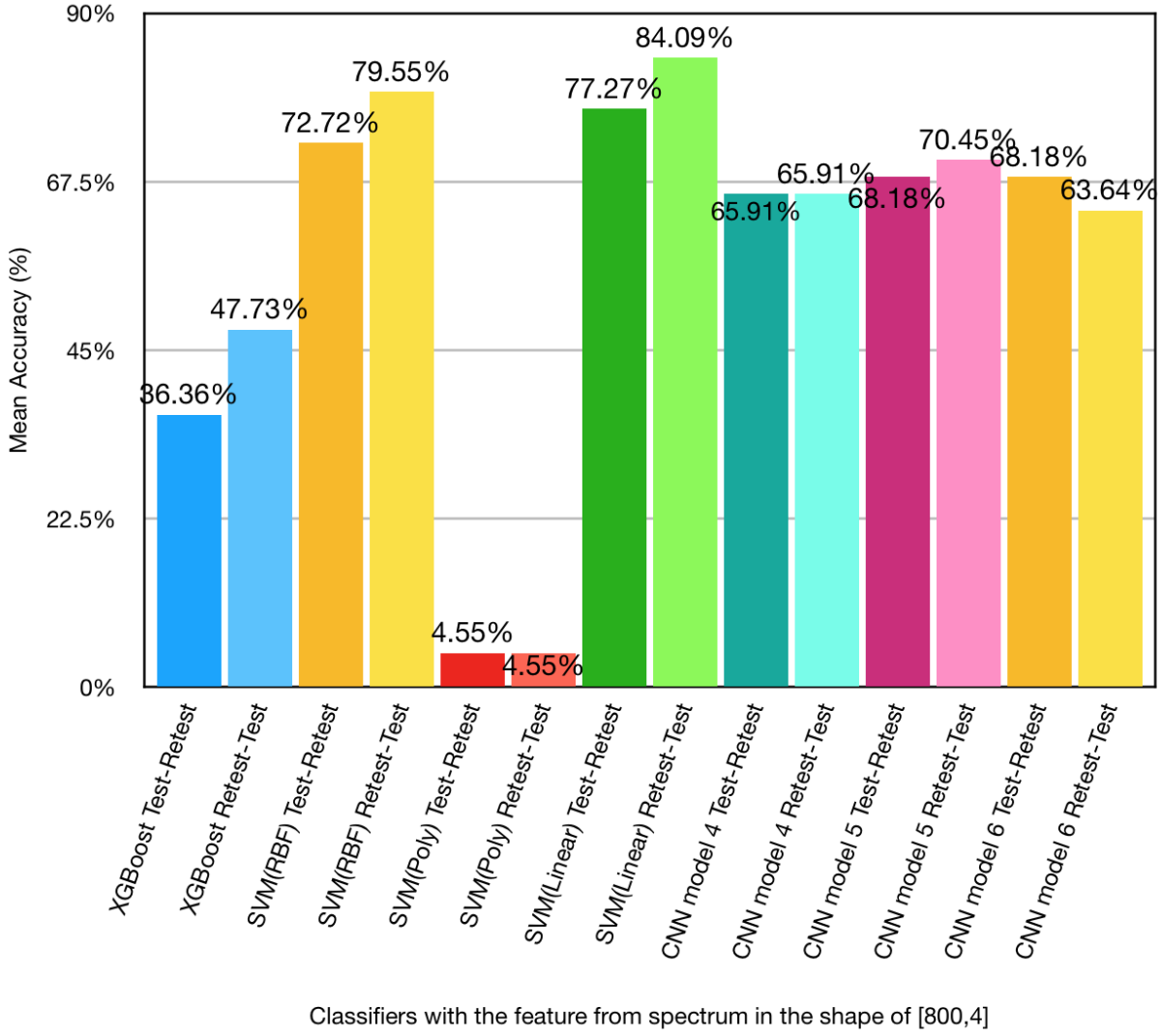

Figure 4-8 : Predictive performance of each classifier trained with either test or retest session with spectrogram feature of size [800,4]

4.3.3 Using Mel spectrograms as features

The results of all classifiers trained on either test or retest session with Mel spectrogram features are summarized in Fig. 4.9.

The input signal series is converted into a Mel spectrogram. The mel-frequency scale is related to biological hearing and has been proven to be successful in the domain of speech processing and recognition (Narayanan & Wang, 2013; Shen et al., 2018). The magnitude spectrogram S is first computed using the time-series signal as input. The S raised to the power of 2 is then mapped onto the Mel-scale with the function:

$$m = 2595 * log_{10}\left(1 + \frac{f}{700}\right) \tag{4.2.23}$$

Lastly, the features corresponding to 0 to 800 Hz were kept so as to be in line with the spectrogram previously used. Here, the Librosa python library provides inbuilt functions for Mel spectrogram calculation.

A maximal subject prediction accuracy of 72.7% was obtained using a CNN model. The SVM with RBF and linear kernels and two other CNN models (CNN model 7 and 8) achieve the best performance here, between 54.5% and 65.9%. Although the average performance of the XGBoost and SVM with polynomial kernel function in the subject classification task ranks at the bottom of the list, it is worth noting that both of the two performances climbed up from 36.36% (47.73%) and 4.55% (4.55%) to 59.09% (47.73%) and 31.8% (18.1%) for test-retest train-test split (and retest-test respectively), compared with the result to time series feature and standard spectrogram.

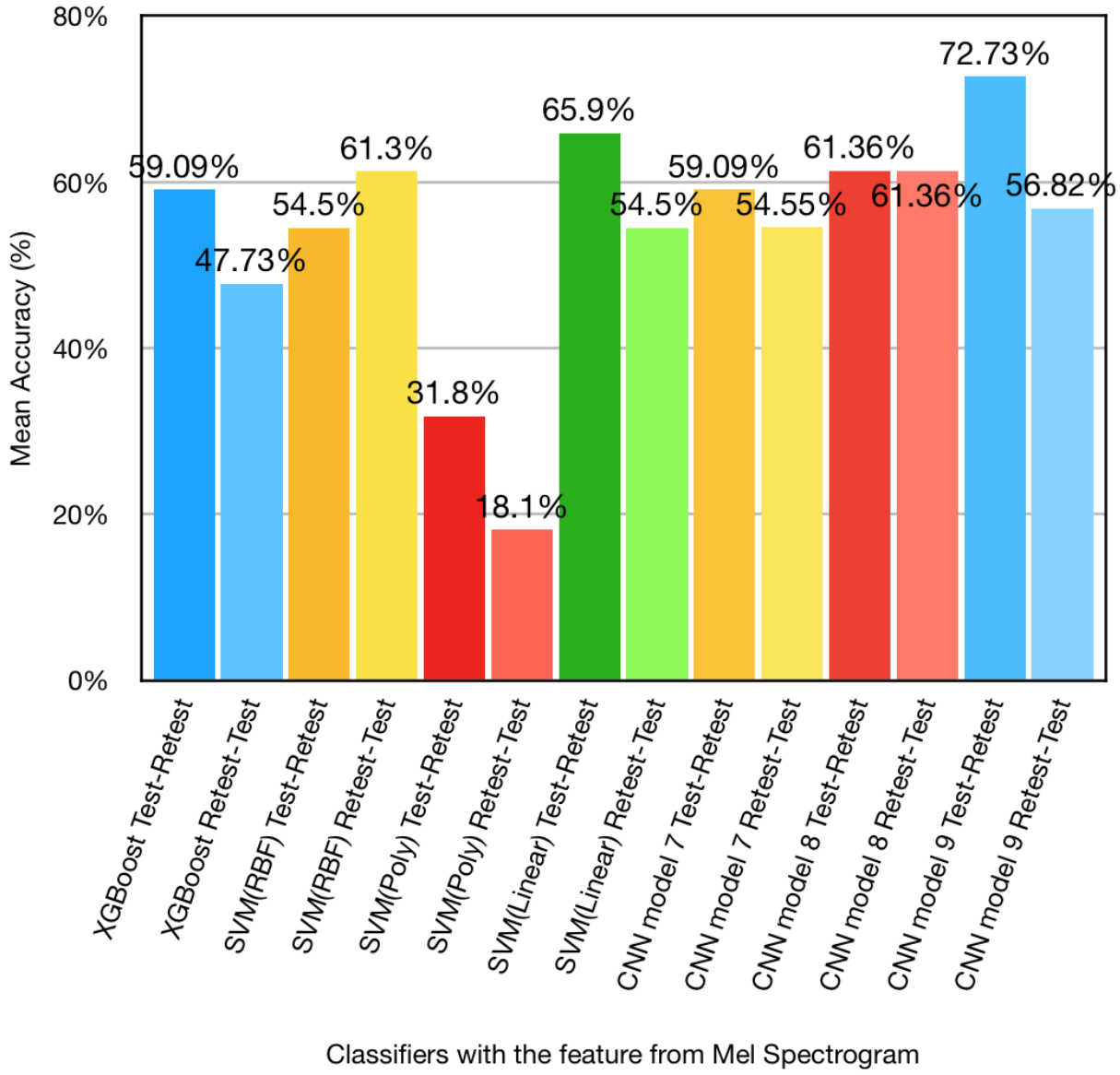

Figure 4-9 : Subject classification performance of each classifier trained on both test and retest session on the Mel spectrogram feature in terms of mean accuracy score.

4.4 Discussion

The classification results indicate that subjects from different test sessions can be predicted well through different classification methods and a variety of time and frequency domain features obtained from the speech-evoked FFR.

The time domain envelope FFRs are obvious feature candidates, as chapter 3 revealed good results using both single vowel and concatenated features.

The classification of the subject-level class membership exhibited a maximum classification accuracy of 86.36% for both Test-Retest and Retest-Test train-test splits, derived from the SVM with linear kernel function with features from the spectrum of size [800, 31] when averaging. Unsurprisingly, it performs 19 times better than chance.

Using the spectrograms as features

The spectrograms of envelope FFR obtained over two different segment lengths and overlap between segments provided the best classification results among all input features. The spectrogram of size [800,31] provides both information about frequency features and the change of signal frequency content over time. In contrast, the spectrogram of size [800,4] only exhibits the frequency features of the four-vowel evoked responses, as it lacks the information about how the frequency appears to be changing over time.

The best obtained classification average accuracy of 86.36% (an average of 86.36% and 86.36%) represents an improvement in performance over the results obtained using time domain series in chapter 3, which provided an average accuracy of 82.95%% (an average of 84.09% and 81.82%). Meanwhile, the overall performance for the spectrogram with size [800,31] is better than for the spectrogram with size [800,4], which indicates that the frequency content over time contains significant information for classification between subjects.

Using Mel spectrograms as features

The Mel spectrogram was selected because of its popularity in speech detection and EEG signal processing (Cantisani et al., 2019; Ofner & Stober, 2018; Yenigalla et al., 2018).

Reducing the dimensionality of the features appears to be helpful in some cases for improving the performance of classifiers. This is evident when comparing the classification accuracy of SVM (polynomial kernel function) with results from other spectrogram features, as the average accuracy increases from 4.5% to 25%. Meanwhile, the performance of the other classifiers is generally lower with the Mel spectrogram than with the standard spectrogram, with an accuracy ranging from 47.73% to 73.73%, which implies that the Mel spectrogram features may lack some of the information useful for discriminating the speech-evoked envelope FFR of subjects.

For the different classifiers used in this study, it is possible to compare their performance with the same features. For XGBoost, the classification performance is higher than chance accuracy but still lower than SVM on most feature sets. When the dimension of the features is reduced, the accuracy of XGBoost shows a slight increase, which may be due to using the same number of trees to fit a smaller number of features.

The performance of the support vector machine is generally very good except when using a polynomial function as kernel. The main reason that the performance of SVM with radial basis kernel function performed worse than others is because the degree d of the kernel function was set to 3, which limits the flexibility of the decision boundary. Overall, the SVM performs with the highest accuracy when using high-dimension features.

The performance of convolutional neural networks rely on parameter tuning to a certain extent. Its performance was limited likely due to the small size of the dataset, even though a simple network structure is used in this study. Unsurprisingly, regularization and dropout were helpful for avoiding overfitting when comparing models. Overall, it is a potentially useful method if the size of the dataset can be expanded in the future.

Overall, the results for machine learning and deep learning methods exhibit good performance from a variety of feature sets extracted from the FFRs of normal hearing subjects. All the classification accuracies exceed chance accuracy, and usually by a very high margin, except SVM with a polynomial kernel function. At the same time, some variation in performance can be observed depending on the classifier and the feature set used.

Furthermore, the ensemble learning, using multiple machine learning algorithms to achieve better predictive performance than any of the constituent learners alone, could be considered to further improve the performance of classification. In our study, the SVM with different kernel function, XGBoost and CNNs can be considered as constituent weak learners as they share the same features for classification. Bootstrap aggregating, abbreviated as bagging, could be used to have each weak leaner in the ensemble vote with a certain weight, which could be considered as future work of this study.

5 Conclusion

5.1 Major Findings

The results of this study provide a fuller understanding of the accuracy with which evoked frequency following responses can be automatically categorized, which supports the feasibility of establishing a subject identification system using these potentials.

The findings of this study follow the three proposed general questions: how the representations of different vowels change in the FFR from different normal hearing adults; whether these representations differ between subjects and are stable across test and retest conditions; and whether or not machine learning and deep learning algorithms can discriminate between FFR responses from different subjects with good accuracy.

The FFR from the presented vowel stimuli provides relatively robust and discernable representations through envelope FFR and spectral FFR in both time and frequency domains. However, the envelope FFR presents more discernable features between individual subjects than the spectral FFR.

The consistency of the signal from a given subject between the test and retest sessions was investigated using a variety of rigorous tests. The Euclidean distance, Pearson Correlation Coefficient and an improved version of Pearson Correlation Coefficient (modified specifically for both the FFR amplitude spectra and the concatenation of time and frequency domain representations of the FFR) were proposed in order to determine the features that can maximize the discernibility of evoked responses among normal hearing subjects. In order to ensure the quality of the signal, the evoked response collected on the same day and from the same subject was verified based on three aspects, namely the tonality coefficient, Pearson Correlation Coefficient and a new proposed peak to noise ratio.

A variety of classification algorithms were shown capable of providing reasonably good classification accuracies, through subject-level classification tasks trained and tested on features extracted from the envelope FFRs of normal hearing subjects. The best classification accuracy of 86.36% was obtained using the support vector machine with radial basis function with the choice of spectrogram as input feature. The performance of the classification algorithms with speech-evoked FFR features supports the potential for establishing identification systems that make use of automatic classification of FFRs.

5.2 Limitations and Future Work

There are several limitations to the work presented here. One of the limitations relates to the fact that only one sound level of the synthetic speech stimuli was utilized, and all the speech stimuli possessed an identical fundamental frequency at 100 Hz. This fundamental frequency is typical for an adult male speaker. Ideally, the evoked responses to natural stimuli with various sound levels across the range of fundamental frequencies typical of human speakers would be selected and analyzed in order to derive a more general comprehension of the FFR across different listeners.

Another limitation is that only the amplitude of the spectra of the evoked responses was considered and analyzed, without taking into account the phase. For example, when making comparison between test and retest conditions, the PCC only made use of the amplitude information of the response without phase information. Both amplitude and phase information may be significant features for subject classification. A future comparison test based on correlation coefficient for complex-valued spectra could be considered.

The result of this study only considers healthy adults with normal hearing. Thus, it may not generalize well to other populations in practice. Other physical or sociocultural factors are

excluded from consideration such as age, hearing loss, musical skills and multilingualism, which have been mentioned or considered in other FFR-related papers. If a diverse population group, including factors like different age groups, subjects with normal hearing and subjects with hearing-loss, native and non-native language speakers, were included in the 22 subjects used for classification, it is assumed that a higher diversity of the features would discriminate the subjects more easily, which would help to improve the performance of classifiers. On the other hand, the training performed with one group may not generalize well to another group (with a higher diversity) not represented in our study, which could be considered in a future study.

Further expansion of the dataset in the future would be helpful to increase the ratio of data size to the number of labels (classes, subjects). In particular a larger dataset could help to improve the classification performance of convolutional neural networks, as these usually rely on a large dataset to avoid overfitting and capture the inherent data distribution more effectively. Furthermore, a larger dataset would allow more flexibility in partioning the data into training and testing sets.

Finally, it should be emphasized that the classification problem we tackled in this study was to identify one subject correctly from a closed set of 22 subjects. In some real-world applications, such as unlocking a personal device, the problem is different. In that case, the system should produce a binary result indicating whether the biometric signal belongs to the intended user or not. In principle, this may be an easier classification problem because it may be possible to accurately learn the characteristics of the biometric signal of the intended user over time, and to more readily differentiate it from another unknown user. Within such a system, it may be possible to obtain higher accuracies with speech-evoked FFRs than those demonstrated in this study. However, such open set recognition tasks include samples at testing which can be quite different from any sample observed at training, which is also a

challenge. Informed forgeries are another potential challenge, although by their nature speech-evoked FFRs would likely be robust to this.

References

Aiken, S. J., & Picton, T. W. (2008a). Envelope and spectral frequency-following responses to vowel sounds. *Hearing Research*, *245*(1–2), 35–47. https://doi.org/10.1016/j.heares.2008.08.004

Aiken, S. J., & Picton, T. W. (2008b). Human cortical responses to the speech envelope. *Ear and Hearing*, *29*(2), 139–157. https://doi.org/10.1097/AUD.0b013e31816453dc

Akhoun, I., Gallégo, S., Moulin, A., Ménard, M., Veuillet, E., Berger-Vachon, C., Collet, L., & Thai-Van, H. (2008). The temporal relationship between speech auditory brainstem responses and the acoustic pattern of the phoneme /ba/ in normal-hearing adults. *Clinical Neurophysiology*, *119*(4), 922–933. https://doi.org/10.1016/j.clinph.2007.12.010

Ananthakrishnan, S., Krishnan, A., & Bartlett, E. (2016). Human Frequency Following Response: Neural Representation of Envelope and Temporal Fine Structure in Listeners with Normal Hearing and Sensorineural Hearing Loss. *Ear and Hearing*, *37*(2), e91–e103. https://doi.org/10.1097/aud.0000000000000247

Bidelman, G. M. (2018). Sonification of scalp-recorded frequency-following responses (FFRs) offers improved response detection over conventional statistical metrics. *Journal of Neuroscience Methods*, *293*, 59–66. https://doi.org/10.1016/j.jneumeth.2017.09.005

Bidelman, G. M., & Powers, L. (2018). 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*, *57*(9), 665–672. https://doi.org/10.1080/14992027.2018.1470338

Billings, C. J., Tremblay, K. L., Souza, P. E., & Binns, M. A. (2007). Effects of hearing aid amplification and stimulus intensity on cortical auditory evoked potentials. *Audiology and Neurotology*, *12*(4), 234–246. https://doi.org/10.1159/000101331

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). Training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, *October 2015*, 144–152. https://doi.org/10.1145/130385.130401

Cantisani, G., Essid, S., & Richard, G. (2019). EEG-Based Decoding of Auditory Attention To a Target Instrument in Polyphonic Music. *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, *765068*, 80–84. http://www.tsi.telecom-paristech.fr/aao/en/

Chandrasekaran, B., & Kraus, N. (2010). The scalp-recorded brainstem response to speech: Neural origins and plasticity. *Psychophysiology*, *47*(2), 236–246. https://doi.org/10.1111/j.1469-8986.2009.00928.x

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785

Chimento, T. C. (1990). Selectively eliminating cochlear microphonic contamination from the frequency-following response. *Electroencephalograph and Clinical Neurophysiology*, 88–96.

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1023/A:1022627411411

Cunningham, J., Nicol, T., Zecker, S. G., Bradlow, A., & Kraus, N. (2001). Neurobiologic responses to speech in noise in children with learning problems: Deficits and strategies for improvement. *Clinical Neurophysiology*, *112*(5), 758–767. https://doi.org/10.1016/S1388-2457(01)00465-5

Dolphin, W. F., & Mountain, D. C. (1992). The envelope following response: Scalp potentials elicited in the mongolian gerbil using sinusoidally AM acoustic signals. *Hearing*

Research, *58*(1), 70–78. https://doi.org/10.1016/0378-5955(92)90010-K

Douglas, M., Bailey, K., Leeney, M., & Curran, K. (2018). An overview of steganography techniques applied to the protection of biometric data. *Multimedia Tools and Applications*, *77*(13), 17333–17373. https://doi.org/10.1007/s11042-017-5308-3

Dubnov, S. (2004). Generalization of spectral flatness measure for non-Gaussian linear processes. *IEEE Signal Processing Letters*, *11*(8), 698–701. https://doi.org/10.1109/LSP.2004.831663

Furui, S. (1997). Recent advances in speaker recognition. *Pattern Recognition Letters*, *1206*, 859–872. https://doi.org/10.1016/S0167-8655(97)00073-1

Golding, M., Pearce, W., Seymour, J., Cooper, A., Ching, T., & Dillon, H. (2007). The relationship between obligatory cortical auditory evoked potentials (CAEPs) and functional measures in young infants. *Journal of the American Academy of Audiology*, *18*(2), 117–125. https://doi.org/10.3766/jaaa.18.2.4

Greenberg, S., Marsh, J. T., Brown, W. S., & Smith, J. C. (1987). Neural temporal coding of low pitch. I. Human frequency-following responses to complex tones. *Hearing Research*, *25*(2–3), 91–114. https://doi.org/10.1016/0378-5955(87)90083-9

Heffernan, B. (2019). *Characterization and Classification of the Frequency Following Response to Vowels at Different Sound Levels in Normal Hearing Adults*, PhD. thesis, University of Ottawa, Canada, January 2019

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., & Wilson, K. (2017). CNN architectures for large-scale audio classification. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 131–135. https://doi.org/10.1109/ICASSP.2017.7952132

Hill, R. B. (1978). Apparatus and method for identifying individuals through their retinal vasculature patterns . *United States Patent*, *19*, 3–7.

Hong, L., & Jain, A. (1997). Integrating faces and fingerprints for personal identification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *1351*(12), 16–23. https://doi.org/10.1007/3-540-63930-6_99

Huis, F., Osterhammel, P., & Terkildsen, K. (1977). The frequency selectivity of the 500 hz frequency following response. *Scandinavian Audiology*, *6*(1), 35–42. https://doi.org/10.3109/01050397709044996

Jain, A. K., Hong, L., Pankanti, S., & Bolle, R. (1997). An identity-authentication system using fingerprints. *Proceedings of the IEEE*, *85*(9), 1365–1388. https://doi.org/10.1109/5.628674

Jain, Hong, L., & Pankanti, S. (2000). Biometric Identification. *Communications of the ACM*, *43*(2), 91–98. https://doi.org/10.1145/328236.328110

Johnson, K. L., Nicol, T. G., Zecker, S. G., & Kraus, N. (2007). Auditory brainstem correlates of perceptual timing deficits. *Journal of Cognitive Neuroscience*, *19*(3), 376–385. https://doi.org/10.1162/jocn.2007.19.3.376

Johnson, K. L., Nicol, T., Zecker, S. G., Bradlow, A. R., Skoe, E., & Kraus, N. (2008). Brainstem encoding of voiced consonant-vowel stop syllables. *Clinical Neurophysiology*, *119*(11), 2623–2635. https://doi.org/10.1016/j.clinph.2008.07.277

Johnston, J. D. (1988). Transform Coding of Audio Signals Using Perceptual Noise Criteria. *IEEE Journal on Selected Areas in Communications*, *6*(2), 314–323. https://doi.org/10.1109/49.608

King, C., Warrier, C. M., Hayes, E., & Kraus, N. (2002). Deficits in auditory brainstem

pathway encoding of speech sounds in children with learning problems. *Neuroscience Letters*, *319*(2), 111–115. https://doi.org/10.1016/S0304-3940(01)02556-3

Klare, B. F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Burge, M., & Jain, A. K. (2015). 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*, *07-12-June*, 1931–1939. https://doi.org/10.1109/CVPR.2015.7298803

Korczak, P. A., Kurtzberg, D., & Stapells, D. R. (2005). Effects of sensorineural hearing loss and personal hearing aids on cortical event-related potential and behavioral measures of speech-sound processing. *Ear and Hearing*, *26*(2), 165–185. https://doi.org/10.1097/00003446-200504000-00005

Krishnan, A. (1999). Human frequency-following responses to two-tone approximations of steady-state vowels. *Audiology and Neuro-Otology*, *4*(2), 95–103. https://doi.org/10.1159/000013826

Krishnan, A. (2002). Human frequency-following responses: Representation of steady-state synthetic vowels. *Hearing Research*, *166*(1–2), 192–201. https://doi.org/10.1016/S0378-5955(02)00327-1

Laroche, M., Dajani, H. R., Prévost, F., & Marcoux, A. M. (2013). Brainstem auditory responses to resolved and unresolved harmonics of a synthetic vowel in quiet and noise. *Ear and Hearing*, *34*(1), 63–74. https://doi.org/10.1097/AUD.0b013e31826119a1

Lee Rodgers, J., & Alan Nice Wander, W. (1988). Thirteen ways to look at the correlation coefficient. *American Statistician*, *42*(1), 59–66. https://doi.org/10.1080/00031305.1988.10475524

Llanos, F., Xie, Z., & Chandrasekaran, B. (2019). Biometric identification of listener identity

from frequency following responses to speech. *Journal of Neural Engineering*, *16*(5), 056004. https://doi.org/10.1088/1741-2552/ab1e01

Moushegian, G., Rupert, A. L., & Stillman, R. D. (1973). Scalp-recorded early responses in man to frequencies in the speech range. *Electroencephalography and Clinical Neurophysiology*, *35*(6), 665–667. https://doi.org/10.1016/0013-4694(73)90223-X

Narayanan, A., & Wang, D. (2013). Ideal ratio mask estimation using deep neural networks for robust speech recognition. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 7092–7096. https://doi.org/10.1109/ICASSP.2013.6639038

Ofner, A., & Stober, S. (2018). Shared generative representation of auditory concepts and EEG to reconstruct perceived and imagined music. *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018*, 392–399.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep Face Recognition. In *British Machine Vision Conference* (Issue Section 3, pp. 41.1-41.12). https://doi.org/10.5244/c.29.41

Picton, T. W. (2010). *Human Auditory Evoked Potentials*. Plural Publishing INC.

Plyler, P. N., & Ananthanarayan, A. K. (2001). Human frequency-following responses: Representation of second formant transitions in normal-hearing and hearing-impaired listeners. *Journal of the American Academy of Audiology*, *12*(10), 523–533.

Rance, G., Cone-Wesson, B., Wunderlich, J., & Dowell, R. (2002). Speech perception and cortical event related potentials in children with auditory neuropathy. *Ear and Hearing*, *23*(3), 239–253. https://doi.org/10.1097/00003446-200206000-00008

Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects.

Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, *336*(1278), 367–373. https://doi.org/10.1098/rstb.1992.0070

Russo, N., Nicol, T., Musacchia, G., & Kraus, N. (2004). Brainstem responses to speech syllables. *Clinical Neurophysiology*, *115*(9), 2021–2030. https://doi.org/10.1016/j.clinph.2004.04.003

Sadeghian, A., Dajani, H. R., & Chan, A. D. C. (2011). Classification of English vowels using speech evoked potentials. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 5000–5003. https://doi.org/10.1109/IEMBS.2011.6091239

Sadeghian, A., Dajani, H. R., & Chan, A. D. C. (2015). Classification of speech-evoked brainstem responses to English vowels. *Speech Communication*, *68*, 69–84. https://doi.org/10.1016/j.specom.2015.01.003

Sanfins, M. D., Garcia, M. V., Biaggio, E. P. V., & Skarzynski, H. (2012). The Frequency Following Response: Evaluations in Different Age Groups. *Intech*, 13. https://doi.org/10.1016/j.colsurfa.2011.12.014

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-ryan, R. J., Saurous, R. A., Agiomyrgiannakis, Y., & Wu, Y. (2018). Natural TTS Synthesis By Conditioning Wavenet On Mel Spectrogram Predictions. 4779–4783. https://doi.org/10.1109/ICASSP.2018.8461368

Skoe, & Kraus. (2010). Auditory brainstem response to complex sounds: a tutorial Erika. *Ear Hear*, *31*(3), 302–324. https://doi.org/10.1142/9789814623094

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 1929–1958. https://doi.org/10.1016/0010-4361(73)90803-3

Stapells, D. R., Makeig, S., & Galambos, R. (1987). Auditory steady-state responses: threshold prediction using phase coherence. *Electroencephalography and Clinical Neurophysiology*, *67*(3), 260–270. https://doi.org/10.1016/0013-4694(87)90024-1

Strait, D. L., Kraus, N., Skoe, E., & Ashley, R. (2009). Musical experience and neural efficiency - Effects of training on subcortical processing of vocal expressions of emotion. *European Journal of Neuroscience*, *29*(3), 661–668. https://doi.org/10.1111/j.1460-9568.2009.06617.x

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the gap to human-level performance in face verification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1701–1708. https://doi.org/10.1109/CVPR.2014.220

Wei, X.-S., Xie, C.-W., & Wu, J. (2016). *Mask-CNN: Localizing Parts and Selecting Descriptors for Fine-Grained Image Recognition*. http://arxiv.org/abs/1605.06878

Won, J. H., Tremblay, K., Clinard, C. G., Wright, R. A., Sagi, E., & Svirsky, M. (2016). The neural encoding of formant frequencies contributing to vowel identification in normal-hearing listeners. *The Journal of the Acoustical Society of America*, *139*(1), 1–11. https://doi.org/10.1121/1.4931909

Yellamsetty, A., & Bidelman, G. M. (2019). Brainstem correlates of concurrent speech identification in adverse listening conditions. *Brain Research*, *1714*(October 2018), 182–192. https://doi.org/10.1016/j.brainres.2019.02.025

Yenigalla, P., Kumar, A., Tripathi, S., Singh, C., Kar, S., & Vepa, J. (2018). Speech emotion recognition using spectrogram & phoneme embedding. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2018-Septe*(September), 3688–3692. https://doi.org/10.21437/Interspeech.2018-1811

Yi, H. G., Xie, Z., Reetzke, R., Dimakis, A. G., & Chandrasekaran, B. (2017). Vowel decoding from single-trial speech-evoked electrophysiological responses: A feature-based machine learning approach. *Brain and Behavior*, *7*(6), 1–8. https://doi.org/10.1002/brb3.665