Classification of Speech Evoked Responses to English Vowels

By

Amir Sadeghian

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of

Masters of Applied Science
in
Biomedical Engineering

Ottawa-Carleton Institute for Biomedical Engineering

Department of Systems and Computer Engineering
Carleton University
Ottawa, Ontario, Canada
August 2012

The undersigned recommend to the
Faculty of Graduate Studies and Research
acceptance of the thesis

Classification of Speech Evoked Responses to English Vowels

Submitted by Amir Sadeghian
in partial fulfillment of the requirements for
the degree of Master of Applied Science in Biomedical Engineering

Adrian D. C. Chan

Hilmi R. Dajani

Chair, Department of Systems and Computer Engineering
Dr. Howard Schwartz

2012, Carleton University

Abstract

The objective of this study is to investigate whether brainstem Speech Evoked Potentials (SpEPs), contain information that can be used to distinguish different speech stimuli. We used the five English language synthetic vowels as stimuli, and investigated the difference between their SpEPs by looking at the features contained in two types of responses: 1) the transient response which reflects the response to the onset of the stimulus, and 2) the sustained response which follows the acoustical features of the periodic speech stimulus. These features include the pitch, which is reflected in the Envelope Following Response (EFR), and the formants, which are reflected in the Frequency Following Response (FFR). We performed linear discriminant analysis and obtained classification accuracies of 38.33% for transient response features and 80.33% for sustained response features. This result demonstrates that brainstem neural responses in the region of onset, F0, and F1 contain valuable information for discriminating vowels.

Statement of Originality

This thesis describes the results of the author's research conducted at Carleton University during the course of the Master of Applied Science (M.A.Sc) program. Portions of this research have been disseminated in the following publications,

Sadeghian A, Dajani HR, Chan ADC, "Classification of English vowels using speech evoked potentials", *Proceedings of the 32nd Annual International Conference of the IEEE-EMBS*, Boston MA, USA, pp. 5000-5003, 2011.

The results of this conference paper constitute the first half of Chapter 4 and Chapter5. The author collected all auditory response samples, performed the data analysis, prepared the manuscript for publication, and made all necessary revisions based on feedback from the co-authors and the paper reviewers. The data collection was performed at the University of Ottawa Health Campus, in the School of Rehabilitation Sciences (Audiology Program).

Acknowledgements

I would like to thank my supervisors Dr. Adrian D.C. Chan and Dr. Hilmi Dajani sincerely for giving me the opportunity to perform this study under their supervision. Without their guidance, support, and feedback this work would not be completed.

I would also like to thank my lovely wife Morvarid for her enormous support.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ABR	Auditory Brainstem Response
ANOVA	Analysis of Variance
CN	Cochlear Nucleus
EFR	Envelope Following Response
FFR	Frequency Following Response
F0	Fundamental (Pitch) frequency
F1	First formant
F2	Second formant
F3	Third formant
ICC	Inferior Colliculus
LDA	Linear Discriminant Analysis
MGB	Medial Geniculate Body
SpEPs	Speech Evoked Potentials
ULDA	Uncorrelated Linear Discriminant Analysis

1 Introduction

1.1 Motivation

Over the past few decades, there has been a strong interest in analyzing brain signals in response to various stimuli, particularly images and sounds. The ultimate goal is to understand such brain processing in order to be able to come up with better clinical tools for both the diagnosis and treatment of sensory and cognitive impairments. For instance, a possible clinical application would be to create a thought-reading system for individuals with communication disabilities to help them to communicate with outside world. A use-case for this system would be to recognize a segment of speech that those individuals intend to say and turn it into audio signal or text.

As part of the effort to achieve the aforementioned goal, a recent study on the brain's visual processing demonstrated that it is possible to partially reconstruct a short natural movie from the blood flow patterns of participants' brain (Nishimoto et al., 2011). A similar study was performed to model auditory neural processing using speech stimuli. It has shown that different English words and sentences being presented to 15 subjects can be reconstructed with an accuracy of up to about 50% from intracranial recordings (i.e. direct recordings from the surface of the auditory cortex) (Pasley et al., 2012). Although these studies have made significant contributions towards modeling brain's neural processing, there still remains a long way to understanding it fully. The focus of this thesis is to provide a better understanding of auditory neural processing of speech by identifying distinctive features of the brainstem signals non-invasively recorded in human subjects using surface electrodes. The result of this study can help to improve the existing models for the brain's auditory processing.

The auditory system is a complex sensory system and there has been a lot of effort to understand how it works in order to be able to diagnose and treat auditory related disorders such as hearing impairments and language learning problems that result from central auditory processing impairment (Møller, 2006a). In the 1967, it was found that Auditory Brainstem Responses (ABRs) can be measured from the scalp using surface electrodes. ABRs are generated from the synchronous electrical activity of auditory neural system at the brainstem in response to sounds. Analyzing ABRs to simple short duration stimuli (i.e. clicks and tone bursts) revealed that ABRs reflect auditory neural activity along the ascending auditory pathway.

When the stimulus signal is more complex comprising a periodic or quasi-periodic component, such as a pure vowel or a consonant-vowel, the ABR can be divided into a transient response and a sustained steady-state response (Jewett et al., 1970; Jewett and Williston, 1971). The transient response refers to the initial non-periodic component of the ABR (usually up to 20 ms) after the onset of the stimulus. The sustained response is an additional response that is formed after the transient response. Further research on ABRs has shown that hearing thresholds and some auditory neural malfunctions can be determined from the peak amplitudes and inter-peak latencies, particularly in the transient response (Hecox and Galambos, 1974; Starr and Anchor, 1975). As a result, the use of ABRs to clicks and tone bursts has become a key tool for clinicians and researchers for diagnosing hearing impairments and understanding the human auditory neural system (Burkard et. al., 2007).

Although ABRs to simple artificial stimuli have been widely used to study auditory neural processing, they provide a poor understanding about the processing of complex stimuli such as speech sounds. This is due to the fact that acoustical information of complex sounds is mainly encoded in the sustained part of ABRs (Skoe and Kraus, 2010). Greenberg used speech as stimuli and showed that auditory neural responses to speech formants are present in the ABRs (Greenberg, 1980).

Additional studies have demonstrated that speech-evoked ABRs (referred to in this thesis as Speech Evoked Potentials or SpEPs) provide additional information regarding the state of the central auditory neural system, which can be used to help children with language and learning problems and potentially in other populations with central processing disorders (Russo et al., 2007; Johnson et al., 2008).

Previous studies on SpEPs have focused on understanding the underlying auditory neural activity during speech processing, the origin of SpEPs, and new techniques for diagnosis of hearing impairment (Kraus and Nicol, 2005; Martin et al., 2008). A possible clinical application of these studies would be in the assessment of infants suspected of hearing impairment. Currently hearing assessment is limited by diagnostic tests, which usually employ artificial signals like tones or clicks (i.e. simple stimuli) that do not allow a clear assessment of auditory function for speech communication. While there are tests of speech perception that rely on subjective responses, these are of no value for assessing the hearing of infants and uncooperative individuals. SpEPs could thus fill the need to objectively assess auditory performance in these cases. Another clinical application would be to optimize technologies used in hearing aids and cochlear implants. The fitting of existing technologies is often based on aforementioned diagnostic tests, using simple stimuli, which do not allow for selective acoustic treatments (Johnson et. al., 2005). Currently, however, there is limited understanding of SpEPs and how they relate to processing of different speech sounds.

1.2 Thesis Objectives

The objective of the research in this thesis is to assess the acoustical information contained in SpEPs in order to better understand speech encoding in the auditory system. We have performed this assessment by applying a basic classification method on two different SpEP feature sets: 1) temporal features of the transient

response, 2) spectral features of the sustained response. We provided a quantitative measure for discriminating SpEPs using transient and sustained response features. Moreover, we investigated the relation between different speech stimuli and their corresponding SpEPs. To the best of our knowledge, this work is the first attempt in speech recognition using SpEPs. The next section provides more details about contributions made in this thesis.

1.3 Contributions

The major contributions towards understanding the auditory neural processing from this thesis research are:

1. **Demonstrated that SpEPs from five English vowels carry sufficient spectral and temporal information for classifying the SpEPs**

   We were able to classify SpEPs from five English vowels with a good accuracy using spectral and temporal features of SpEPs separately. This result indicates that SpEPs contain useful speech encoding information which can be used to better understand the functionality of the auditory system. Also, the classification result provides a quantitative measure on how much speech-specific information is available in both spectral and temporal features. To the best of our knowledge, this is the first attempt in speech recognition using SpEPs measured using surface electrodes.

2. **Demonstrated that the brainstem response to both speech envelope and speech formants can be used to classify SpEPs from five English vowels**

   From the sustained steady-state response we chose two amplitude spectral feature sets corresponding to brainstem responses to source and filter

characteristics of speech (i.e. speech envelope and speech formants). The classification results showed that both feature sets provided high classification accuracy; however, the source-related features provide higher accuracy than the filter-related features. This is a novel finding because it is generally thought that the filter characteristics of speech make the main contribution to perceptual discrimination of different vowels (Peterson and Barney; 1952).

3. **Demonstrated that the transient response features can be used to classify SpEPs from five English vowels**

   We were able to classify SpEPs of five English vowels using temporal features of the transient response. We found that the latency and amplitude of significant peaks carry speech-specific information. This is a potentially important finding because the transient response to a vowel has been thought to carry general sound onset information and not vowel-specific information.

4. **Collected SpEPs in response to five English vowels**

   We collected SpEPS of five English vowels (\a\,\ae\,\e\,\i\,\u\) from 8 subjects with the specifications which are explained in section 3.3. These data can be used in future studies of SpEPs.

Portions of the research have been disseminated in the following publication:

Sadeghian A, Dajani HR, Chan ADC, "Classification of English vowels using speech evoked potentials", *Proceedings of the 32nd Annual International Conference of the IEEE-EMBS*, Boston MA, USA, pp. 5000-5003, 2011.

1.4 Organization of Thesis

This thesis consists of six chapters. The remaining five chapters are organized as follows,

- In Chapter 2, we provide an overview of the human auditory system, the structure of SpEPs. We also provide a discussion of related works, and outline

the main differences between previous work and our work. This chapter provides fundamental background information to help understand the following chapters.

- In Chapter 3, we describe methodologies used in various parts of this work. We explain techniques used for generating stimuli and we provide details on experimental procedure. Moreover, we discuss data analysis approaches, including features extraction and classification methods.
- In Chapter 4, we present the results of the SpEP classification. In addition, we provide additional data analysis that supports the results.
- In Chapter 5, we discuss the results presented in Chapter4 and explain our novel findings. In addition, we provide explanations for misclassified samples.
- In Chapter 6, we conclude with a brief summary of the work presented in this thesis along with a discussion on possible enhancements in the future work.

2 Background & Literature Review

2.1 Anatomy of the Auditory System

2.1.1 The Structure of the Ear

The ear consists of three main structures: 1) the outer ear, 2) the middle ear, and 3) the inner ear. Figure 2-1 shows the simplified anatomy of the human ear. The outer ear includes the Pinna which re-shapes a sound stimulus to provide additional information to help brain for sound localization and the external auditory canal that amplifies sounds within a frequency range between 3 and 12 kHz. The middle ear consists of three small bones (Malleus, Incus, and Stapes) that act as an impedance transformer from air to fluid. The middle ear is separated from the outer ear by the Tympanic membrane (Eardrum) which transmits acoustic energy from air to the three bones. The snail-shaped inner ear (the Cochlea) contains the organ of Corti in which hair cells transduce mechanical fluid waves into electrical nerve signals that travel through cochlear nerve (Møller, 2006a).

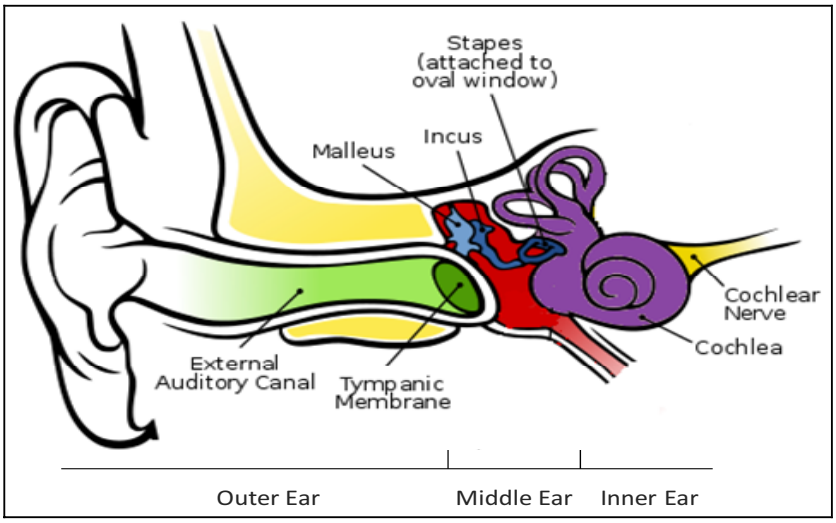

Figure 2-1: Anatomy of the human ear consisting of three main parts, outer, middle, and inner ear (Adapted from the Wikimedia Commons file "Anatomy_of_the_Human_Ear.svg" http://upload.wikimedia.org/wikipedia/commons/d/d2/Anatomy_of_the_Human_Ear.svg).

2.1.2 The Auditory Nervous System

The central auditory nervous system connects the cochlear nucleus, which carries nerve signals from the cochlea in the inner ear, to the auditory cortex where the sound signal is processed. The central auditory neural system consists of two pathways: 1) the ascending auditory pathway and 2) the descending auditory pathway. The ascending pathway describes the auditory neural centres that link sensory information from the ear to higher centres in the brain, whereas the descending pathway provides feedback from higher neural centres to lower neural centres and eventually to the ear (Møller, 2006b). The following two subsections (2.1.2.1 and 2.1.2.2) describe each of the two pathways in further detail.

2.1.2.1 The Ascending Auditory Pathways

The ascending auditory pathways consist of two neural systems, 1) the classical pathway and 2) the non-classical pathway. Both pathways have connections at different auditory neural levels up to the cortex but the classical pathway is thought to dominate auditory sensory processing and involves parallel and hierarchical organization (Møller, 2006b). On the other hand, the non-classical pathway has connections to the somatosensory system (e.g., touch and pain); however, its functionality in auditory processing is not well understood (Møller and Rollins, 2002). As such, it is assumed that it plays a less important role in auditory sensory processing. Therefore, for the purposes of this study, we will focus on the classical ascending auditory pathway and in particular the known neural connections between the inner ear and the brainstem (Møller, 2006b).

The classical ascending auditory pathway for both ears can be viewed as two partly independent systems that have neural connections at different brains levels via several nuclei. As sound information travels from the inner ear to the cortex, it passes through intermediate nuclei where various types of auditory processing occur.

Not all auditory neural fibres go through every nucleus; however, all auditory neural fibres get interrupted by three of the nuclei in ascending order namely, 1) the cochlear nucleus (CN), 2) the central nucleus of the contralateral inferior colliculus (ICC) in the upper brainstem, and 3) the medial geniculate body (MGB). Figure 2-2 shows the simplified schematic of the classical ascending auditory pathways in which the main auditory neural connections are shown at different brain levels.

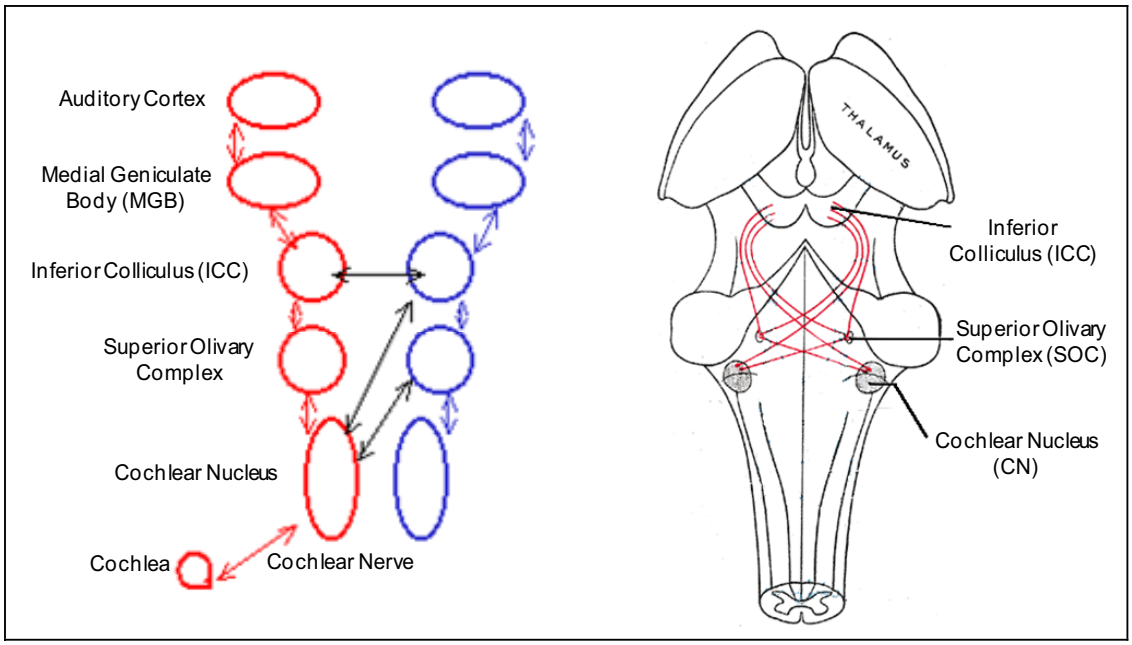

Figure 2-2: Simplified schematic of the classical ascending auditory pathways (Adapted from the Wikimedia Commons file "Anatomy_of_the_Human_Ear.svg" http://upload.wikimedia.org/wikipedia/commons/d/d2/Anatomy_of_the_Human_Ear.svg).

2.1.2.2 The Descending Auditory Pathways

The descending auditory pathway can be viewed as a parallel counterpart to the ascending pathway. It consists of two separate systems known as the corticofugal system and the olivocochlear system. The corticofugal system originates in the primary auditory cerebral cortex and reaches the inferior colliculus. The olivocochlear system starts from the superior oilvary complex (located in the pons) and end at the hair cells of the cochlea (Møller, 2003). Studies show that the descending pathway

may help in speech decoding or in musical perception. Moreover, these pathways can be effective in enhancing particular acoustic features such as the extracted pitch.

2.2 Acoustics of Vowels

Speech generation by the human vocal apparatus can be described by the source-filter model (Kraus and Nicol, 2005). According to this model, the sound source refers to the vibration of the vocal folds reacting to airflow that is generated by the lungs whereas the filter refers to the transfer function of the vocal cavity and organs such as the tongue, lips, and jaw, which shape the spectrum of speech. The sound source determines the fundamental frequency of speech and consequently the pitch of speech. In English and other non-tonal languages, the source mainly characterizes nonlinguistic information relating to speaker identity and prosody. On the other hand, the filter determines the formant frequencies (i.e. resonant peaks of the filter) and so is believed to determine the linguistic content of speech (Johnson et al., 2005; Kraus and Nicol, 2005). In general, the fundamental frequency (F0) depends on the gender, age, and emotion of the speaker, and ranges from about 75 to 175Hz for males and about 175 to 300Hz for females (Greenberg and Ainsworth, 2004). Moreover, formant frequencies change with gender and age as well because the filter properties vary (Peterson and Barney, 1952). Previous studies show that generally, formants are lower in men than in women (Peterson and Barney, 1952). To illustrate the source-filter model of speech production, the following two figures show the time domain waveform and spectrum of a synthetically generated vowel \a\ as spoken by a male. Figure 2-3 shows the first 100 ms of the vowel \a\ with a sound source fundamental period of T0=10ms (Appendix A illustrates five English vowels in the time domain). Figure 2-4 shows the amplitude spectra of the vowel \a\ up to 2000Hz with sound source fundamental frequency of F0=100Hz (F0=1/T0). The first two formants, F1=700Hz and F2=1200Hz, are the resonance peaks of the filter.

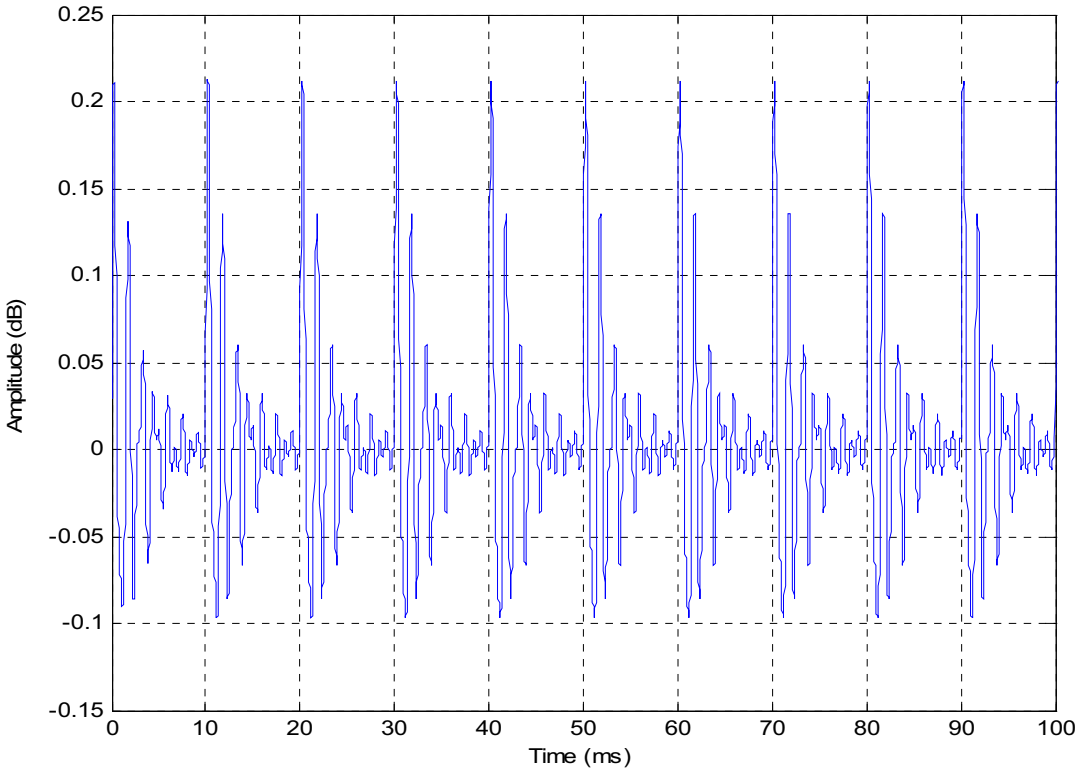

Figure 2-3: Time domain representation of the synthetic vowel \a\ as spoken by a male with T0=10ms.

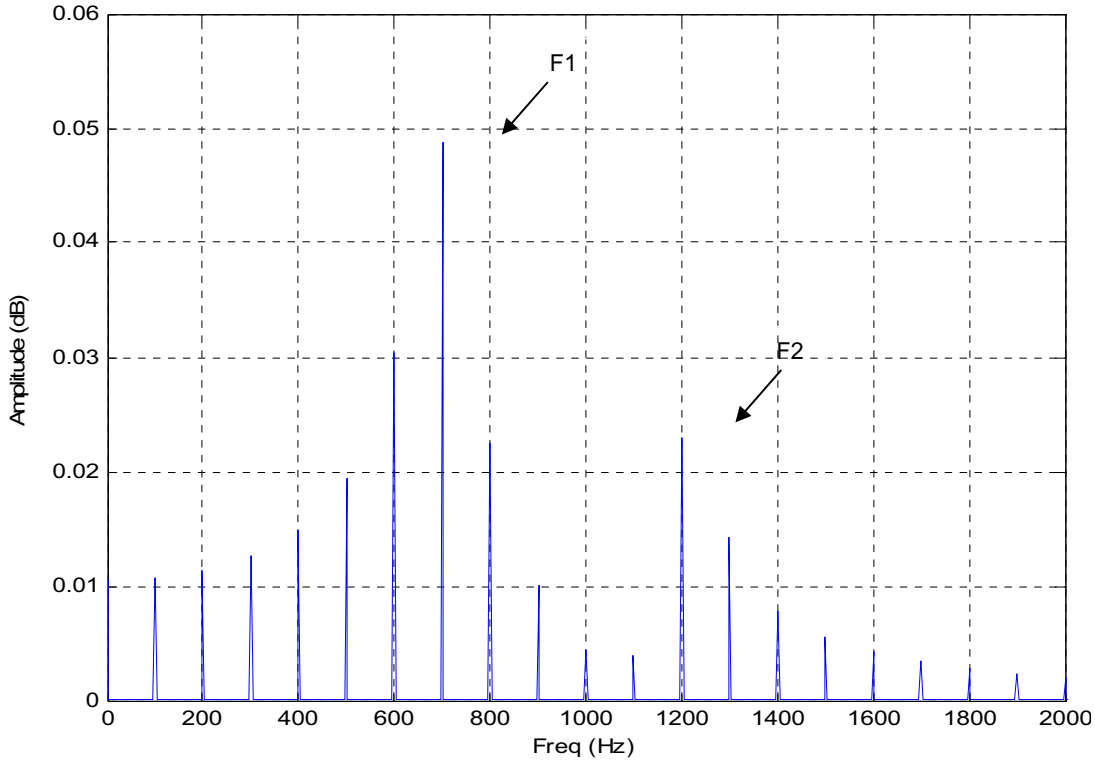

Figure 2-4: Amplitude spectra of synthetic vowel \a\ as spoken by a male with F0=100Hz, F1=700Hz and F2= 1220Hz.

Previous studies have shown that the first three formants (F1, F2, and F3) are the dominant formants in terms of intensity. In English, vowel identification is possible using only the first two formants (Peterson and Barney, 1951). Figure 2-5 shows the frequency plot of first formant versus second formant for five English vowels spoken by 76 speakers. Each point represents the average F1-F2 coordinate for a vowel and the boundary around it illustrates the distribution of that vowel where about 90% of F1-F2 values occurred. As can be seen, the vowels can be easily separated using the first two formants.

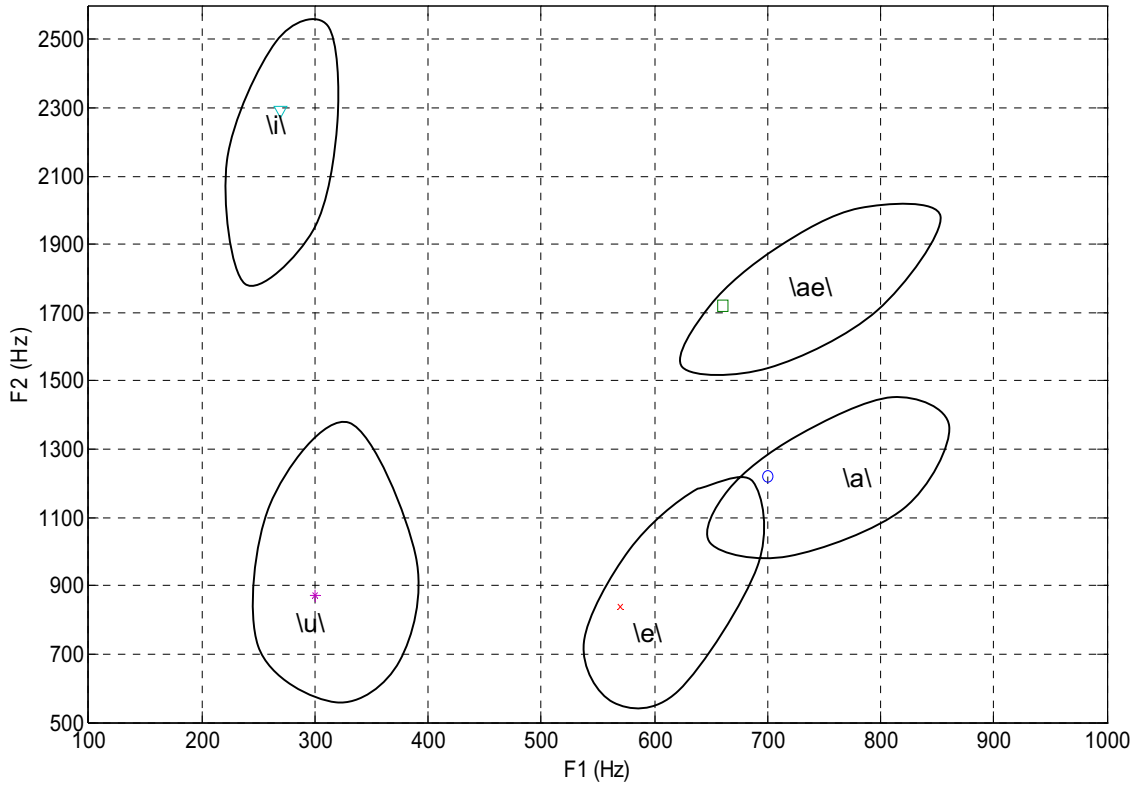

Figure 2-5: Frequency of first formant versus second formant for five English vowels. Each boundary illustrates about 90% distribution of a single vowel and the point inside the boundary is the average F1-F2 coordinate for that vowel (based on Peterson and Barney, 1952).

2.3 Speech Encoding in the Auditory System

Two theories have been proposed to explain speech encoding in the auditory system: 1) place coding theory and 2) temporal coding theory. It is thought that a combination of these two mechanisms contribute to speech encoding as each of them alone has its own limitations (Møller, 2006b). The place coding theory states that different frequencies are perceived along a spatial array of neurons throughout the auditory system. This is due to the fact that the hair cells are arranged tonotopically along the basilar membrane in the cochlea (i.e. respond best to one frequency or a narrow range of frequencies), to generate neural action potentials. This arrangement is preserved in the auditory nuclei throughout the ascending auditory pathway (Aiken, 2008; Møller, 2006b). As is shown in Figure 2-6, the basilar membrane is more sensitive to low frequencies at the distal region (the thicker area) and it becomes more sensitive to higher frequencies towards the base region (the thinner area).

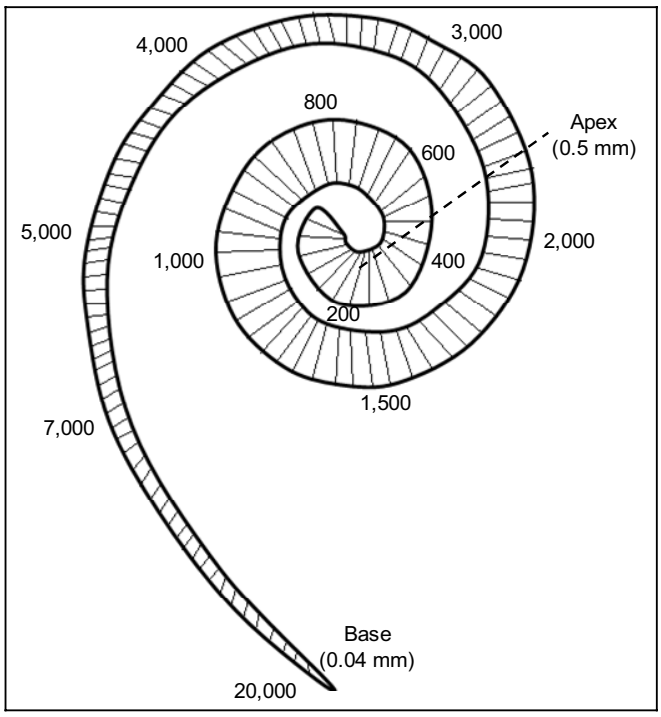

Figure 2-6: Basilar membrane with frequency selectivity in different regions (based on Møller, 2006b).

Although the place coding theory can generally describe speech encoding in the auditory system, there are cases that cannot be explained by this theory. For instance, it does not explain why normal speech perception is possible when the pitch frequency is suppressed or when the speech is presented at a high level (above 70 dB) that saturates the firing rate in auditory neurons. The temporal coding theory addresses these limitations.

Temporal coding theory suggests that frequency discrimination is performed through phase-locking of auditory neural activity to frequency components of the stimulus signal, which usually correspond to frequencies with highest spectral magnitude (Greenberg and Ainsworth, 2004; Møller, 2006b). Phase-locking means that neurons synchronously fire at a particular phase angle of the periodic components of the speech stimulus. Phase-locking is strong for frequencies below 1 kHz and it starts degrading for higher frequencies up to about 4 kHz, where it becomes non-existent (Rhode and Greenberg, 1994; Assmann and Summerfield, 2004; Greenberg and Ainsworth, 2004; Bear et al., 2007). For frequencies above 4 kHz, speech encoding is thought to be achieved via place coding (Griffiths et al., 2001; Bartlett and Wang, 2007; Krishnan and Gandour, 2009). Previous studies have shown that the main speech components, such as the fundamental frequency and low frequency formants, are reflected in the temporal patterns of auditory neural responses. For example, a few studies have shown that formant peaks are preserved in nerve fibre responses recorded invasively in animals at harmonics near the formant frequencies (Delgutte and Kiang, 1984; Sachs and Young, 1980), and in compound auditory brainstem responses recorded non-invasively in humans (Krishnan, 2002).

2.4 Speech Evoked Potentials (SpEPs)

Speech evoked potentials (i.e. SpEPs), and specifically the auditory brainstem response to vowel stimuli, consist of two parts: 1) transient response and 2) sustained response. The transient response is short (< 20 ms) and is similar to the transient response to click stimuli. As such it may be thought to be an undifferentiated response to stimulus onset (Skoe and Kraus, 2010). However, it is also thought to depend on the "attack" characteristics of the stimulus (i.e. how sharply and suddenly it starts). The transient response can differ depending on the initial consonant, when the stimulus is consonant-vowel stimulus (Johnson et al, 2008; Skoe and Kraus, 2010), but there has been no previous work on whether it is able to convey phonetic information when the stimulus is a pure vowel. The sustained response, on the other hand, reflects the periodic content of the vowel. The transient response is generated by any type of stimulus (speech and non-speech such clicks or tone bursts); however the sustained response is formed when periodic sound stimuli are used.

2.4.1 Transient Response

The transient response refers to the initial part of the SpEP (typically less than 20 ms) after the start of the stimulus and it reflects neural activiy of the ascending auditory pathway in response to the onset of the stimulus. The transient response usually contains five prominent peaks which signify activities of major nuclei along the ascending auditory pathway between the cochlear nerve and ICC in midbrain. Figure 2-7 illustrates those peaks, generated in response to a tone burst, and their corresponding points of origin along the auditory ascending pathway. The VA complex signifies auditory processing transition beyond the upper brainstem (Chandrasekaran and Kraus, 2010).

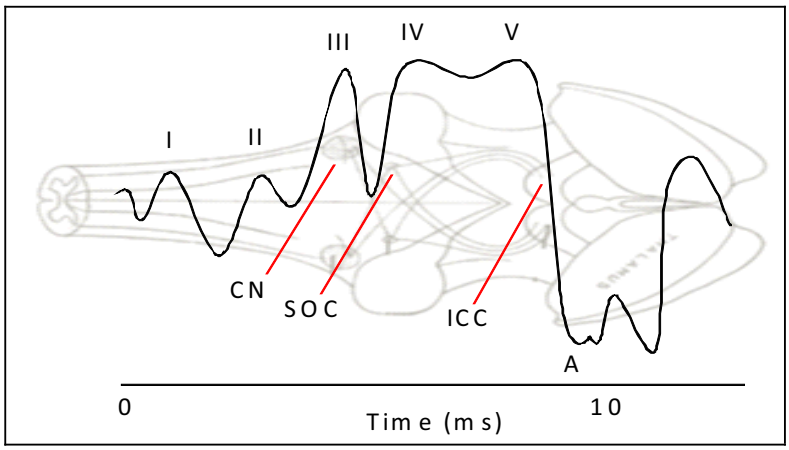

Figure 2-7: Five prominent peaks in transient response which correspond to neural activities of five key points along auditory pathways in response to a tone burst (based on: Adapted from the Wikimedia Commons file "Anatomy_of_the_Human_Ear.svg" http://upload.wikimedia.org/wikipedia/commons/d/d2/Anatomy_of_the_Human_Ear.svg; J.W. Hall. (1992). *ABR* [Online]. Available: http://www.biosemi.com/abr.htm).

Clinicians and researchers commonly use the transient response to clicks or short tone bursts for assessing the integrity of the auditory pathway (including as an objective hearing test for infants), and for studying people with learning problems and language impairments related to impairments of central auditory processing. These assessments are done by examining of the amplitude and latency of the five peaks and the VA complex. For instance, one study demonstrated that the VA duration is longer for people with language-based learning problems (Wible et al., 2005). Also, other studies have shown that adding noise to stimuli prolongs the VA duration (Song et al., 2006; Wible et al., 2005; Laroche, 2010). Although there have many studies on the behaviour of the transient response under different conditions, none of them used this it for vowel identification purposes and generally the transient response has been thought to reflect an undifferentiated response to the sudden onset of a sound stimulus. In other words, it has been thought that the transient response may not contain phonetic information (Skoe and Kraus, 2010). In

this study we attempt to show that the transient response contains useful information for vowel identification.

2.4.2 Sustained Response

The sustained response follows the transient response and it follows the acoustical features of the periodic speech stimulus including T0, F0, and formants. In the case of vowel stimuli, the sustained response provides information about the neural encoding of the vowel whereas the transient response has been thought to a response to the onset of the sound stimulus. The sustained response can be viewed as two separate responses: 1) the Envelope Following Response (EFR) and 2) the Frequency Following response (FFR). The EFR represents the neural response that follows the speech envelope, while the FFR represents the neural response that directly follows the harmonic content of speech, and in particular the speech formants.

The sustained response of SpEPs can correspond to the Envelope Following Response (EFR) or Frequency Following Response (FFR) depending on how the response signals are analyzed. Figure 2-8 illustrates a simplified model for generating the EFR and FFR. Figure 2-8-(a) (left-hand panel) shows the EFR and FFR generated using a 200 Hz stimulus tone (A), and Figure 2-8-(b) (right-hand panel) shows EFR and FFR generated using a 2 kHz tone modulated at 200 Hz (G). Stimuli are presented sequentially in alternating polarities (A,B and G,H) and their corresponding brainstem responses are shown in C,D and I,J. Note that the half-wave rectification is due to non-linear properties of hair cells. The EFR (E and K) is calculated by taking the average between the responses to the original stimulus and the inverted polarity stimulus, while the FFR (F and L) is calculated by taking the average between the response of the original stimulus and the negative of the response to the inverted polarity stimulus (Aiken and Picton, 2008). As can be seen

in K, the EFR follows the envelope of the modulated stimulus although as seen in E, it may also contain additional components with twice frequency of harmonics found in the stimulus. On the other hand, as shown in F, the FFR directly follows the harmonic content of the stimulus and as seen in L, it suppresses any response to the envelope of the stimulus.

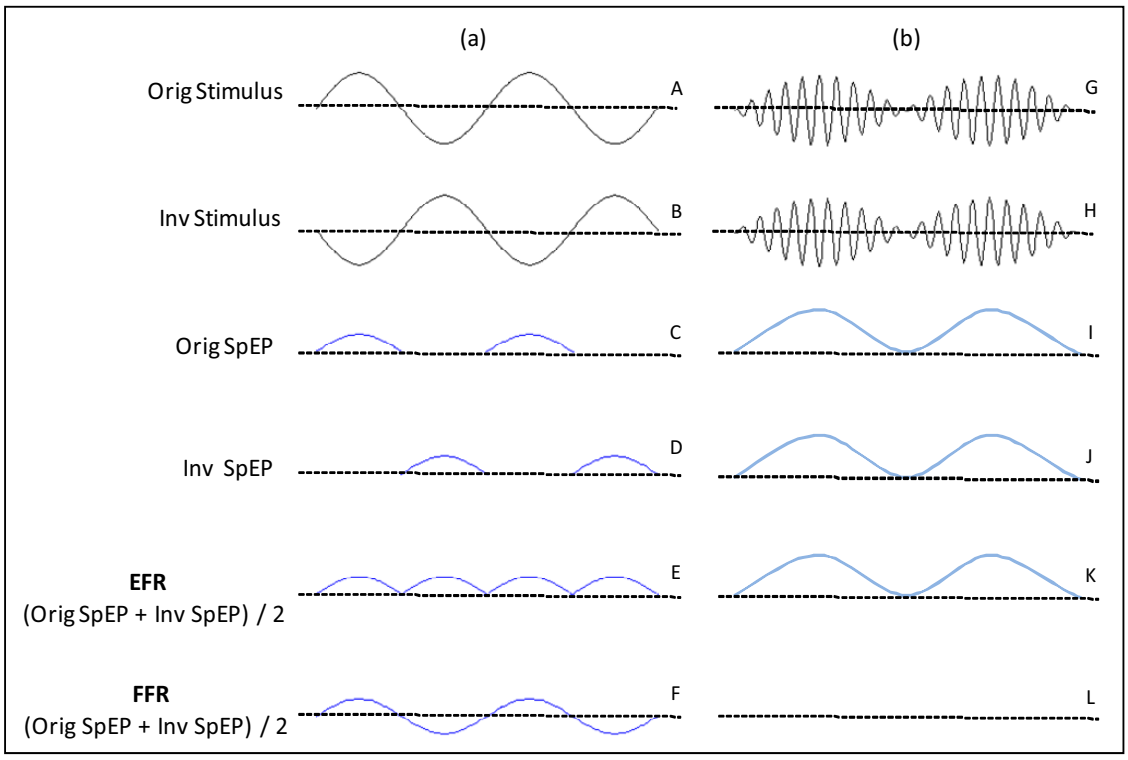

Figure 2-8: Simplified model for generating EFR and FFR. Panel (a) on the left shows the brainstem response to a 200 Hz tone and panel (b) on the right panel shows the brainstem response to a 2 kHz tone modulated at 200Hz (based on Aiken and Picton, 2008).

Consequently, the EFR primarily reflects auditory neural activity that is phase-locked to the envelopes of the speech stimuli, which are modulated at the fundamental frequency (F0=1/T0) as seen for example in Figure 2-3 (Aiken and Picton, 2008; Dajani et al., 2005). Although we would expect to see the EFR spectral content only at the modulation frequency (F0), the spectral contents are also seen at harmonics of F0 (i.e. multiple integer of F0) (Cebulla et al., 2006). These spectral contents are generated because of some non-linearities which are introduced by the

rectification process of the speech envelope within the cochlea and by non-linearities in neural processing (Aiken, 2008; Cebulla et al., 2006). This phenomenon can sometimes causes stronger EFR energy contents to appear at the harmonics of F0 than at F0. This case can be seen in Figure 2-9, where spectral amplitudes at multiple integer of F0 (F0=100Hz) are strong. Figure 2-9 illustrates the EFR spectral amplitude of vowel \i\ which is averaged across 48 trials collected from 8 subjects (more information about data collection can be found in section 3.3).

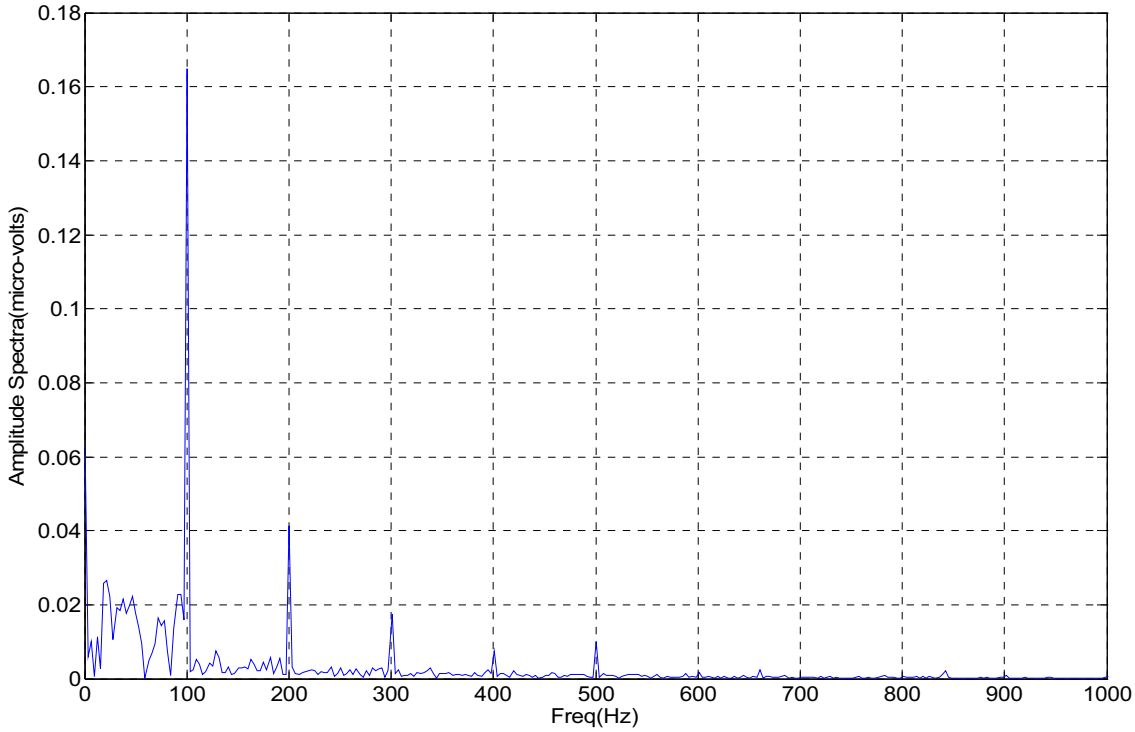

Figure 2-9: Amplitude spectrum of Envelope Following Response (EFR) for vowel \i\ averaged across all trials and subjects (i.e. grand-average EFR). This figure confirms that there are robust peaks at F0 (100Hz) and a few of its harmonics.

The FFR is formed as a result of auditory neural phase-locking that directly follows the harmonics of a speech stimulus, and in particular near the first formant F1 since these harmonics are typically the most intense in the stimulus and are usually well within the phase-locking frequency limit of auditory neurons. Spectral analysis of the FFR shows that strong peaks occur at harmonics of F0 near the formant frequencies (Krishnan, 2002; Skoe and Kraus, 2010), as shown in Figure 2-10 where the strong

peaks occurred at 2F0, 3F0, and 4F0 which are the harmonics near F1=270Hz. Figure 2-10 illustrates FFR spectral amplitude of vowel \i\ which is averaged across 48 trials collected from 8 subjects (more information about data collection can be found in section 3.3).

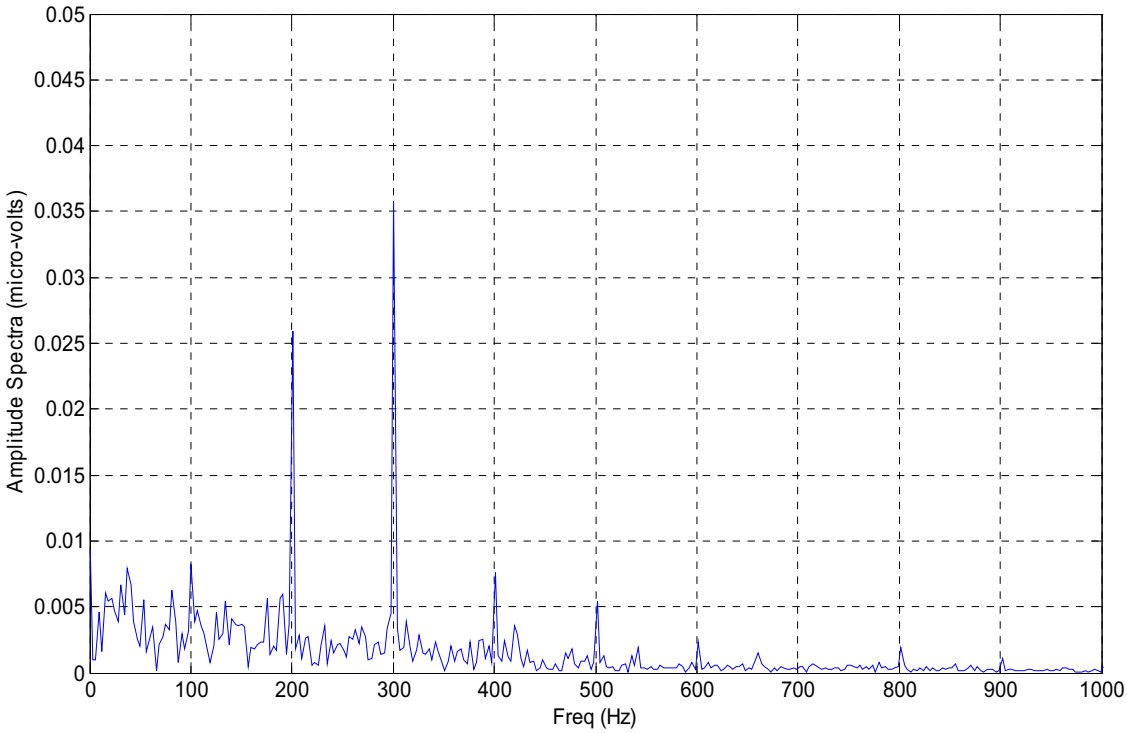

Figure 2-10: Amplitude spectrum of the Frequency Following Response (FFR) for vowel \i\ averaged across all trials and subjects (the grand-average FFR) with F1 = 270 Hz. The neural response to F1 can be observed from dominant peaks near F1 (i.e. 200, 300, and 400 Hz).

2.5 Related Work

Previous studies on SpEPs can be divided into two main groups: 1) studies which have tried to explain how the human auditory system processes speech by determining key signal features of SpEPs in both the time and spectral domains (Kraus and Nicole, 2005; Krishnan, 2002; Russo et al., 2004; Skoe and Kraus, 2010), 2) studies which have investigated the relationship between hearing-related disorders (e.g., hearing and language-learning impairments) and the key temporal and spectral features of SpEPs found in the first group of studies (Song et al., 2006; Wible et al., 2005). The results from the latter group of studies have been proposed for improving objective hearing assessments, especially in some children with learning impairments, and to provide better rehabilitation strategies. Since the focus of this work is closer to the first group of studies, the second group of studies will not be discussed further.

The first group of studies have identified some prominent spectral and temporal peaks on SpEPs as major landmarks for human processing of speech. These peaks were identified both in the transient and sustained responses. Figure 2-11 illustrates significant temporal peaks in an averaged SpEP (black) from 24 normal hearing subjects using a 40ms synthetic \da\ consonant-vowel stimulus (grey). Moreover, the stimulus has been shifted in time to account for the onset of the stimulus (i.e. auditory neural transmission delay). In the transient response, peaks I, III, V, A, C are identified as temporal features among which the first four peaks were explained in section 2.4.1 and peak C corresponds to the onset of the voicing part of the stimulus (Johnson et al., 2005). In the sustained response, peaks D, E, and F correspond to the periodic acoustical landmarks of the stimulus (i.e. responses at pitch periods of the stimulus (~10ms). Peak O corresponds to the end of the

stimulus. The sustained response was best explained using spectral features and in particular EFR and FFR which were discussed in section 2.4.2 (Aiken, 2008).

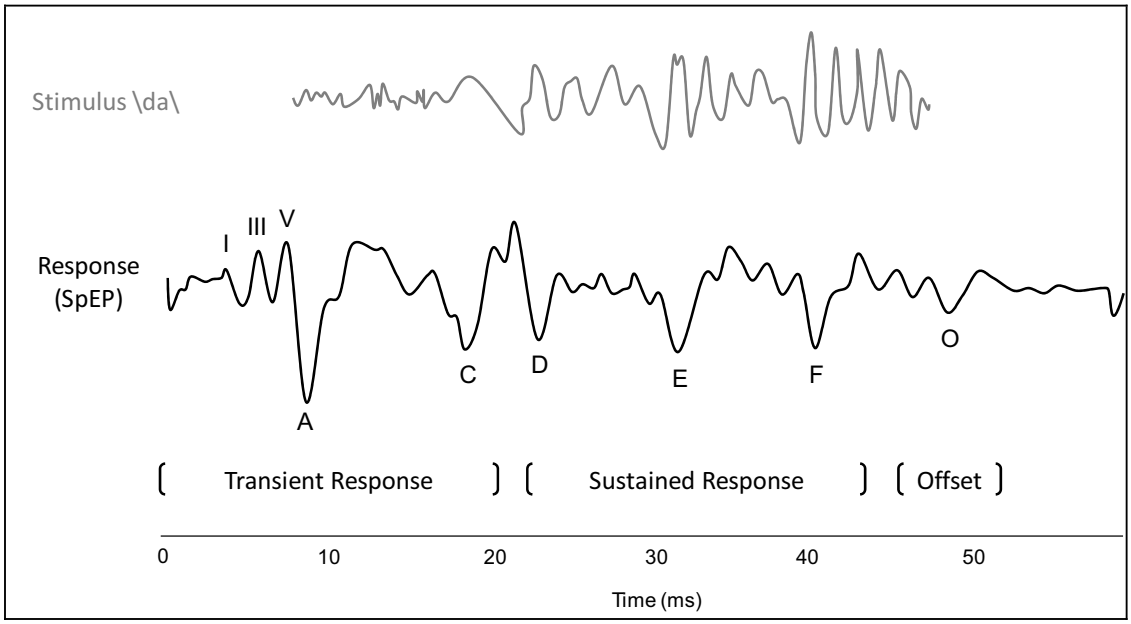

Figure 2-11: Time-domain illustration of an averaged brainstem response (black) to a 40ms synthetic \da\ stimulus (grey) from 24 normal hearing subjects. The stimulus has been shifted to account for the onset delay. The major peaks are shown in the transient and sustained response (based on Skoe and Kraus, 2010).

Some of the more recent studies have compared SpEP features of different speech sounds to better understand how the auditory system encodes temporal and spectral acoustical features of stimuli. For example, one study used consonant-vowel syllables starting with three stop consonants ([ba], [da], and [ga]) as stimuli to compare the latency of significant temporal peaks over the first 60ms of their SpEPs which were collected from 22 normal hearing children (Johnson, et al., 2008). The stimuli were generated synthetically such that they had the same pitch frequency, formant frequencies after the initial transition, and first formant transition; however, they had different transitions to the second and third formants. The transitions to F2 and F3 occupied frequency ranges in the following order from lowest to highest: [ba], [da], [ga]. The latency differences started as relatively large and they diminished over time until they vanished around 60ms, when all responses went to their steady state. The differences found in latency were in-line with the place coding

theory which states that higher frequencies get processed earlier because they are processed in the initial (basal) part of the basilar membrane (Johnson, et al., 2008). Another study considered the same set of stimuli but this time the only difference among stimuli was their second formant. They showed that SpEPs of the three stop consonants ([ga], [ba], and [da]), which were collected from 52 normal hearing children, can be distinguished by looking at SpEP phase changes over time (Skoe et al., 2011).

In this work we also studied SpEP features to help better understand how the human auditory system processes speech. However, there are a few differences between this work and previous work which are listed below,

1- *The purpose of this study was different:* We investigated the automatic classification of SpEPs. This was done using the sustained and transient response features separately. The classification result provides a quantitative measure for discriminating SpEPs using transient and sustained response features.

2- *The sustained response features were different:* In previous studies temporal features were used to investigate the differences between SpEPs of different speech sounds whereas in this study spectral features were used. We looked at the EFR and FFR response features to provide a better understanding about the phase-locking mechanism of the auditory neural system.

3- *The stimuli were different:* In most previous work, consonant-vowel syllables were used as stimuli to study SpEPs, whereas in this study we used five pure English vowels as stimuli. Generally, consonants are short and contain rich acoustic information mainly on the non-voiced portion of speech (low spectral energy). Vowels, on the other hand, contain voiced information of speech, therefore it can provide more comprehensive understanding about the encoding of speech (Laroche, 2010; Johnson et. al., 2008).

3 Methodology

3.1 Subjects

Eight subjects (six males and two females, 25-45 years old) participated in this experiment. The subjects had no known hearing problems, and normal hearing thresholds of 15 dB HL or less were confirmed in both ears through a pure-tone audiometric test using a Clinical Audiometer (model AC40, Interacoustics, Eden Prairie MN, USA) at 500, 1000, 2000, and 4000 Hz. Participants were asked to read and sign a consent form before starting the experiment and it was made sure that they understood the content of the consent form and the experimental procedures. This research was approved by the University of Ottawa Research Ethics Board.

3.2 Stimuli

In this study we use English vowels as stimuli because their SpEPs can be readily recorded and analyzed compared to words or sentences, while they provide rich information about speech processing in the central auditory system.

3.2.1 Methods for Generating Synthetic Speech

Synthetically-generated speech has applications in various fields such as text-to-speech and systems for people with speech impairments. There are two methods for generating speech synthetically, Concatenative and Formant synthesis. In the concatenative method, different segments (words or diaphones) of human-generated speech are recorded in a database and they get concatenated to one another as needed (Dutoit, 1997; Van Santen et al., 1997; Holmes and Holmes, 2001). The formant method, on the other hand, follows the source-filter model and it requires specification of some characteristics such as formants, bandwidths of resonators, and

pitch frequency as input. The latter technique provides more flexibility for setting speech parameters to produce a desired speech property (eg. male vs. female speech); however, it is more complex compared to the concatenative method. In this study we used a simplified version of Klatt's cascade/parallel formant synthesizer (Klatt, 1980), implemented by Laroche (2010) to generate the five English vowels (\a\,\e\,\ae\,\i\,\u\).

The parameters of the stimuli, namely the first 3 formant frequencies, formant bandwidths, and relative formant amplitudes, are shown in Table 3-1. These parameters followed those determined in previous work for male speakers (Klatt, 1980; Peterson and Barney, 1952). The fundamental frequency (F0) of all vowels was set to 100 Hz. Only the first three formants (F1, F2, and F3) of each vowel were included, since these formants are the most dominant. All the vowel stimuli are 300 ms in duration, and they were generated with a sampling frequency equal to 48 kHz.

Vowels	F1 (Hz)	F2 (Hz)	F3 (Hz)	BW1 (Hz)	BW2 (Hz)	BW3 (Hz)	A1 (dB)	A2 (dB)	A3 (dB)
\a\	700	1220	2600	130	70	160	-1	-5	-28
\ae\	660	1720	2410	70	150	320	-1	-12	-22
\e\	570	840	2410	100	60	110	0	-7	-34
\i\	270	2290	3010	50	100	140	-4	-24	-28
\u\	300	870	2240	65	110	140	-3	-19	-43

Table 3-1: Formant frequencies, bandwidths and amplitudes of the five synthetic vowels used as stimuli.

Figure 3-1 illustrates a simplified parallel/cascade vowel synthesizer that is used in this study to generate the five vowels. In this system, the source of all vowels is represented by train of impulses at the rate of F0 (100 Hz), and each formant is represented by a resonator acting as a narrow band-pass filter with center frequency and bandwidth equal to their corresponding values in Table 3-1. To obtain a synthetic vowel, the source impulses passed through three parallel resonators and

the resulting outputs were added together. The output signal, whose spectrum is shown on the furthest block on the right, was saved as a sound (.wav) file.

The transfer function of each formant (resonator) was calculated using the following equations,

$$T(f) = \frac{A}{1 - Bz^{-1} - Cz^{-2}}$$

With $z = \exp\left(\frac{j \times 2 \times \pi \times (\text{formant freq})}{Fs}\right)$

Where the constants of the transfer functions are:

$$C = -\exp\left(\frac{-2 \times \pi \times BW}{Fs}\right)$$

$$B = 2 \times \exp\left(\frac{-\pi \times BW}{Fs}\right) \times \cos\left(\frac{2 \times \pi \times (\text{formant freq})}{Fs}\right)$$

$$A = 1 - C - B$$

in which BW signifies the bandwidth for each formant, and Fs signifies the sampling frequency.

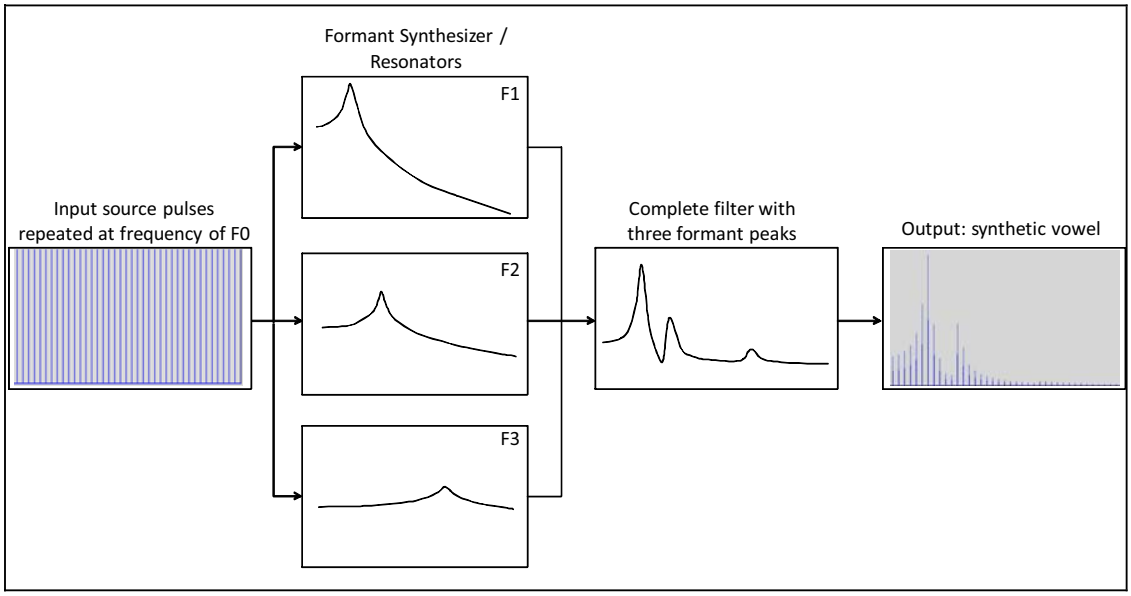

Figure 3-1: Simplified Parallel/Cascade vowels synthesizer.

Figure 3-2 shows single-sided amplitude spectra of the five synthetic English vowels which were generated using the formant synthesizer. Since we only analyze the brainstem response up to the frequencies in the region of the first formant (F1) in this study, this graph shows the amplitude spectra up to 1000 Hz to emphasize F1 of all the vowels.

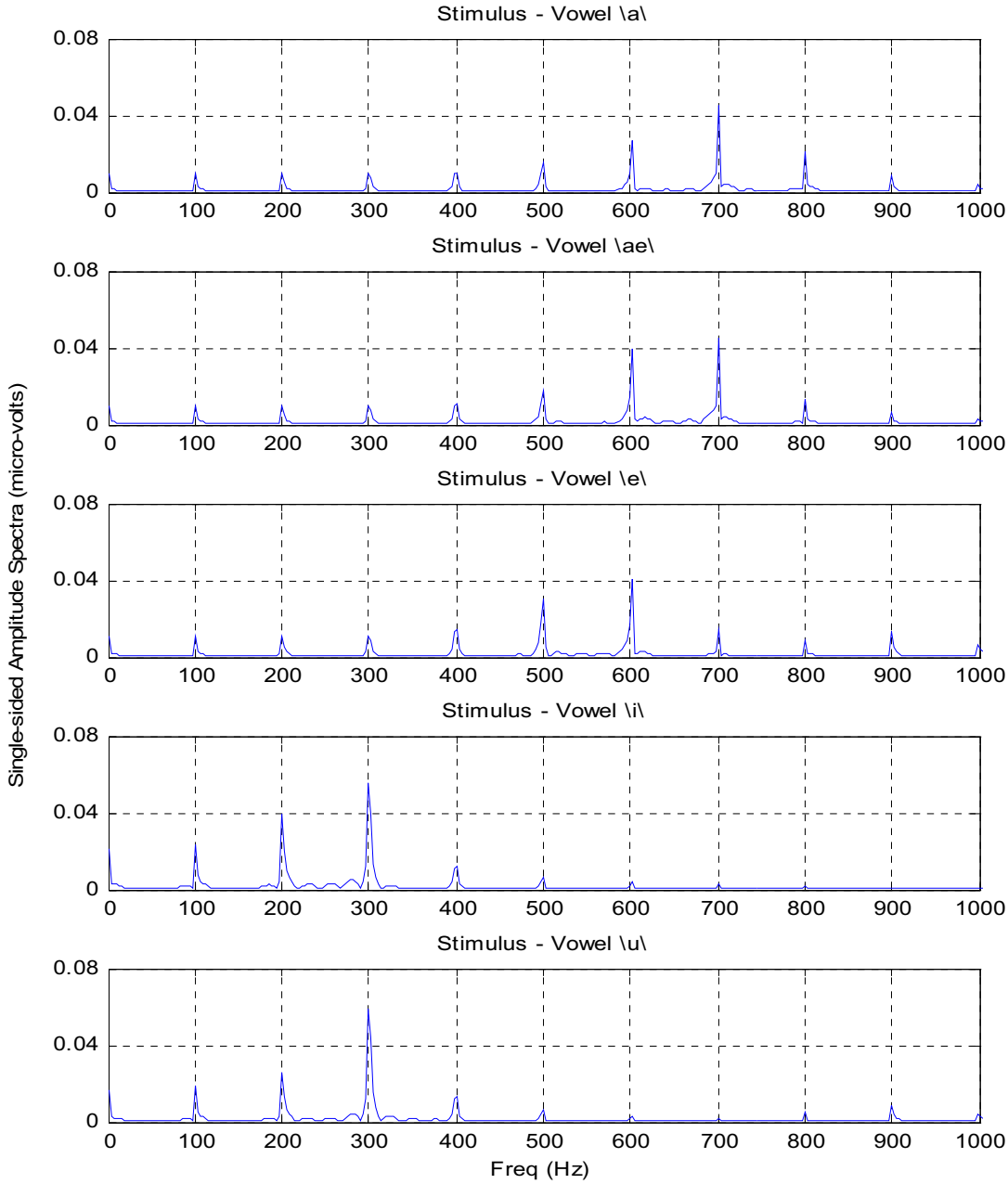

Figure 3-2: Single-sided amplitude spectra of the stimuli (five synthetic English vowels) up to 1000 Hz.

3.3 Experimental Protocol

Subjects were seated comfortably in an acoustical booth located at the University of Ottawa Health Campus, in the School of Rehabilitation Sciences (Audiology Program). During a recording session, the subjects were asked to stay relaxed, while minimizing their movements in order to minimize noise in the recordings (i.e. to avoid artefacts). They were also asked to keep their eyes open, and to help them stay awake a muted movie with subtitles was shown during the experiment. The reason for asking subjects to stay awake was to reduce aterfacts due to physical movements, which are discussed in the next subsection (3.4.2). Also, for the consistency of data collection it was easier to ensure that subjects stay awake than asking them to fall sleep for the duration of data recording.

A single recording session consisted of six trials, and in each trial, subjects were presented 500 repetitions of a single vowel at a repetition rate of 3.1/sec. Responses were coherently averaged over the 500 repetitions to give one EFR and one FFR prior to further analysis. A BioMARK v.7.0.2 system (Biological Marker of Auditory Processing, Bio-logic Systems Corporation) was used to present the stimuli and record the SpEPs. Each vowel was presented at a calibrated level of 80.5 dB SPL by adjusting an internal calibration factor in the BioMARK system, with the calibration performed by connecting the earphone to a 2cc coupler attached to a Brüel & Kjaer Artificial Ear type 4152, and a Sound Level Meter (SLM) Type 2230.

Figure 3-3 depicts the experimental set-up for collecting SpEPs. Stimuli were presented using Etymotic ER 2 insert earphones. Three gold-plated G.R.A.S. electrodes were used in this experiment; the recording electrode was placed at the vertex (Cz), the reference electrode was placed on right earlobe, and the ground electrode was placed on the left earlobe. Before placing the electrodes, the skin of the participants was scraped with cotton q-tips and Nuprep cream to lower the

impedance. Electrode impedances were kept below 5 kΩ during the recording by monitoring the impedance at the start and end of each trial, and discarding the trials with impedance higher than 5 kΩ. Vowels were presented in alternate polarities (i.e. 180 degree phase difference as shown in Figure 2-8) to both ears in order to allow the calculation of EFR and FFR which was discussed in section 2.4.2. Also, the vowels were played at 48 kHz with a 16-bit resolution. SpEPs were recorded with a sampling frequency of 3202 Hz for a duration of 319.8ms starting at stimulus onset.

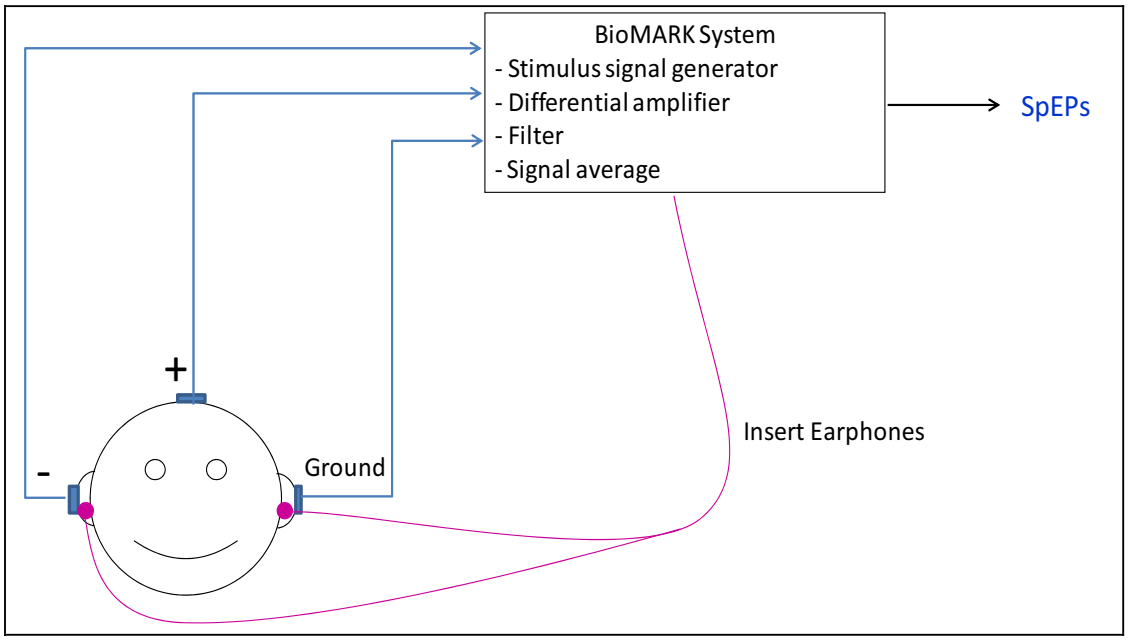

Figure 3-3: Schematic diagram of the experimental set-up.

3.4 Data Pre-processing

3.4.1 Coherent Averaging

Generally, the SpEP in response to a single stimulus presentation is a small signal that is buried in noise. The most commonly-used approach to reduce the noise of SpEPs (i.e. to increase SNR) is coherent averaging over many repetitions (typically >1000) (Chandrasekaran and Kraus, 2010). The underlying assumption is that SpEP samples are synchronized to the stimuli and are similar in shape over the repetitions, while noise is random and uncorrelated between the different samples(Deutsch and Micheli-Tzanakou, 1987). Therefore, averaging over many response samples suppresses the noise and so increases the SNR of the SpEP samples. The coherent averaging formula is shown below, where $s_i(t)$ represents a single SpEP sample and n is the number of repetitions. In this study we chose n to be 500 repetitions because it gave us desired robustness without having to select a higher number of repetitions.

$$\text{Coherent Average} = \frac{\sum_{i=1}^{n} s_i(t)}{n}$$

3.4.2 Artefact Reduction

As part of data pre-processing, it is important to eliminate or at least reduce artefacts in order to obtain cleaner and more robust SpEP samples. The artefacts can be categorized into three types based their sources, 1) physical movements, 2) electromagnetic (EM), and 3) Cochlear Microphonic (CM). As the name suggest, artefacts from physical movements can be caused by any body movements such as tightening of the jaw or the neck muscles. The EM arifacts occur because of surrounding ambient electromagnetic noise that can be due to nearby equipment or

machinery and power line noise, and potentially electromagnetic leakage from the stimulus generating equipment to the recording electrodes. Lastly, the CM artefacts originate from the cochlear hair cells and, similar to the FFR, the CM follows temporal shape of the stimulus (Skoe and Kraus, 2010).

We used a few techniques to eliminate/reduce the aforementioned artefacts. The first technique was averaging responses of alternating stimulus polarities when estimating the Envelope Following Response (EFR). This technique eliminates CM artefacts, because CMs from opposite polarities cancel out each other (Skoe and Kraus, 2010). The second technique was using foam insert earphones and plastic tubes for connecting earphones to the BioMark stimulus transducer. This reduces the risk of EM leakage from the stimulus generator equipment to the recording electrodes (Akhoun et al., 2008; Skoe and Kraus, 2010). The third technique was presenting the stimulus at a rate that does not contain multiple integers of powerline noise cycles to ensure the 60Hz artefact shifts in phase from one response to the other. The artefact was then suppressed after taking coherent averaging over repetitions of SpEP samples (Picton and John, 2004). As it was mentioned in Section 3.3, we presented the stimulus at the rate of 3.1 Hz or in 322.6 ms intervals, which is not an integer multiple of the powerline peridiodicity. The fourth technique was decreasing body movements by providing a comfortable seat and asking subjects to stay relaxed and still during data collection. In the event of the occurrence of muscular and other types of artefacts, epochs in which the response exceeded 23.8 μV were discarded by the BioMark system. Moreover, care was taken to repeat trials which contained more than 20 discarded epoches. Finally, we tried to reduce the electromagnetic artefact increasing the distance separating electrical equipment, such as DVD player, and the subject.

3.5 Classifying SpEPs of Five English Vowels

3.5.1 Feature Selection

We selected two different sets of features for classification of the SpEPs. The first set of features was obtained from the frequency domain representation of the sustained response and the second set of features was obtained from the time domain representation of the transient response.

3.5.1.1 Sustained Feature Selection

For the first set of features, we examined both amplitude and phase of EFR and FFR spectra at F0 and its harmonics. The reason for selecting these frequencies is that we expected robust responses at harmonics of F0 due to the neural phase-locking (discussed in section 2.4.2 Sustained Response). Also, by choosing discrete frequency points, we wanted to keep the feature selection approach straightforward in order to avoid working with complex transformations that are outside of the scope of this study.

We further reduced the number of frequency points by considering the spectral values (i.e. amplitudes and phases) between 100 Hz and 700 Hz, inclusive, for EFR, and between 200 Hz and 800 Hz, inclusive, for FFR. Therefore, the sustained feature vectors had 7 amplitude and 7 phase feature elements for the EFR or FFR. The reason why we selected frequencies below 1000 Hz is that neural phase-locking degrades above 1000 Hz and also the upper cut-off frequency of the band-pass filter on the BioMark system is 1000Hz. Moreover, by looking at the grand average amplitude spectra of the EFR and FFR (Figure 3-4 and Figure 3-5) we realized that the peaks at harmonics start diminishing after 700 Hz for EFR and after 800 Hz for FFR. Also, the spectral component of the FFRs at 100 Hz was excluded because there were no robust peaks at this frequency.

We assessed amplitude and phase features by comparing the classification results for different combinations of the features and found that using the phase features do not improve the classification accuracy by much (only 2-3% improvement). Thus, we decided to only consider the amplitude features to avoid the so-called "*curse of dimensionality"* which could increase the risk of overfitting and classification complexity (Duda et al., 2001). This means the sustained feature vectors was reduced to 7 amplitude feature elements for the EFR or FFR.

The frequency spectrum was determined using the Discrete Fourier Transform (DFT) of the coherently averaged response in each trial, containing 1024 data points, and the amplitude spectra was calculated using the following formula,

Amp Spectrum = | FFT (EFR or FFR signal in time domain, NFFT) / signal length |

Single-sided Amp Spectrum = 2 x first or second half of Amp Spectrum

where FFT stands for Fast Fourier Transform and NFFT = 1024.

Figure 3-4 shows the amplitude spectra of the EFR for each vowel averaged across all trials and subjects (i.e. grand-average EFRs). As can be seen, there are robust peaks at harmonics of F0. Figure 3-5 illustrates the amplitude spectra of the FFR for each vowel averaged across all trials and subjects (the grand-average FFRs). This figure confirms that peaks at harmonics of F0 are dominant near the F1 frequencies listed in Table 3-1. In this work, the responses to the second and third formants probably play little role in the analysis, because several of them were beyond the upper cut-off frequency of the band-pass filter on the BioMARK system of 1000 Hz, and beyond the phase-locking limit of the probable main generator of SpEPs in the upper brainstem (Johnson et al., 2005).

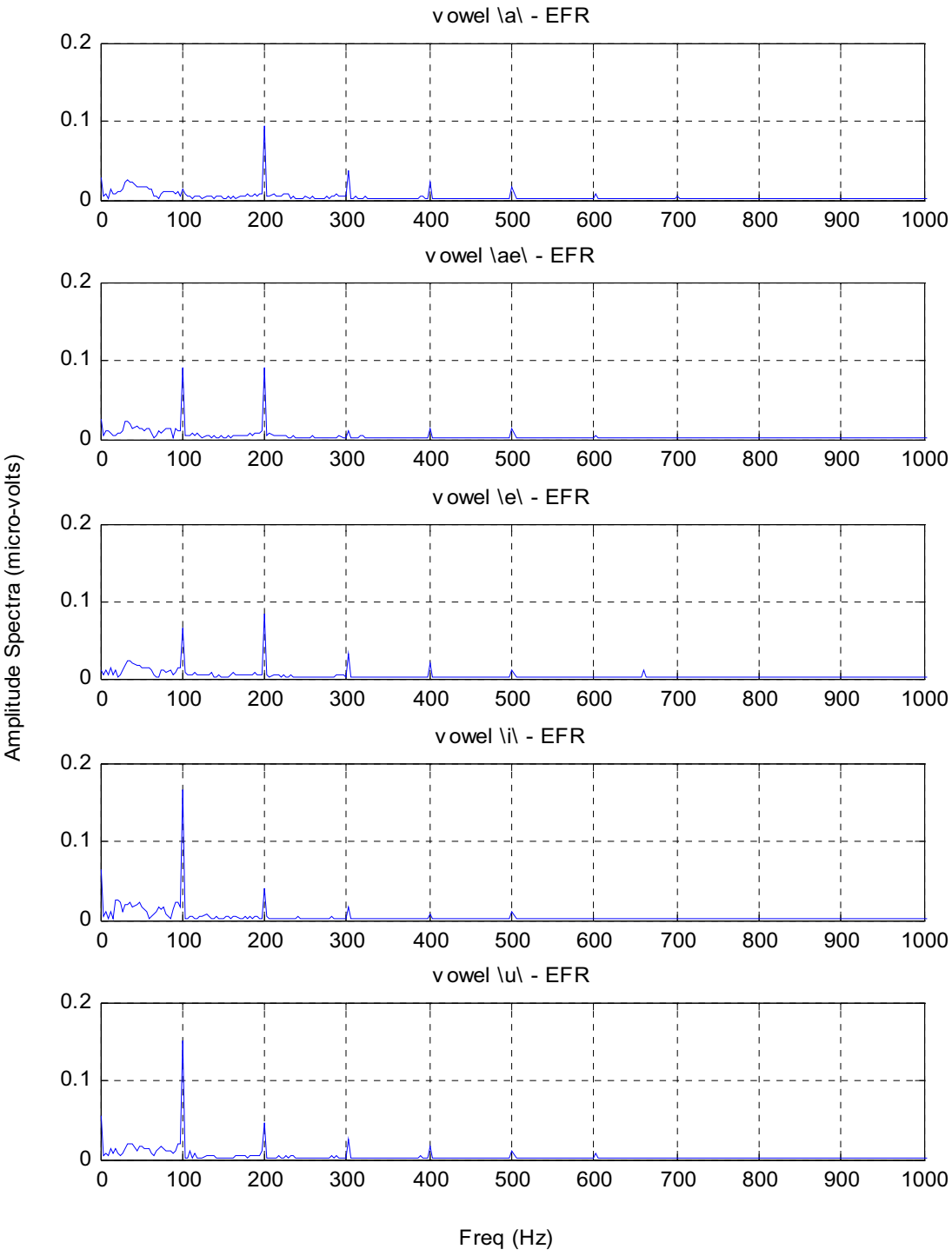

Figure 3-4: Single-sided amplitude spectra (up to 1000 Hz) of the SpEPs for all vowels averaged over all trials and all subjects (grand-averages) for the Envelope Following Response (EFR).

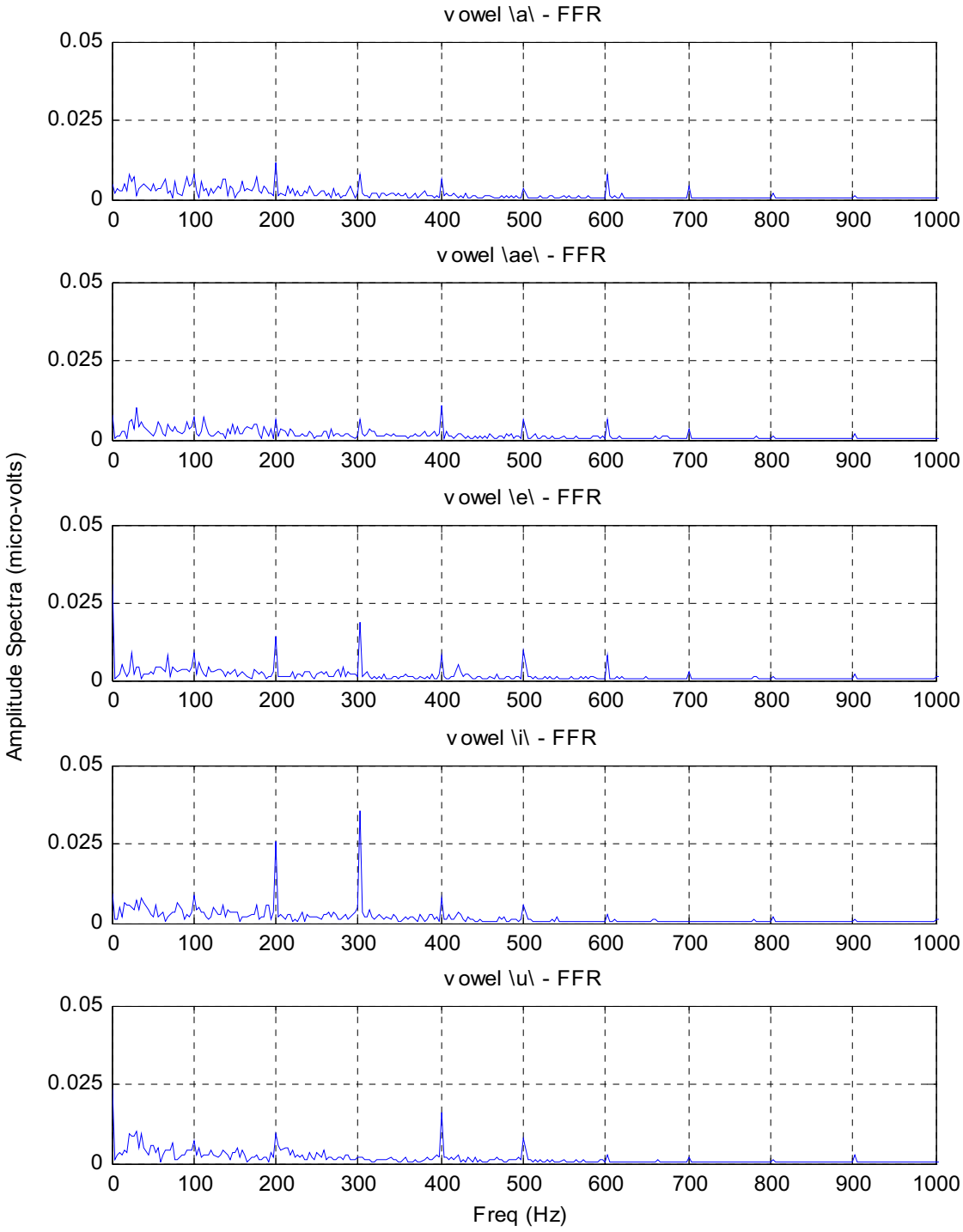

Figure 3-5: Single-sided amplitude spectra (up to 1000 Hz) of the SpEPs for all vowels averaged over all trials and all subjects (grand-averages) for the Frequency Following Response (FFR).

3.5.1.2 Transient Feature Selection

For the second set of features, we focused on the significant transient peaks from the EFRs (i.e. EFR from all trials) in the time domain. As was discussed in Section 2.4.1 on the Transient Response, the significant transient peaks correspond to neural responses of major landmarks along the ascending auditory pathways and they are known as peak I, II, III, IV, V, and A. Of these, peaks V and A (*a.k.a.* the VA complex) have the most clinical use in speech ABR. The reason for choosing the transient responses only from the EFR, rather than the FFR or both EFR and FFR, is that the EFR does not contain CMs and so it provides a clean transient responses, whereas this is not the case for the FFR (Aiken, 2008; Russo *et al.,* 2004).

We identified all the peaks using a Matlab program that we implemented by finding local maxima/minima around the time when we expected the peaks to occur. After determining the peaks, we realized that V and A were the most robust peaks relative to the background noise across all trials, whereas other peaks were not clearly present all the time. Hence, we only assessed features of the VA complex, namely amplitude and latency of V and A, VA duration, VA height, and VA slope (Russo *et al.,* 2004). The classification accuracy obtained with all possible non-dependent combinations of the 7 features (i.e. highly linearly correlated features were not combined to avoid singularity in training data) showed that the 4 features of latency of V and A, and height of V and A provided the highest classification accuracy. As such, we present the classification results with transient feature vectors containing these 4 elements.

3.5.2 Classification Method: Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) was employed for classification using Matlab (v. 7.9.0.529, The Mathworks, Natick, MA, U.S.) (Duda et al., 2001). We had five classes corresponding to the five different vowels and each class had 48 sets of

SpEPs trials (6 samples each corresponding to 500 stimulus repetitions per subject × 8 subjects). Leave-one-out with no replacement was used to train and test. That is, training was performed on all samples except one, which was used to test. The leave-one-out was repeated such that each of the 240 SpEP samples (5 vowels x 6 trials/vowel x 8 subjects = 240) was tested. We used leave-one-out, because our data set is small and this approach provides more accurate training (i.e. reduces error rates) especially for small data sets (Duda et al., 2001).

Generally, LDA separates classes from one to another by drawing hyper-planes between classes such that the ratio of between-class variance over within-class variance is maximized in order to optimize class separation (Duda et al., 2001a). One of the biggest advantages of LDA over more complex classifiers like artificial neural networks is that it prevents over-fitting especially when the data set is small.

4 Results

We evaluated the classification results using a *classification accuracy* measure which is the aggregate of correctly classified samples over the total number of samples (Duda et al., 2001). It is important to note that the chance level accuracy is 20% since we have 5 classes (100% / 5 vowels = 20%). The next two sections present the results using the features from the sustained response (section 4.1) and transient response (section 4.2). Finally, section 4.3 presents the classification result for combination of the sustained and transient response features.

4.1 Classification of SpEPs using Sustained Response Features

Table 4-1 shows LDA classification accuracies per subject for three different sets of amplitude features including, 1) EFR (7 elements), 2) FFR (7 elements), and 3) EFR+FFR (combining both EFR and FFR features to generate a new feature set with 14 elements). The last row of this table illustrates the overall classification accuracy (i.e. average accuracy over 8 subjects) and as can be seen EFR+FFR amplitude features provided the highest accuracy of 80.83% followed by the individual EFR and FFR amplitude features.

Since the dimensions of the three feature sets are relatively large (i.e. dimension of size 7 and 14) we also performed linear Support Vector Machine (SVM) classification to confirm the LDA classification results and to see if the classification accuracies improve. Similar to LDA, SVM draws hyper-planes between classes; however, SVM generates a hyper-plane between two classes by considering nearest points of the two classes and not the whole training set (Duda et al., 2001). Table A-1 in Appendix-B illustrates the overall classification accuracies for the SVM classification. Since the classification accuracies achieved from the SVM approach are

similar and in some cases a bit lower than those obtained from the LDA approach, we will focus on the LDA results. Other SVM results are also provided in this Appendix and are discussed below wherever it is applicable.

	Classification Accuracy		
Subjects	**EFR + FFR**	**EFR**	**FFR**
Sub1	70.00%	60.00%	56.67%
Sub2	80.00%	66.67%	53.33%
Sub3	83.33%	76.67%	53.33%
Sub4	53.33%	70.00%	26.67%
Sub5	80.00%	53.33%	63.33%
Sub6	100.00%	76.67%	80.00%
Sub7	93.33%	70.00%	66.67%
Sub8	86.67%	83.33%	36.67%
All Subs	**80.83%**	**69.58%**	**54.58%**

Table 4-1: LDA classification accuracies of three different amplitude feature sets.

In order to visualize the distribution of data samples, we plotted EFR and FFR amplitudes from the subject who provided the highest accuracy rate for each EFR and FFR feature sets (i.e. subjects 8 and 6 who provided the highest accuracies for EFR and FFR feature sets respectively). Figure 4-1 shows log-scaled EFR amplitude distribution for three trials from subject 8. This figure presents 3 amplitude values for each of 7 frequency points (i.e. harmonics of F0) which were discussed in section 3.5.1.1. We chose to show the amplitude spectra from only the first 3 trials, instead of 6 trials, to provide a clearer and less cluttered plot. Similarly, Figure 4-2 shows log-scaled FFR distribution for subject 6. This way we can visualize 7 features of each EFR and FFR, which were discussed in section 3.5.1.1, in 2-dimensional plot; however, we cannot clearly see how each class is distributed with respect to other classes. To see that we performed Uncorrelated Linear Discriminant Analaysis (ULDA), using Matlab (v. 7.9.0.529, The MathWorks, Natick, MA, U.S.A), on all trials

for each of the three feature sets listed in Table 4-1 and plotted the first two features. Figure 4-3, Figure 4-4, and Figure 4-5 illustrate these plots for EFR+FFR, EFR, and FFR feature sets respectively. ULDA is an effective feature reduction method which generates up to M-1 statistically uncorrelated features (i.e. minimizes redundancy) for a problem that has M classes (Jin et al., 2001a; Jin et al., 2001b).

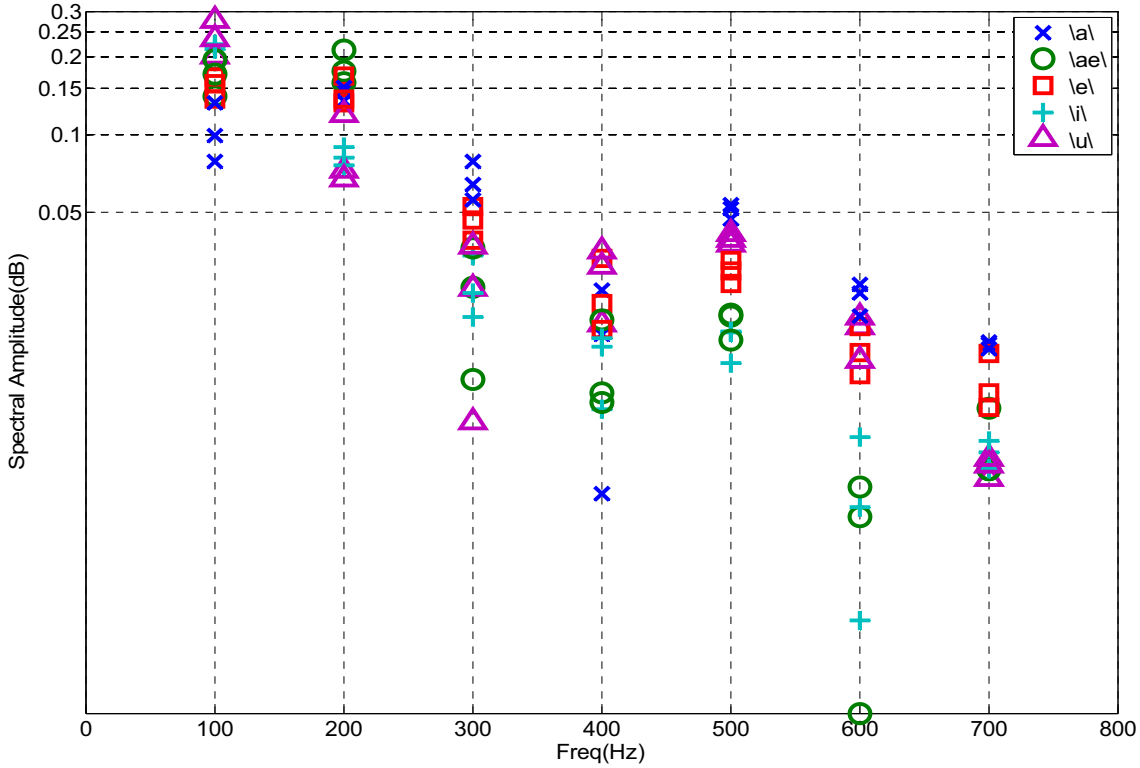

Figure 4-1: Log-scaled EFR amplitude features for the first 3 trials from Subject 8.

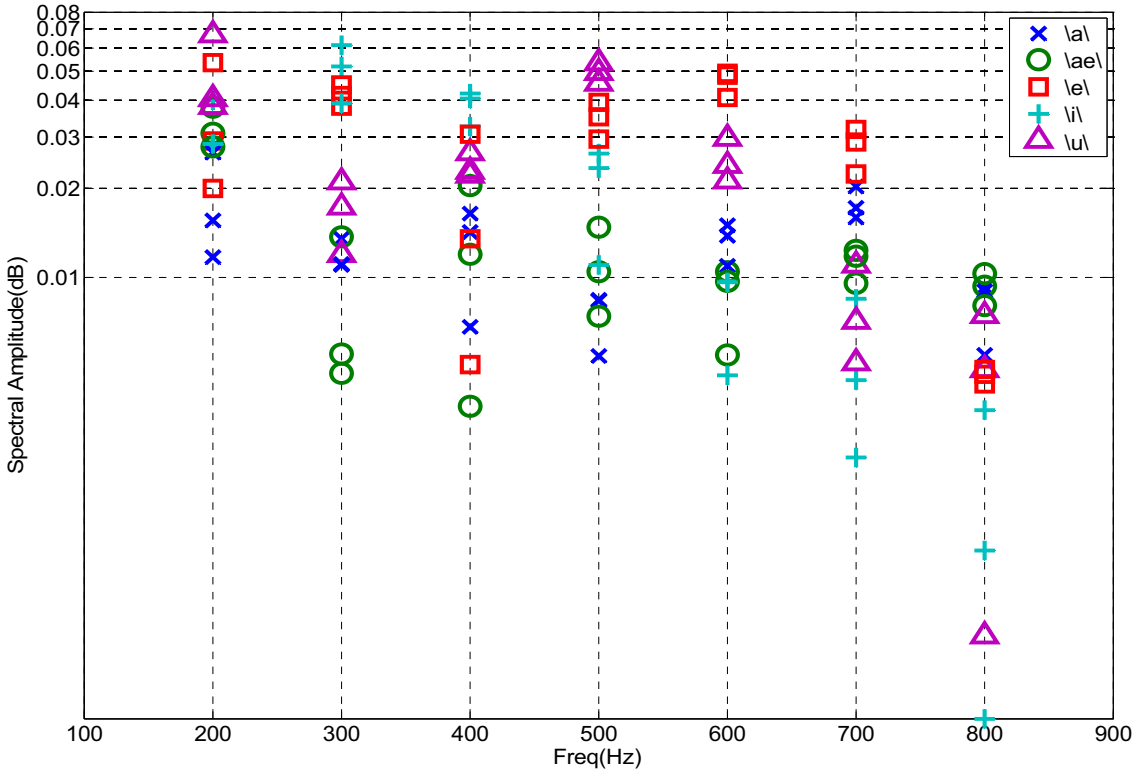

Figure 4-2: Log-scaled FFR amplitude features for the first 3 trials from Subject 6.

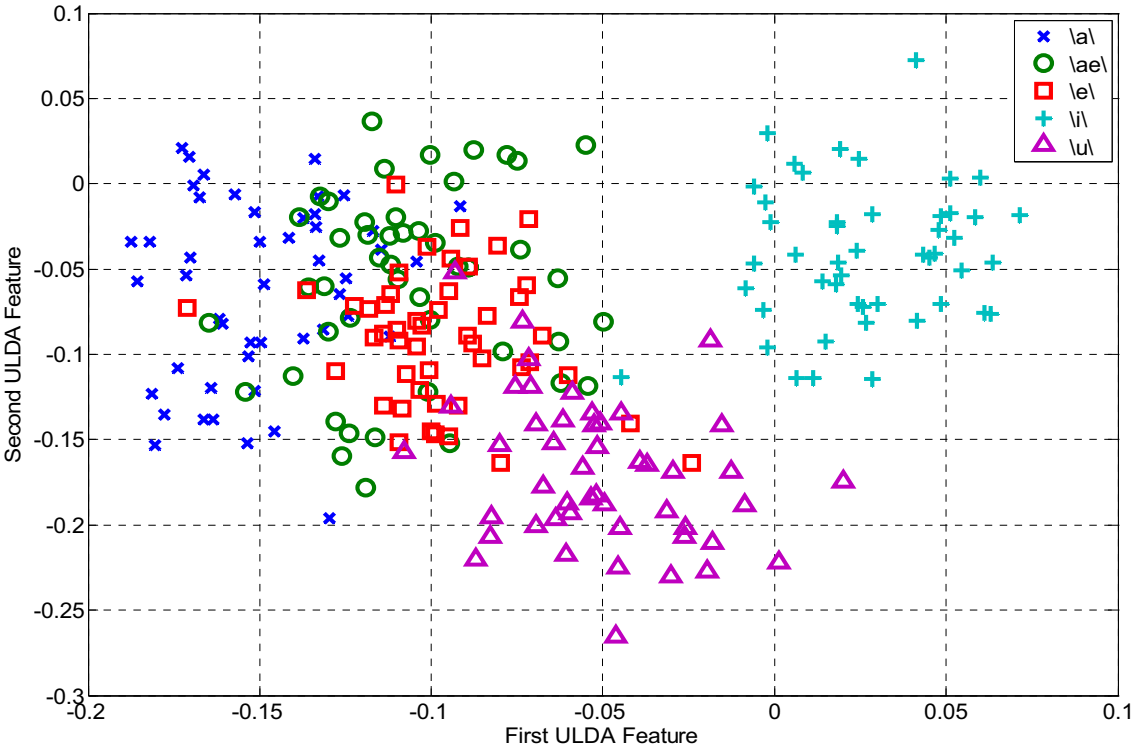

Figure 4-3: Feature1 versus feature2 of the ULDA analysis on the original EFR+ FFR features for all trials.

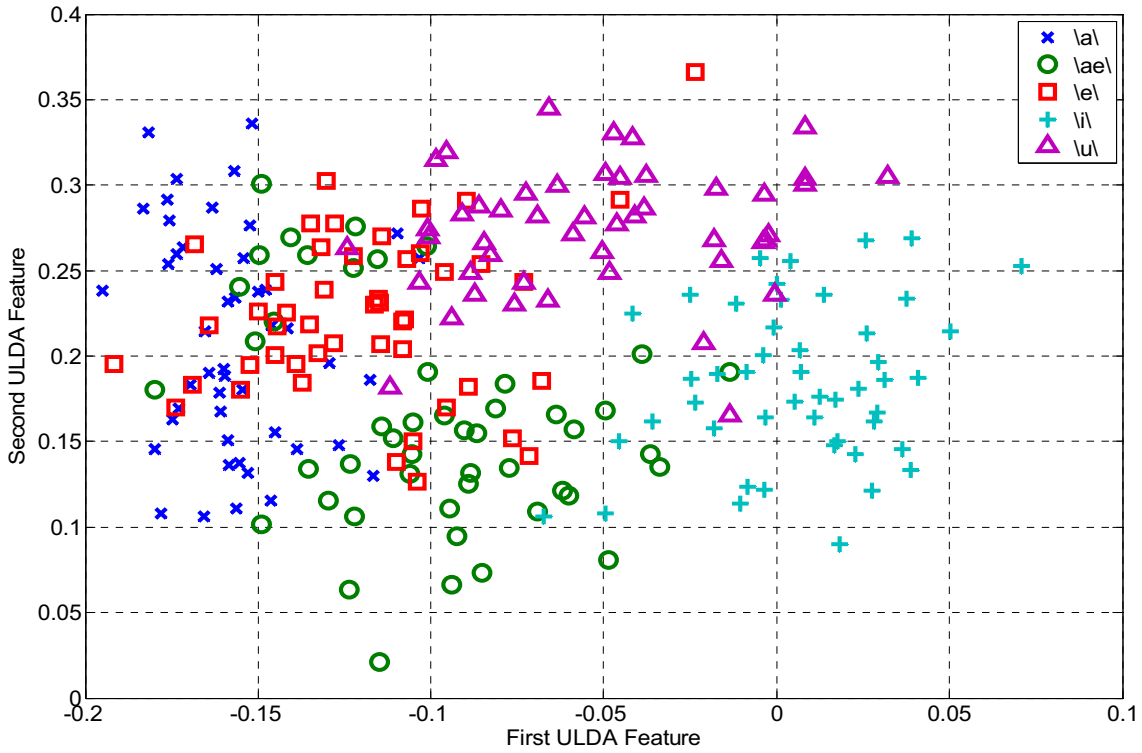

Figure 4-4: Feature 1 versus feature 2 of the ULDA analysis on the original EFR feature set for all trials.

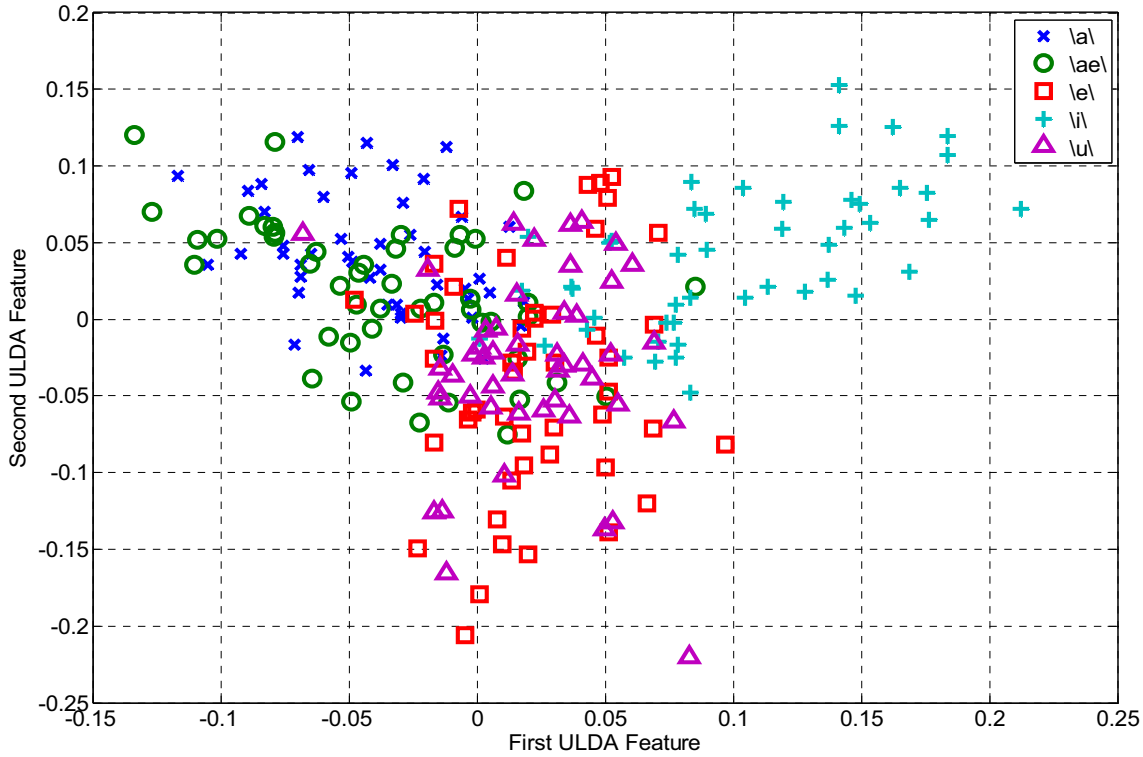

Figure 4-5: Feature 1 versus feature 2 of the ULDA analysis on the original FFR feature set for all trials.

Table 4-2 provides more details on the classification results by showing confusion matrices and Mahalanobis distances for all trials over the three feature sets (also, Table A-2 in Appendix-B provides confusion matrices for the SVM approach). The left column of Table 4-2-((a), (b), and (c)) illustrates confusion matrices for the three different amplitude feature sets. Each value in the confusion matrices represents the number of samples, from one class, which got classified to a particular class. For example, the value in row 1 column 2 (1,2) of Table 4-2-(a) indicates that 3 of 48 SpEP samples from the vowel \a\ was being heard as the vowel \ae\ (i.e. 4 samples of vowel \a\ were misclassified to vowel \ae\). Another example is the value in cell (3,3) of Table 4-2-(a) which indicates that 36 of 48 SpEP samples of the vowel \e\ was being heard as the vowel \e\ (i.e. 36 samples of vowel \e\ were correctly classified). Table4-2-(a) shows that vowel \i\ and \ae\ have provided the highest and lowest accuracies of 97.9% (47 correctly classified samples/48 samples in each class) and 66.66% (32/48), respectively. Also, the lowest classification accuracy (22.9%) among the three confusion matrices belongs to vowel \ae\ in the case of using FFR amplitude features (Table 4-2-(c)).

The right column of Table 4-2-((d), (e), and (f)) illustrates the Mahalanobis distances between all possible pair-wise combinations of the five vowels for the three different amplitude feature sets (Note that all Mahalanobis distances are normalized with Mahalanobis distance of each vowel from itself since the distance was the same for all vowels). Each value in "Mahal Dist" matrices represents the averaged Mahalanobis distances from every SpEP sample of one class in "Sample Vowels" to a reference class in "Reference Vowels". For example, the value in cell (2,1) of Table 4-2-(d) shows the averaged Mahalanobis distance of 48 vowel \ae\ SpEP samples from the reference class vowel \a\. The smallest Mahalanobis distance in all three matrices is 1, which corresponds to cases where the reference and the sample vowels are the same (i.e. the diagonal cells of the matrices) and the largest distance

(13.45) belongs to the distance between the sample vowel \i\ and the reference vowel \a\ in the EFR Mahalanobis Distance matrix (Table 4-2- (e)). In order to simplify the comparison between confusion matrices and Mahalanobis distances, five different shades of grey are used that signify five different ranges for classification distribution and Mahalanobis distances. In general, the darker the grey, the higher the classification rate and the longer is the distance.

Conf Matrix EFR + FFR		Predicted Vowels				
		\a\	\ae\	\e\	\i\	\u\
Actual Vowels	\a\	41	3	4	0	0
	\ae\	7	32	6	0	3
	\e\	6	4	36	0	2
	\i\	0	0	0	47	1
	\u\	0	1	7	2	38

(a)

Mahal Dist EFR + FFR		Sample Vowels				
		\a\	\ae\	\e\	\i\	\u\
Reference Vowels	\a\	1.00	4.04	5.19	12.29	8.55
	\ae\	3.49	1.00	4.63	6.19	3.73
	\e\	4.50	4.26	1.00	6.43	4.13
	\i\	12.12	8.37	10.09	1.00	6.55
	\u\	5.23	4.59	4.25	5.88	1.00

(d)

Conf Matrix EFR		Predicted Vowels				
		\a\	\ae\	\e\	\i\	\u\
Actual Vowels	\a\	34	5	9	0	0
	\ae\	5	32	6	4	1
	\e\	12	11	22	0	3
	\i\	0	1	0	46	1
	\u\	0	0	9	6	33

(b)

Mahal Dist EFR		Sample Vowels				
		\a\	\ae\	\e\	\i\	\u\
Reference Vowels	\a\	1.00	3.78	2.48	13.45	8.31
	\ae\	3.64	1.00	2.64	3.99	3.80
	\e\	1.77	1.88	1.00	3.76	2.71
	\i\	12.66	8.12	8.96	1.00	6.40
	\u\	3.73	3.37	2.22	2.81	1.00

(e)

Conf Matrix FFR		Predicted Vowels				
		\a\	\ae\	\e\	\i\	\u\
Actual Vowels	\a\	22	17	1	0	8
	\ae\	23	11	5	1	8
	\e\	1	6	24	7	10
	\i\	1	0	2	37	8
	\u\	2	1	4	4	37

(c)

Mahal Dist FFR		Sample Vowels				
		\a\	\ae\	\e\	\i\	\u\
Reference Vowels	\a\	1.00	1.53	5.24	6.21	3.11
	\ae\	1.73	1.00	4.09	4.53	2.10
	\e\	2.98	2.80	1.00	2.48	1.56
	\i\	4.91	4.19	5.65	1.00	2.71
	\u\	4.03	2.96	4.44	4.17	1.00

(f)

Classification Accuracy
>40% (20-48)
20%-40% (10-19)
10%-20% (5-9)
0.02%-10% (1-4)
0% (0)

Mahalanobis Distamce
1
1-2
2-5
5-8
>8

Table 4-2: Left column shows confusion matrices for a) EFR+FFR, b) EFR, and c) FFR. The darker the grey, the higher the classification rate is and vice versa. Right column shows Mahalanobis distances between all possible pair-wise combinations of vowels for d) EFR+FFR, e) EFR, f) FFR. The darker the grey, the smaller the Mahalanobis distance is and vice versa.

4.2 Classification of SpEPs Using Transient Response Features

We performed LDA classification on all possible combinations of the seven transient features including latency and height of V and A waves, VA complex duration, VA height, and VA slope. We compared the corresponding classification accuracies and found that a feature set containing the combination of the four features derived from the latency and height of V and A waves provides the highest overall accuracy of 38.33% (SVM provided a classification accuracy of 34.58%). Table 4-3 shows classification accuracies per subject for the set of these four transient features.

To visualize the distribution of the five classes with respect to each other for the aforementioned feature set, we performed ULDA on all trials and plotted the first two features (this is shown in Figure 4-6).

Subjects	Classification Accuracy for V,A latency + V,A height
Sub1	33.33%
Sub2	33.33%
Sub3	46.67%
Sub4	26.67%
Sub5	50.00%
Sub6	36.67%
Sub7	30.00%
Sub8	50.00%
All Subs	**38.33%**

Table 4-3: LDA classification accuracies per subject using 4 transient features (latency and height of V and A).

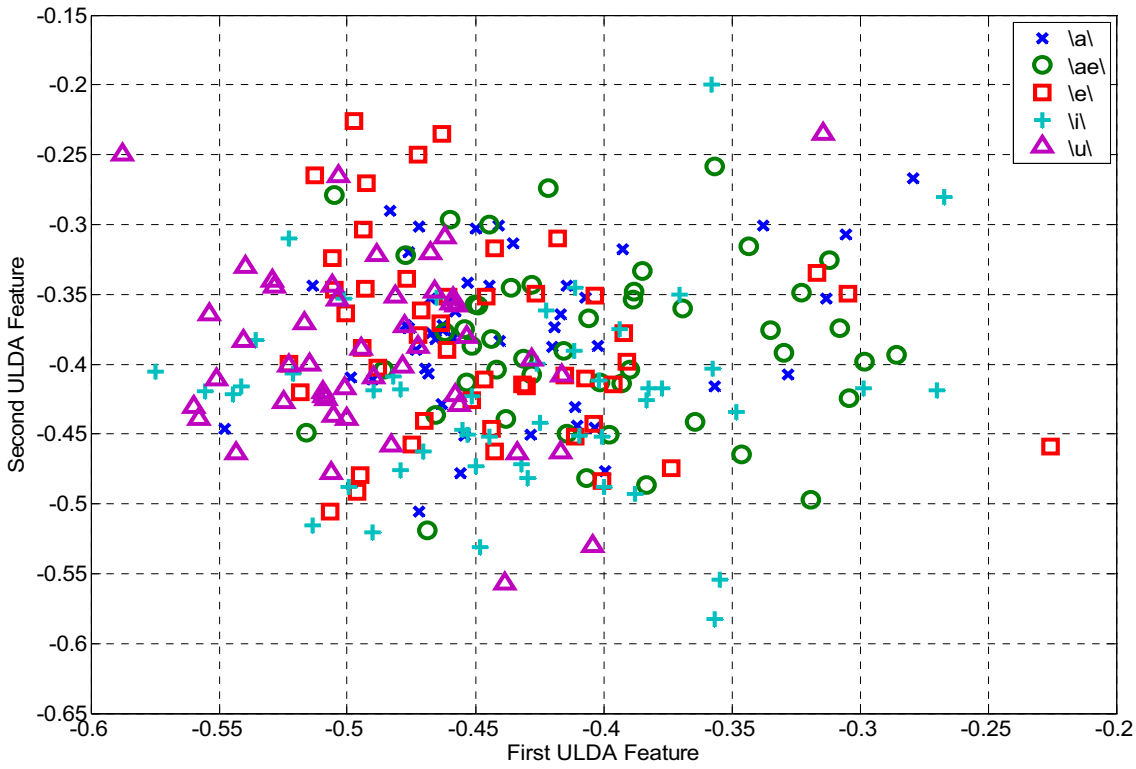

Figure 4-6: Feature 1 versus feature 2 of the ULDA analysis on the 4 transient features (latency and height of V and A) for all trials.

Table 4-4 provides more details about the classification results for each class by illustrating confusion matrices, on the left side (a) and (b), and Mahalanobis distances, on the right side (c), for the four selected transient features (also, Table A-3 in Appendix-B provides a confusion matrix for the SVM approach). As can be seen in confusion matrix (a), vowel \ae\ and \u\ have provided the highest and lowest accuracies of 62.5% (30/48) and 14.58% (7/48) respectively. Since vowel \a\ provides the lowest classification accuracy among all vowels, we re-examined the trials without vowel \a\ samples to see if the classification accuracies improve. The result of this analysis is shown in the confusion matrix (c).

Table 4-4-(c) shows Mahalanobis distances between all possible pair-wise combinations of the five vowels for the 4 transient features (Note that all Mahalanobis distances are normalized with the Mahalanobis distance of each vowel

from itself since the distance was the same for all vowels). The longest and shortest Mahalanobis distances are 2.28 and 0.74 which belong to the distance of vowel \ae\ from \u\ and vowel \a\ from \i\, respectively. Similar to Table 4-2 in the sustained response feature analysis section, we used three different shades of grey to better show the relation between classification distribution and Mahalanobis distances. In general, the darker the grey, the higher the classification rate and the longer the distance is.

Conf Matrix Transient EFR		Predicted Vowels				
		\a\	\ae\	\e\	\i\	\u\
Actual Vowels	\a\	7	11	15	7	8
	\ae\	8	21	6	9	4
	\e\	2	7	16	12	11
	\i\	5	9	5	18	11
	\u\	1	2	8	7	30

(a)

Mahal Dist EFR		Sample Vowels				
		\a\	\ae\	\e\	\i\	\u\
Reference Vowels	\a\	1.00	1.76	1.94	2.05	1.55
	\ae\	0.92	1.00	1.58	1.63	1.52
	\e\	0.91	1.44	1.00	1.49	1.05
	\i\	0.74	1.08	1.20	1.00	0.97
	\u\	1.42	2.28	1.89	2.20	1.00

(c)

Conf Matrix Transient EFR		Predicted Vowels			
		\ae\	\e\	\i\	\u\
Actual Vowels	\ae\	24	12	11	1
	\e\	13	15	7	13
	\i\	10	14	11	13
	\u\	10	7	10	21

(b)

Classification Accuracy
>30% (15-48)
10%-30% (5-14)
< 10% (0-4)

Mahalanobis Distamce
<= 1
1-2
> 2

Table 4-4: (a) The confusion matrix for the group of 4 transient features which provide the highest classification accuracy among all possible combinations of the transient features (i.e. latency and height of V and A). (b) The confusion matrix without vowel \a\. The darker the grey the higher the classification rate is and vice versa. (c) Mahalanobis distances between all possible pair-wise combinations of vowels for the 4 transient features. The darker the grey, the smaller the Mahalanobis distance is and vice versa.

Table 4-5 shows means and Standard Error of the Mean (SEM) of the transient features for all vowels and Figure 4-7 to Figure 4-10 show the corresponding graphs for each feature. To further investigate the transient features, one-way ANOVA was performed on each feature in order to identify features with significant mean differences of the means among the five classes. The last row of Table 4-5 lists p-

values of one-way ANOVA for each feature in which three shades of grey are used to signify the p-values lower than three standard thresholds (the darker the grey, the lower the threshold is).

Class Labels	Stats	V latency (ms)	A Latency (ms)	V Height (μV)	A Height (μV)	VA duration (ms)	VA height (μV)	VA slope (μV/ms)
vowel \a\	Mean	8.06	9.78	0.41	-0.05	1.72	0.46	-0.27
	SEM	0.11	0.14	0.02	0.02	0.10	0.03	0.01
vowel \ae\	Mean	7.78	9.87	0.38	-0.10	2.08	0.48	-0.23
	SEM	0.10	0.18	0.02	0.04	0.15	0.04	0.02
vowel \e\	Mean	8.25	10.04	0.46	-0.05	1.78	0.51	-0.31
	SEM	0.11	0.15	0.03	0.03	0.10	0.04	0.02
vowel \i\	Mean	8.33	10.37	0.45	-0.01	2.05	0.46	-0.23
	SEM	0.15	0.21	0.03	0.03	0.13	0.03	0.01
vowel \u\	Mean	8.79	10.34	0.37	-0.05	1.54	0.42	-0.28
	SEM	0.09	0.14	0.02	0.03	0.11	0.03	0.01
ANOVA: p-value		< 0.001	< 0.05	< 0.05	0.3248	< 0.01	0.4293	< 0.01

Table 4-5: Means and Standard Error of the Mean (SEM) of the seven transient response features for all 48 trials for each class. The ANOVA p-values are listed on the last row of the table. The three shades of grey represent significance level such that the darker the grey the lower the significance level.

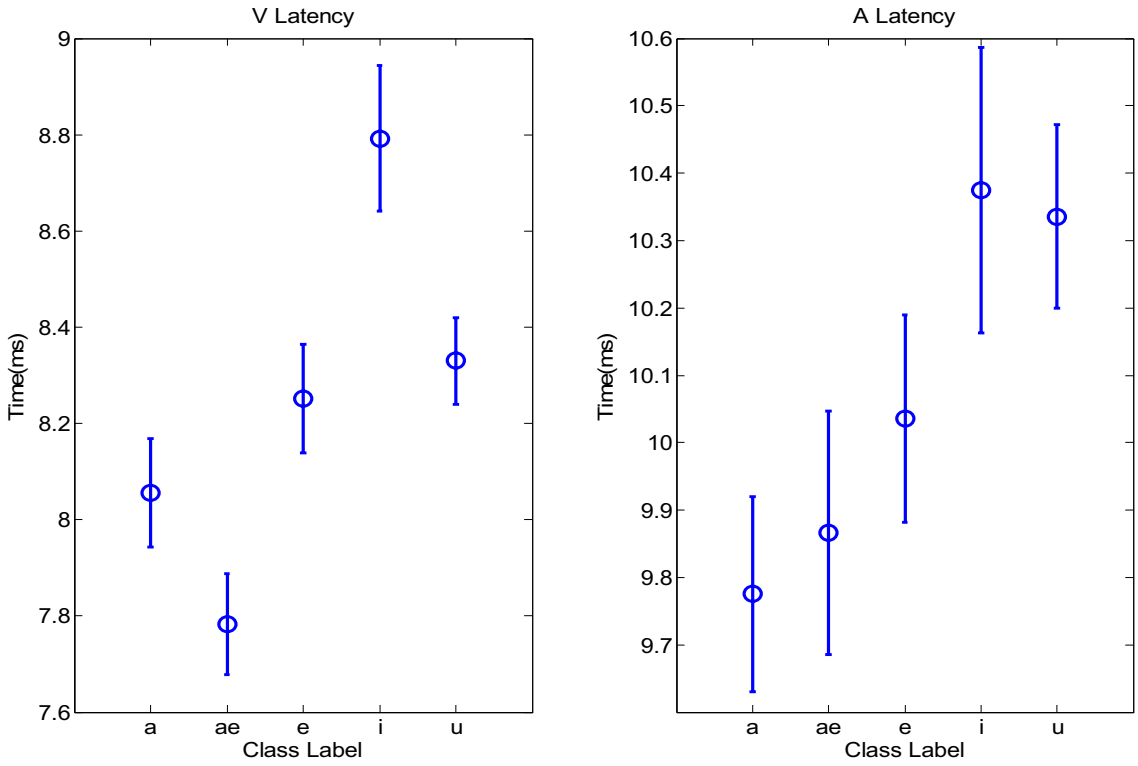

Figure 4-7: Mean and SEM of the latency of waves V and A for 48 trials of each class.

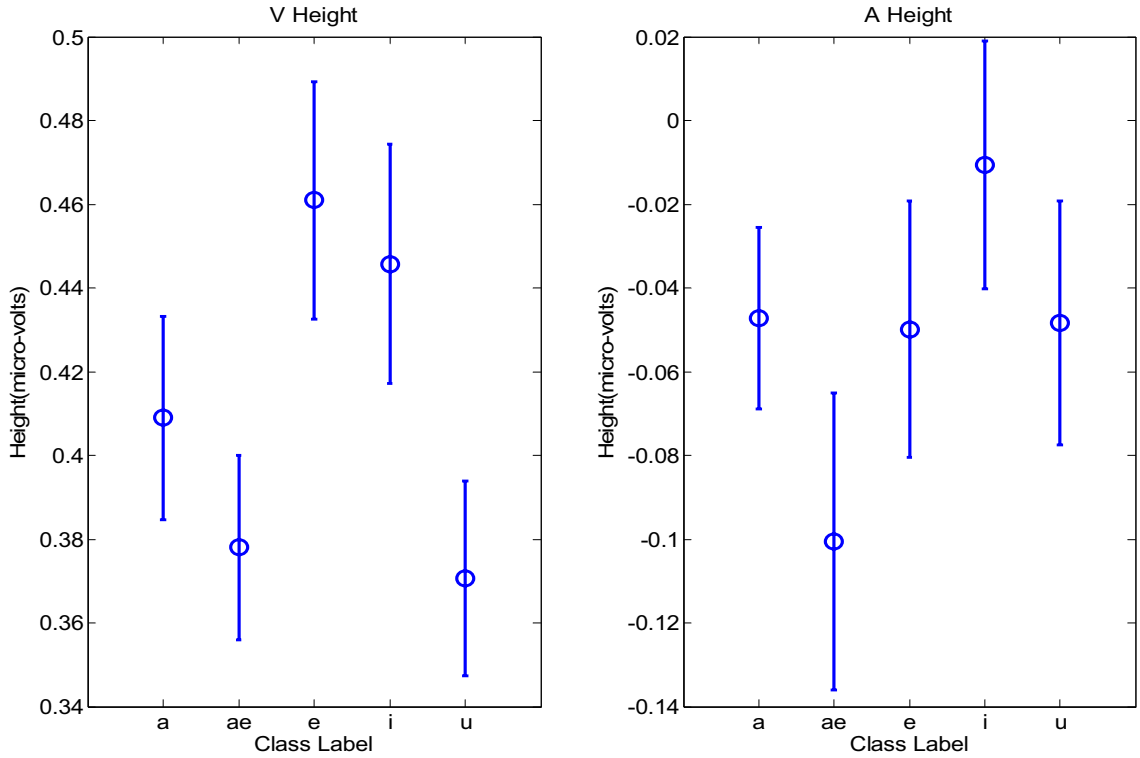

Figure 4-8: Mean and SEM of the height of waves V and A for 48 trials of each class.

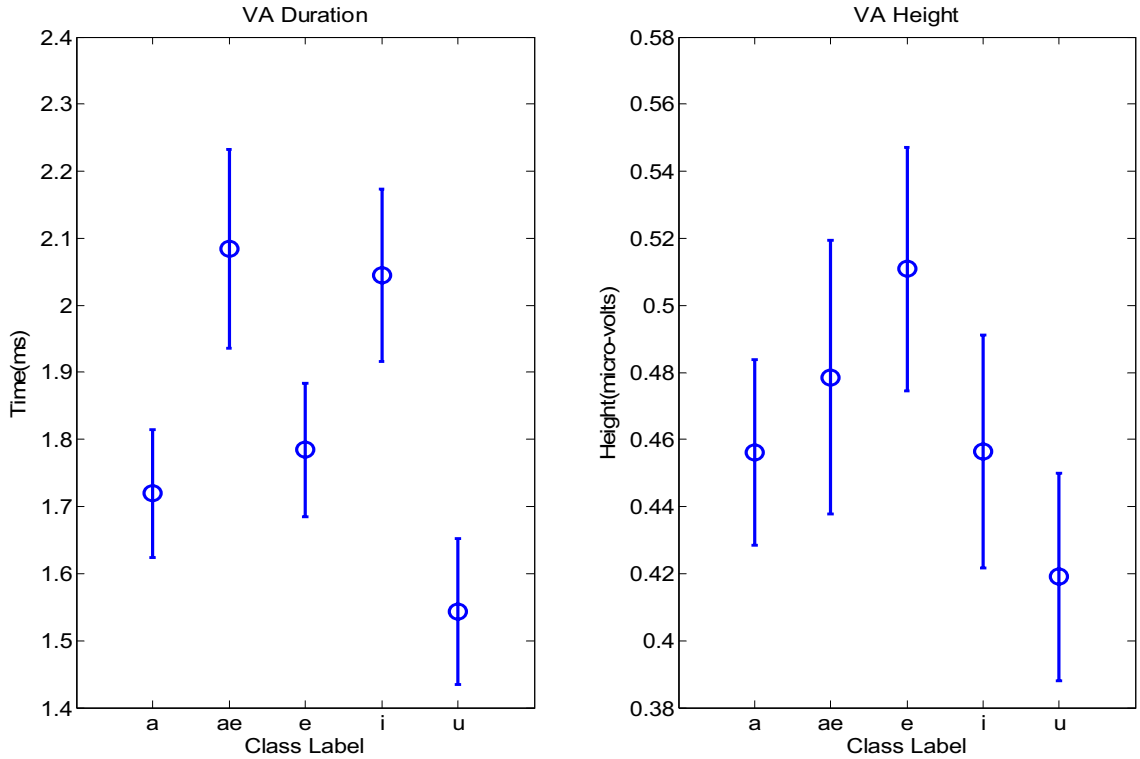

Figure 4-9: Mean and SEM of the duration and height of the VA complex for 48 trials of each class.

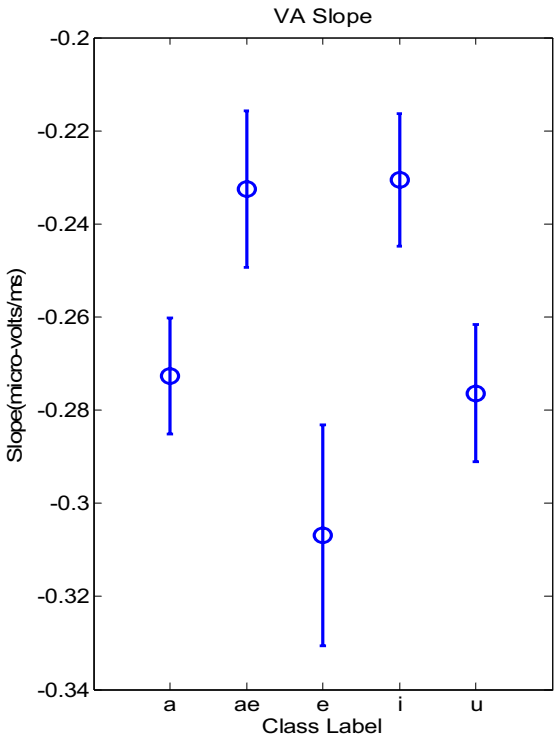

Figure 4-10: Mean and SEM of the slope of the VA complex for 48 trials of each class.

4.3 Classification of SpEPs using Sustained & Transient Response Features

The combination of all sustained ("EFR+FFR") and the best 4 transient response features provide a classification accuracy of 83.33% which is 3% higher than the highest classification accuracy obtained using the "EFR + FFR" feature set (also, SVM provided a classification accuracy of 81.25%). Table 4-6 shows the corresponding confusion matrix along with four shades of grey (Table A-4 in Appendix-B shows SVM confusion matrix).

Conf Matrix Sustaind + Transient		Predicted Vowels				
		\a\	\ae\	\e\	\i\	\u\
Actual Vowels	\a\	40	4	4	0	0
	\ae\	5	34	7	0	2
	\e\	4	4	38	0	2
	\i\	0	0	0	47	1
	\u\	0	1	6	0	41

Classification Accuracy
>40% (20-48)
20%-40% (10-19)
10%-20% (5-9)
0.02%-10% (1-4)
0% (0)

Table 4-6: Confusion matrix for the combination of sustained and transient response features.

5 Discussion

The classification results show that the sustained and transient components of SpEPs with five English vowels can be identified through a basic classification method, which to the best of our knowledge, is the first attempt in speech recognition using SpEPs measured using surface electrode. This finding suggests that the different components of SpEPs carry useful information for discriminating speech stimuli. In the next three subsections (5.1, 5.2, and 5.3) we discuss about the classification results and properties of the sustained and transient response features.

5.1 Classification of SpEPs using Sustained Response Features

Classification accuracy results obtained from the sustained response experiment demonstrate that we were able to successfully classify SpEPs of five different vowels with an accuracy of 80.83%, which is considerably higher than the chance level of 20% (100% / 5 vowels = 20%).

The ability to classify the five English vowels using SpEPs demonstrates that brainstem neural responses in the region of F0 and F1 contain valuable information for discriminating vowels. Both the EFR and FFR features provide a classification accuracy that is considerably higher than chance, demonstrating speech information in both signals. Together the EFR and FFR features provide the highest classification accuracy followed by the individual EFR and FFR features. The fact that the EFR features provide higher classification accuracy than the FFR features is a key finding, because so far it has been thought that the phonetic information is mainly contained in the speech formants, which get reflected in the FFR, and not in the speech envelope, which gets reflected in the EFR (Greenberg et al., 2004). The following two subsections (5.1.1 and 5.1.2) provide more details about the properties of the EFR

and FFR amplitude features. In subsection 5.1.3, we provide a discussion on assessing the classification distributions of all three feature sets.

5.1.1 Investigation on the Properties of the EFR Amplitude Features

The classification accuracy of 69.58% with the EFR was relatively high especially in comparison with the FFR. This finding implies that the EFR amplitude features may contain more perceptually relevant information compared to the FFR amplitude features. This is a novel finding since neural activity that corresponds to the envelope of speech and its harmonics can be used to distinguish the vowels, whereas vowels are usually thought to be perceptually discriminated based mainly on the formant frequencies, and in particular the relative frequencies of F1 and F2 (Advendano et al., 2004; Assmann and Summerfield, 2004; Peterson and Barney; 1952). Moreover, this finding is different from the model that Kraus and Nicol proposed in which the "source" of speech, reflected in components of the response in the region of F0, and "filter" of speech, reflected in components of the response in the region of F1, are processed separately (Kraus and Nicol, 2005). In our study, the differences in the envelope shapes of the vowels can only be due to differences in the formant content since the impulse train source signal used to synthesize all the vowels was the same. Therefore, differences among the responses to vowels in the region of F0 in the EFR cannot be said to reflect any differences in the "source". Instead, they correspond to neural activity that results from differences in the "filter". This neural activity occurs at frequencies mostly well below the first formant and yet allows for vowel discrimination.

By looking at the grand average of the EFR and FFR amplitude spectra, shown in Figure 3-4 and Figure 3-5, it appears vowels may be differentiated using the EFR spectrum better than using the FFR spectrum. This can be better seen from the 2-D

ULDA representation of the EFR and FFR features in Figure 4-4 and Figure 4-5 respectively.

The strength of the EFR amplitude features can be due to two reasons. The first reason is that we expect to get different EFR components for different vowels since the shape of envelope is different for each vowel as shown in Appendix A. The difference in the shape of envelopes is caused by the three formants of each vowel. The second reason is that the EFR contents which are introduced by non-linearities of the cochlea and the different neural centres leading up to the upper brainstem during processing of speech envelope. In general, as it was mentioned in section 2.4.2, for the EFR we would expect to see the highest spectral amplitude at F0 and the lower spectral amplitudes at harmonics of F0. However, in this work we observed some cases which do not follow this assumption. For example, in Figure 3-4 the EFR spectral amplitude of vowel \a\ at F0 (100Hz) is very small compared to other harmonics. Other examples are the EFR spectral amplitudes of vowel \ae\ and \e\ at F0 and 2F0 which are almost equal. These behaviours can be due to some non-linearities which are introduced by the rectification process of the speech envelope within the cochlea and by non-linearities in neural processing (Aiken, 2008). These effects could result in EFR spectral content that differentiates well between the vowels.

5.1.2 Investigation on the Properties of the FFR Amplitude Features

Although the FFR amplitude features provided a classification accuracy of 54.58%, which is more than 2.5 times the chance accuracy, these features demonstrated less phonetic information compared to the EFR features. The strength of the FFR amplitude features can also be partly due to FFR content which is introduced by non-linearities of the cochlea. Ideally, as it was mentioned in section 2.4.2, we would expect to see strong spectral amplitudes at harmonics of F0 near the first formant

(F1) of each vowel; therefore, we would expect to observe similar FFR spectral amplitudes for the vowels with similar F1. However, this statement is not true for all cases in this study. For example, vowels \i\ and \u\ have similar F1 values (270Hz and 300Hz respectively) while their SpEPs do not look similar in Figure 3-5 and they have the highest classification accuracies (77%) among all vowels. Such dissimilarities could be explained by distortion products which are caused by the nonlinearity characteristic of the cochlea (Aiken, 2008). This phenomenon generates spectral amplitudes at 2Fa-Fb and Fb-Fa, Fa and Fb can refer to any multiple integer of F0 as long as Fa<Fb. A good example of distortion products can be seen in Figure 3-5 where the spectral amplitude at 2F0 (200Hz) is very strong in "vowel \a\-FFR" graph while we would expect to only observe strong peaks at harmonics of F0 that are close to the first formant of vowel \a\ (i.e. 700Hz). This component could be caused by 2F1-F2 (F1 and F2 are the first and second formant of the vowel \a\). Another example is the spectral amplitude at 2F0 in "vowel \e\-FFR" graph possibly generated by the effect of distortion product in FFRs.

The weakness of the FFR amplitude features could be due to three reasons. The first reason is that for most of the vowels (\a\, \ae\, and \i\), only the FFR in the region of F1 was included in the analysis; responses at F2 and F3 were generally omitted, as explained in section 3.5.1.1. Adding the responses of higher formants (especially F2), if they are available in the SpEP, might improve the FFR features by providing additional distinct information specific to each vowel (Peterson and Barney; 1952). This reasoning can be supported by considering accuracies of vowels \a\ and \u\ in confusion matrix of FFR (Table 4-2-(c)) which have the two highest accuracies among all the tested vowels. The second formants of these two vowels are near 800 Hz (Table 3-1) which means that F2 information could have been included in the FFR features thus potentially contributing to the highest accuracies.

The second reason for the weakness of the FFR amplitude features is that the F1 frequencies for vowels \a\, \ae\, and \e\ are similar and also similar for vowels \i\ and \u\ (Table 3-1). This could generate overlapping response peaks at harmonics of F0 around F1 frequencies. For instance, vowels \a\ and \ae\ have the closest F1 frequencies among all vowels. As shown in Table 4-2-(c), they were mainly misclassified with each other. However, there are instances in Table 4-2-(c) that show this may not be applicable to all vowels. For example, vowels \a\, \ae\, and \e\ were highly misclassified with vowel \u\ even though their F1 frequencies were not similar.

The third reason for the weakness of the FFR amplitude features is that the FFR is generally a weaker response (in amplitude) than the EFR and so the natural biological variability could impact the FFRs such that the auditory phase-locking to F1 was not consistent across subjects. As a result, the FFR spectra of different vowels may not have been strongly distinguishable when the responses were combined from all subjects. This claim can be confirmed by comparing the variability of the individuals' classification accuracies for the EFR and FFR features in Table 4-1. The difference between the highest and lowest individual's accuracies is 30% (83.33% - 53.33%) for the EFR features and 53.33% (80% - 26.67%) for the FFR features. This indicates that the FFR features are more subject dependent compared to the EFR features. Another observation to support this claim is that subject 4 gave the lowest classification accuracy of 26.67% using the FFR features while this subject provided a classification accuracy of 70% using the EFR features. The reason is that the SpEP recordings from this subject contained more artefact/noise compared to other subjects' recordings. This was due to the fact that he had very dense hair and it was difficult to maintain the recording electrode directly on his scalp while keeping the impedance lower than 5 kΩ as was discussed in section 3.3. This indicates that the FFR is more vulnerable to noise than EFR.

5.1.3 Investigation on the Classification Distribution

In order to support the results of confusion matrices, Mahalanobis distances were calculated for the same pairs of vowels. As can be seen in Table 4-2, for the most part, the shorter Mahalanobis distances generally correspond to higher misclassification rates and longer Mahalanobis distances generally correspond to lower misclassification rates. This can be better observed by comparing the shades of grey of the matching cells. Although the colours are not the same for all matching cells, they are close enough to confirm that when there is high misclassification rate, the Mahalanobis distance is low and vice versa. However, there are a few cases which do not follow this rule. Two good examples are cells (5,1) and (5,2) in Table 4-2-(b) and (e) which show that the LDA classifier was able to discriminate the vowels and correctly classify them despite the low Mahalanobis distance. Another example of this is cell (2,1) in Table 4-2-(c) and (f) which shows that vowel \ae\ was more classified to vowel \a\ than vowel \ae\, while the Mahalanobis distance of vowel \ae\ from itself (cell (2,2) in Table 4-2-(f)) is smaller than the Mahalanobis distance of vowel \ae\ from vowel \a\ (cell (2,1) in Table 4-2-(f)). Such discrepancies between the two analysis approaches can be explained based on two reasons. The first reason is that the main difference between applying LDA and Mahalanobis distance on two classes is that LDA uses the pooled covariance matrix to build a hyperplane between the two classes, whereas Mahalanobis distance uses a covariance of a single class, depending on which class is taken as a reference, to calculate the distance. The second reason is that we used different methods to test and train for each analysis approach. For LDA, we used leave-one-out, whereas for Mahalanobis distance we considered all samples of one class as "Reference Vowels" and all samples other class as "Sample Vowels".

5.2 Classification of SpEPs Using Transient Response Features

The fact that we were able to classify SpEPs of the English vowels using the transient response features indicates that the process of encoding of vowels begins even before processing the pitch and formants of the stimuli. In other words, this result suggests that the neural response from the lower brainstem response and in particular during transition between the lower and upper brainstem (i.e. VA complex) carries vowel-specific information. As it was explained in section 2.5, previous studies have found differences in the transient response when the stimulus was a consonant-vowel syllable with different initial consonants primarily because of the different gaps that follow the consonant (Johnson, et al., 2008; Skoe et al., 2011). However, to the best of our knowledge, nobody has distinguished between separate vowels using the onset response or suggested that it contains phonetic information.

The LDA classification result obtained from the best combination of the transient response features (V,A latency and height) demonstrates that we were able to successfully classify SpEPs of the five vowels with an accuracy of 38.33%, which is almost double the chance level of 20% (100% / 5 vowels = 20%). We verified that this accuracy is different from chance level using the one-tailed binomial test with the following set of variables,

Total number of samples: n=240,

Number of samples that were correctly classified: k=91 (240x38.33%),

The probability of occurrence of one class: a=0.2,

The probability of occurrence of the other classes: b=1-a=0.8

The Binomial test result gives a p-value < 0.001, which indicates that it is extremely unlikely that the accuracy of 38.33% happened by chance.

As can be seen from the confusion matrix in Table 4-4-(a), all vowels have been classified correctly with an accuracy above 33.33% except for vowel \a\ which has

an accuracy of 14%. The ULDA distribution of vowel \a\ in Figure 4-6 also shows that the samples of vowel \a\ are more scattered compared to the samples of the other four vowels. To assess the impact of vowel \a\ samples on the classification accuracy, we performed LDA classification on the data without vowel \a\ samples. As shown in Table 4-4-(b), after removing the samples of vowel \a\ the accuracy of vowel \ae\ was improved by 6%; however, the accuracies for the three other classes were diminished.

In the next subsection (5.2.1) we provide a discussion on the properties of the transient response features. In subsection 5.2.2, we discuss the significance of the transient response. Finally in subsection 5.2.3, we talk about assessing the classification distribution for the transient response features.

5.2.1 Investigation on the Properties of the Transient Response Features

In general, the transient response features have provided lower classification accuracies compared to the sustained response features. The weakness of the transient features could be due to three reasons. The first reason is that the transient response does not include acoustic information of the vowel stimuli (i.e. pitch and formants); therefore, we would expect to see less vowel-specific information in the transient.

The second reason is that we averaged over 500 repetitions to generate a single trial while previous studies averaged over at least 1000 repetitions for a single trial (Russo *et al.*, 2004; Wible *et al.*, 2004; Johnson *et al.,* 2008). The higher number of repetitions reduces noise and provides more robust transient peaks. Using a larger number of repetitions could have helped to obtain other transient peaks including peaks I, II, III, and IV which as a result may have provided a stronger collection of features.

The third reason is that the transient response peaks and trough were identified automatically by finding local maxima and minima, respectively, within a certain time range (as explained in section 3.5.1.2). This could generate outliers, because the shape and location of the transient response landmarks can vary from one subject to another due to biological differences. As such, the results could have improved if the transient landmarks were identified by an auditory clinical expert. Figure C-1 in Appendix C shows transient responses corresponding to 6 trials of vowel \a\ from subject 1. As can be seen, the shape of the waveforms and the positions of the peaks are not consistent among the 6 trials.

5.2.2 Investigation on the Significance of the Transient Response Features

The transient features were further investigated by looking at the mean and SEM of all transient features for each class (Table 4-5). The ANOVA test was performed to identify features which have different means for the five vowels at three different significant levels (i.e. 0.05, 0.01, and 0.001). The result of this test is shown on the last row of the Table 4-5. As can be seen the V latency provided the lowest significant level followed by VA duration and slope, and A latency and V height. It is important to note that three features (i.e. V,A latency and V height) out of the four selected transient features (i.e. V,A latency and height) are among the features with low ANOVA p-value. This confirms that most of the selected features follow ANOVA results. Although VA duration and slope provide the second lowest p-value among all features they could not be used in combination with V latency, which provides the lowest p-value, because these features are highly correlated and they cause singularity as was discussed in section 3.5.1.2.

The fact that the ANOVA test shows significantly small p-value with V latency implies that the V latency distributions of the five classes are different from one another. This difference could be due to the place of activation on the basilar

membrane; higher frequencies get processed earlier because they activate neural responses on the basal (initial) portion of the basilar membrane (Johnson, et al., 2008; Skoe et al., 2011). From the left plot in Figure 4-7, it can be seen that there are significant differences between mean value of different pairs of vowels except for vowels \e\ and \u\. These differences may indicate the effect of the place coding; however, the decreasing order of the mean latencies (/ae/</a/</e/</u/</i/) cannot be correlated to increasing order of any of the three formants of the five vowels. This could be due to the fact that we used three different formants for each vowel so it is not obvious in what order and combination the stimulus formants affect the latency of the transient response. The reason for the similarity of mean values for vowels \e\ and \i\ could be 1) errors in the automatic peak selection, and 2) the low number of repetitions per trial.

Also, we looked at the latency of wave A and sketched its mean and SEM for each class in Figure 4-8 to verify the effect of place coding. Although the ANOVA result states that the means are different with a p-value < 0.05, SEM overlaps can be observed between vowels \a\, \ae\, and \u\ on the one hand and vowels \e\ and \i\ on the other hand. As a result, it is hard to judge the presence of the effect of place coding for this feature.

5.2.3 Investigation on the Classification Distribution

To support the LDA classification results shown in the confusion matrices (Table 4-4-(a) and (b)), we calculated Mahalanobis distances. Also, we used three shades of grey to simplify the comparison between the two measures. As can be seen in Table 4-4-(c), the small Mahalanobis distances on the non-diagonal cells correspond to large number of misclassifications on the matching cells in Table 4-4-(a) and (b). However, there are a few exceptions to this statement. A good example is cell (\e\,\a\) which shows a very small Mahalanobis distance in Table 4-4-(c) while it has

a very low misclassification rate in Table 4-4-(a). Such exceptions can be explained by the difference between LDA classification and Mahalanobis distance which was discussed in the previous section.

5.3 Classification of SpEPs Using Sustained & Transient Response Features

The classification accuracy was increased by 3% when the combination of the sustained and transient response features was used as a feature set. Comparing Table 4-2-(a) and Table 4-6, it can be seen that the classification accuracies in Table 4-6 have increased for the vowels with lower accuracies in Table 4-2-(a) (i.e. vowels \a\, \ae\, \e\, and \u\) than the vowel with the highest classification accuracy (i.e. vowel \i\). For example, the classification accuracy of vowel \u\ has improved by 6% while the classification accuracy of vowel \i\ has not changed.

In order to assess the significance of this improvement, we performed a paired t-test in which the first sample included classification accuracies of the five vowels in Table 4-2-(a), and the second sample included classification accuracies of the five vowels in Table 4-6. The result of the t-test was not statistically significant showing that the 3% increase is likely due to chance. Therefore, we cannot conclude that there could be phonetic information that is mutually exclusive between the two sets of features.

6 Conclusion and Future Work

6.1 Summary of Conclusions

We have demonstrated that the Speech Evoked Potentials (SpEPs) of five English vowels can be classified with fairly high accuracy of 80.33% for the sustained response features and 38.33% for the transient response features, using a Linear Discriminant Analysis (LDA) classifier. We used amplitudes of EFR and FFR as the sustained response features and temporal properties of the VA complex as the transient features. Results show that the EFR amplitude features represents each vowel better compared to the FFR amplitude features. The advantage of the EFR amplitude features was explained to be possibly due to different envelope shape of the vowels and significant EFR content at harmonics of F0 which are introduced by non-linearities of auditory neural system during processing of speech envelope. The disadvantage of the FFR amplitude features was suggested to be due to the limitation of examining only the responses of F1, having similar F1 frequencies for some vowels, and the biological variability of the subjects.

Moreover, we attributed the weakness of the transient response features to the low number of repetitions used to generate a single trial and the automatic approach for selecting the significant transient peaks.

Results obtained from this study demonstrate that SpEPs contain useful information which can be used to distinguish different speech stimuli. Therefore, this work is a solid baseline for further study of SpEP classification using more complex stimuli, such as words. In addition, the high accuracy with the EFR spectral features is a novel finding since it has been thought that the filter characteristics of speech make the main contribution to perceptual discrimination of different vowels. Moreover, the ability to classify the vowels with the transient response features is a

potentially novel finding because the transient response to a vowel has been thought to carry general sound onset information and not vowel-specific information.

6.2 Future Work

The future directions for this study are,

1- Using more complex speech stimuli

In this study we have demonstrated that SpEPs of the five synthetic vowels contain valuable information which can be used for classification of the SpEPs. For future work, natural vowels and more complex speech sounds such as words and sentences can be investigated to provide a better understanding of the speech processing in the auditory system.

2- Applying more complex classifiers

We have obtained a good classification result with a simple classifier like LDA. Using more complex pattern classification methods such as neural network and hidden Markov model may help to achieve a better result especially when more complex speech stimuli are used. Note that given the limited data set in this study, a simple classifier like LDA probably helped to prevent over-fitting.

3- Employing feature selection and feature reduction methods

In this work we simply performed feature selection by choosing spectral amplitudes at particular frequency points. Employing more advanced feature selection and feature reduction methods (e.g., information gain ratio) may help to obtain a higher classification accuracy by providing optimal and consistent features.

4- Collecting data from a larger number of subjects

In general, having a large collection of data is always ideal for classification problems because of two reasons: 1) a part of data can be used only as a test

set, 2) the possibility of over-fitting can be prevented. As such, the classification results of this study may be improved by collecting additional data from a larger number of subjects.

Appendix A: Stimuli in Time Domain

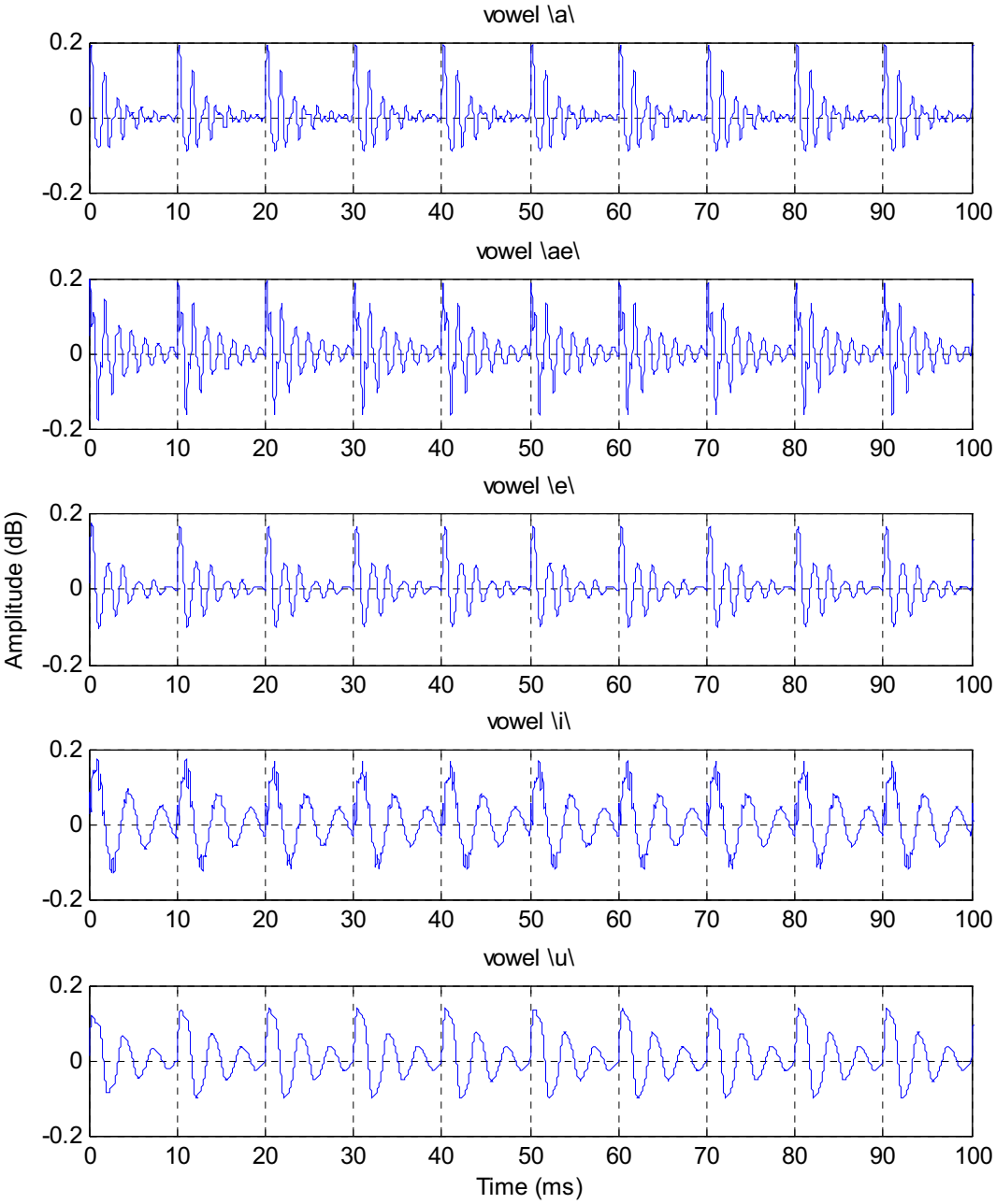

Figure A-1: Time domain representation of five synthetic vowels as spoken by a male with T0=10ms.

Appendix B: SVM Classification Results

Table B-1 shows the overall SVM classification accuracies for the three different sets and Table B-2 provides more details on the classification results by showing confusion matrices.

Amplitude Features	Classification Accuracy
EFR + FFR	80.00%
EFR	69.17%
FFR	53.33%

Table B-1: SVM classification accuracies of three different amplitude feature sets for all trials

Conf Matrix EFR + FFR		Predicted Vowels				
		\a\	\ae\	\e\	\i\	\u\
Actual Vowels	\a\	37	4	6	0	1
	\ae\	3	33	6	0	6
	\e\	5	3	36	0	4
	\i\	0	0	0	46	2
	\u\	0	2	4	2	40

(a)

Conf Matrix EFR		Predicted Vowels				
		\a\	\ae\	\e\	\i\	\u\
Actual Vowels	\a\	34	4	9	0	1
	\ae\	3	28	9	6	2
	\e\	11	10	21	0	6
	\i\	0	0	2	44	2
	\u\	0	0	6	3	39

(b)

Conf Matrix FFR		Predicted Vowels				
		\a\	\ae\	\e\	\i\	\u\
Actual Vowels	\a\	13	22	2	0	11
	\ae\	13	14	10	2	9
	\e\	0	4	28	4	12
	\i\	0	0	3	37	8
	\u\	1	1	4	6	36

(c)

Classification Accuracy
>40% (20-48)
20%-40% (10-19)
10%-20% (5-9)
0.02%-10% (1-4)
0% (0)

Table B-2: SVM confusion matrices for a) EFR+FFR, b) EFR, and c) FFR. The darker the grey, the higher the classification rate is and vice versa.

Conf Matrix Transient EFR		Predicted Vowels				
		\a\	\ae\	\e\	\i\	\u\
Actual Vowels	\a\	6	4	2	17	19
	\ae\	0	12	4	22	10
	\e\	1	3	7	18	19
	\i\	2	5	1	25	15
	\u\	2	1	2	10	33

Classification Accuracy
>30% (15-48)
10%-30% (5-14)
< 10% (0-4)

Table B-3: SVM confusion matrix for the group of 4 transient features which provide the highest classification accuracy among all possible combinations of the transient features (i.e. latency and height of V and A). The darker the grey, the smaller the Mahalanobis distance is and vice versa.

Conf Matrix Sustaind + Transient		Predicted Vowels				
		\a\	\ae\	\e\	\i\	\u\
Actual Vowels	\a\	38	5	4	1	0
	\ae\	3	35	6	1	3
	\e\	4	1	39	0	4
	\i\	0	1	1	44	2
	\u\	0	2	5	2	39

Classification Accuracy
>40% (20-48)
20%-40% (10-19)
10%-20% (5-9)
0.02%-10% (1-4)
0% (0)

Table B-4: SVM confusion matrix for the combination of sustained and transient response features.

Appendix C: Transient Response

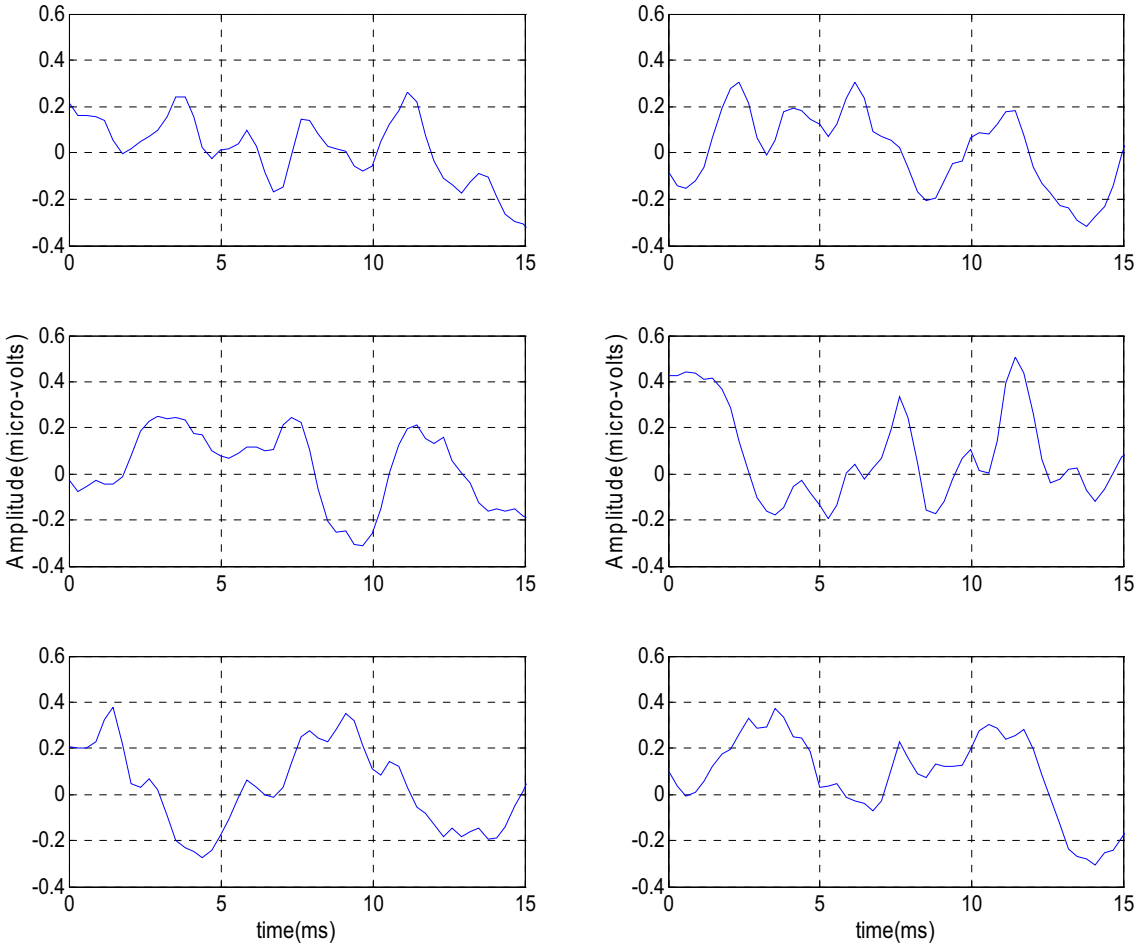

Figure C-2: Transient responses of vowel \a\ obtained from 6 trials of subject 1

References

Advendano, C., Deng, L., Hermansky, H., Gold, B. 2004. *Chapter 2: Analysis and Representation of Speech*. In: Greenberg, S., Ainsworth, W.A., Popper, A.N., Fay,
R.R., (Eds.), *Speech Processing in the Auditory System*, Vol. 18, 1 ed. Springer New York, New York. pp. 63-101.

Akhoun, I., Moulin, A., Jeanvoine, A., Ménard, M., Buret, F., Vollaire, C., Scorretti, R., Veuillet, E., Berger-Vachon, C., Collet, L., and Thai-Van, H. 2008. 'Speech auditory brainstem response (speech ABR) characteristics depending on recording conditions, and hearing status: An experimental parametric study', *Journal of Neuroscience Methods*, vol. 175, no. 2, pp.196-205.

Aiken, S.J. and Picton, T.W. 2008, 'Envelope and spectral frequency-following responses to vowel sounds', *Hearing Research*, vol. 245, pp.35-47.

Aiken, S.J. 2008, Human brain responses to speech sounds, PhD thesis, Institute of Medical Science, University of Toronto.

Assmann, P., Summerfield, Q. 2004. *Chapter 5: Perception of Speech Under Adverse Conditions*. In: Greenberg, S., Ainsworth, W.A., Popper, A.N., Fay, R.R., (Eds.), *Speech Processing in the Auditory System*, Vol. 18, 1 ed. Springer New York, New York. pp. 231-309.

Burkard, R.F., Eggermont, J.J., and Don, M. 2007, Section 1: *What are auditory evoked potentials?.* In: *Auditory evoked potentials: basic principles and clinical application* (*1st ed.*), Lippincott Williams & Wilkins, Philadelphia, pp. 7-20.

Cebulla, M., Stürzebecher, E., and Elberling, C. 2006, 'Objective detection of auditory steady-state responses: comparison of one-sample and q-sample tests', *Journal of the American Academy of Audiology,* vol.17, no. 2, pp. 93-103.

Chandrasekaran, B. and Kraus, N. 2010, 'The scalp-recorded brainstem response to speech: Neural origins and plasticity', *Psychophysiology*, vol. 47, pp.236-246.

Dajani, H. R., Purcell, D., Wong, W., Kunov, H., and Picton, T.W. 2005, 'Recording Human Evoked Potentials That Follow the Pitch Contour of a Natural Vowel', *IEEE Transactions on Biomedical Engineering*, vol.52, pp. 1614-1618.

Deutsch, S. and Micheli-Tzanakou, E. 1987, 'Digital Signal Processing of Evoked Potentials', *Neuroelectric Systems*, New York University Press New York, USA. pp. 450-462.

Duda, R.O., Hart, O.E., and Stok, D.G. 2001a, *Pattern Classification*, 2nd ed. Wiley-Interscience, Toronto (Canada).

Galbraith, G.C., Arbagey, P.W., Branski, R., et al. 1995, 'Intelligible speech encoded in the human brain stem frequency-following response', *Neuroreport*, vol. 6, no. 17, pp. 2363-2367.

Gelfand, S. 2001, *Essentials of Audiology*, Thieme Medical Publishers, 2nd ed. New York. pp. 37-91.

Greenberg, S. 1980, Temporal Neural Coding of Pitch and Vowel Quality, UCLA Working Papers in Phonetics, vol. 52 (Ph.D. Thesis, UCLA).

Greenberg, S., Popper, A.N., Ainsworth, W.A., and Fay, R.R. 2004a, *Speech Processing in the Auditory System*, Vol. 18, 1st ed. Spinger New York, New York.

Jewett, D.L. and Williston, J.S. 1971, 'Auditory-evoked far fields averaged from the scalp of humans', *Brain*, vol. 94, no. 4, pp. 681-696.

Jewett, D.L., Romano, M.N. and Williston, J.S. 1970, 'Human auditory evoked potentials: Possible brain stem components detected on the scalp', *Science*, vol. 167, pp.1517-1518.

Jin, Z., Yang, J. Y., Hu, Z.-S., and Louw, Z. 2001a, 'Face recognition base on the uncorrelated discriminant transformation', *Pattern Recognition*, vol. 34, pp.1405-1416.

Jin, Z., Yang, J. Y., Tang, Z.-M., and Hu, Z.-S. 2001b, 'A theorem on the uncorrelated optimal discriminant vectors', *Pattern Recognition*, vol. 34, pp. 2041-2047.

Johnson, K.L., Nicol, G.T.,and Kraus, N. 2005, 'Brain Stem Response to Speech: A Biological Marker of Auditory Processing', *Ear & Hearing*, vol. 26, pp.424-434.

Johnson, K.L., Nicol, G.T., Zecker, S.G., Bradlow, A.R., Skoe, E., and Kraus, N. 2008, 'Brainstem encoding of voiced consonant–vowel stop syllables', *Clinical Neurophysiology*, vol. 119, pp.2623-2635.

Klatt, H.D. 1980, 'Software for a cascade/parallel formant synthesizer', *Journal of the Acoustical Society of America,* vol. 67, no.33, pp. 971-995.

Kraus, N. and Nicol, G.T. 2005, 'Brainstem origins for cortical 'what' and 'where' pathways in the auditory system', TRENDS in *Neurosciences*, vol. 28, pp.176-181.

Krishnan, A. 2002, 'Human frequency-following responses: representation of steady-state synthetic vowels', *Hearing Research*, vol. 166, pp.192-201

Laroche, M., 2010, A Study of Auditory Speech Processing using Brainstem Evoked Responses under Quiet and Noisy Conditions, M.S.C Thesis, Ottawa-Carleton Institute for Biomedical Engineering, University of Ottawa.

Møller, A.R. and Rollins, P.R. 2002, 'The non-classical auditory pathways are involved in hearing in children but not in adults', *Elsevier, Neuroscience Letters*, vol. 319, pp.41-44.

Møller, A.R. 2003, *Sensory Systems: Anatomy and Physiology*, Academic Press, New York. pp. 97-99.

Møller, A.R. 2006a, *Section I: Hearing: Anatomy, physiology, and disorders of the auditory system*, 2nd ed. Elsevier Science, pp. 3-68.

Møller, A.R. 2006b. *Section II: The Auditory Nervous System, Hearing: Anatomy, physiology, and disorders of the auditory system,* 2nd ed. Elsevier Science London, pp. 75-192.

Nishimoto, S., Vu, A.T., Naselaris,T., Benjamini, T., Yu, B., and Gallant, J.L. 2011, 'Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies', *Current Biology*, vol. 21.

Pasley, B.N., David, S.V., Mesgarani, N., Flinker, A., Shihab, A.S., Corne, N. E., and Knight R.T., and Chang E.F. 2012, 'Reconstructing Speech from Human Auditory Cortex', *PLoS Biology*, vol. 10, no. 1.

Peterson, E.G. and Barney, L.H. 1952, 'Control Methods Used in a Study of the Vowels', *The Journal of the Acoustical Society of America*, vol. 24, pp.175-184.

Picton, T.W., John, M.S. 2004, 'Avoiding electromagnetic artifacts when recording auditory steady-state responses', *Journal of the American Academy of Audiology*, vol. 15, pp.541-554.

Russo, N., Nicol, T., Musacchia, G., and Kraus, N. 2004, 'Brainstem responses to speech syllables', *Clinical Neurophysiology,* vol. 115, pp.2021-2030.

Sachs, M.B. and Young, E.D. 1979, 'Representation of steady-state vowels in the temporal aspects of the discharge patterns of auditory-nerve fibers', *Journal of the Acoustical Society of America*, vol. 66, pp.1381-1403.

Skoe, E. and Kraus, N. 2010, 'Auditory Brain Stem Response to Complex Sounds: A Tutorial', *Ear & Hearing*, vol. 31, pp. 302-324.

Skoe, E., Nicol, T., and Kraus, N. 2011, 'Cross-phaseogram: Objective neural index of speech sound differentiation', *Journal of Neuroscience Methods*, vol.196, no. 2, pp.308-317.

Song, J.H., Banai, K., Russo, N.M., and Kraus, N. 2006,' On the relationship between speech- and nonspeech-evoked auditory brainstem responses', *Audiology and Neurotology*, vol. 11, pp.233-241.

Wible, B., Nicol, T., and Kraus, N. 2005, 'Correlation between brainstem and cortical auditory processes in normal and language-impaired children', *Brain*, vol. 128, pp.417-423.

Wible B, Nicol T, Kraus N. 2004, 'Atypical brainstem representation of onset and formant structure of speech sounds in children with language-based learning problems', *Biology Psychology*, vol. 67, pp.299–317.

Ye, J., Janardan and R., Li, Q. 2004, 'Feature extraction via generalized uncorrelated linear discriminant analysis', *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004.