Automatic Recognition of Speech-Evoked Brainstem Responses to English Vowels

Hamed Samimi

A thesis submitted to the Faculty of Graduate and Postdoctoral Studies in partial fulfillment of the requirements for the M.A.Sc. degree in Electrical and Computer Engineering

Ottawa-Carleton Institute for Electrical and Computer Engineering

School of Electrical Engineering and Computer Science

University of Ottawa

Abstract

The objective of this study is to investigate automatic recognition of speech-evoked auditory brainstem responses (speech-evoked ABR) to the five English vowels (/a/, /ae/, /ao (ɔ)/, /i/ and /u/). We used different automatic speech recognition methods to discriminate between the responses to the vowels. The best recognition result was obtained by applying principal component analysis (PCA) on the amplitudes of the first ten harmonic components of the envelope following response (based on spectral components at fundamental frequency and its harmonics) and of the frequency following response (based on spectral components in first formant region) and combining these two feature sets. With this combined feature set used as input to an artificial neural network, a recognition accuracy of 83.8% was achieved. This study could be extended to more complex stimuli to improve assessment of the auditory system for speech communication in hearing impaired individuals, and potentially help in the objective fitting of hearing aids.

Statement of Originality

This thesis describes the research work that was done by the author at the University of Ottawa for completion of the Master of Applied Science (M.A.Sc) program in Electrical and Computer Engineering. A portion of this research has been reported in the following publication:

Samimi H, Forouzanfar M, Dajani HR, "Automatic Recognition of Speech-Evoked Brainstem Responses to English Vowels", *Proceedings of the 6th IASTED International Conference on Computational Intelligence*, Innsbruck, Austria, 2015.

The results reported in the above mentioned conference paper are reflected throughout this thesis. The author performed various feature extraction and recognition methods along with the data analysis, prepared the document for publication, and made all necessary changes based on feedback from the co-authors and the paper reviewers.

Acknowledgements

I would like to thank my supervisor Dr. Hilmi Dajani for his guidance throughout this work. Without his constant support and feedback this study would not have been completed. I would also like to thank Mr. Amir Sadeghian for his role in collecting the speech-evoked auditory brainstem responses and Dr. Mohamad Forouzanfar for his valuable advice on automatic speech recognition methods.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ABR	Auditory Brainstem Response
CM	Cochlear Microphonic
LDA	Linear Discriminant Analysis
ASR	Automatic Speech Recognition
PCA	Principal Component Analysis
EFR	Envelope Following Response
FFR	Frequency Following Response
ANN	Artificial Neural Network
SOC	Superior Olivary Complex
ANF	Auditory Nerve Fibers
CAP	Cochlear Action Potential
HMM	Hidden Markov Model
ANN	Artificial Neural Network
LPC	Linear Predictive Coding
MFCC	Mel-Frequency Cepstral Coefficients
PLP	Perceptual Linear Prediction
DFT	Discrete Fourier Transform
FFT	Fast Fourier Transform
ICA	Independent Component Analysis
LPCC	Linear Predictive Cepstral Coefficients

1 Introduction

1.1 Motivation

The speech-evoked auditory brainstem response (speech-evoked ABR) can be measured by placing electrodes on the scalp and measuring the compound activity of populations of neurons in the auditory nuclei of the brainstem whose firing patterns follow various components of a speech stimulus (Johnson et al., 2005; Dajani et al., 2005). In fact, the speech-evoked ABR is sufficiently "speech-like" that if played back as an audio signal, it can be heard intelligibly as speech (Galbraith et al., 1995). What makes ABRs different from the cochlear microphonic (CM), that was first observed by Weaver and Bray in 1930, is the fact that unlike the CM, ABRs occur several milliseconds after the stimulus is applied (Galbraith et al., 1995). In contrast, CM occurs almost simultaneously after a sound reaches the tympanic membrane. The delay in response reassures that the ABRs that are recorded from the scalp are of neural origin not the CM (Galbraith et al., 1995).

Recent evidence suggests that the speech-evoked ABR can provide an important window into auditory processing of speech in normal and hearing impaired individuals (Prévost et al., 2013; Anderson et al., 2013). One particular difficulty with hearing assessment is that it is limited by current tests, which usually use artificial signals like tones or clicks that do not allow a clear assessment of auditory function for speech communication (Anderson and Kraus, 2013). Although there are tests of speech perception, these are of no value for assessing the hearing of infants and uncooperative individuals. Understanding speech-evoked ABRs could fill this gap; however, the work done in this area has been limited. In addition, speech-evoked ABRs could

help in the development of objective methods for fitting hearing aids, particularly in patients where subjective tests are not an option, such as newborns and infants (Dajani et al., 2013).

One way in which speech-evoked ABRs could be used in objective hearing aid fitting is to adjust the hearing aid settings so that the evoked neural responses to different speech stimuli are most easily discriminated by an automatic pattern classifier (Dajani et al., 2013). Then, presumably, the brain could learn to readily discriminate these stimuli.

This study focuses on discriminating between speech-evoked ABRs of five English vowels. An earlier study by Sadeghian et al. (2015) investigated the discrimination of speech-evoked ABRs to the vowel stimuli using linear discriminant analysis (LDA), and achieved a best classification accuracy of 83.3% (Sadeghian et al., 2015). However, because the speech-evoked ABR to vowels has many similarities to speech, with a fundamental frequency and a harmonic structure, in this study we investigate the discrimination of the speech-evoked ABRs using techniques used in automatic speech recognition (ASR).

Automatic speech recognition has been studied for more than four decades and there is not a complete solution for it yet. Almost all ASR systems consist of a feature extraction method (front end processing) and a pattern recognition engine. Front end processing is used to extract features in speech signals that ideally uniquely characterize them and so could be used to distinguish between them. The choice of combination of feature extraction method and pattern recognition engine is important as different combinations could provide different overall recognition results.

In this study, the best recognition performance was achieved with the feature set obtained from combining principal component analysis (PCA) applied to the amplitudes of the first ten harmonic components of each of the envelope following response (EFR) and frequency following response (FFR), variants of the speech-evoked ABR obtained with different processing

of the responses, and using the artificial neural network (ANN) as the pattern recognition engine. Our findings show that these techniques are capable of discriminating between speech-evoked ABRs, and that this may help in developing practical applications of these responses in hearing assessment and rehabilitation.

1.2 Objectives

The objective of this study was to evaluate the information carried in speech-evoked ABRs in order to better comprehend the processing of speech by the human auditory system, and so explore the feasibility of using speech-evoked ABRs in new ways for auditory rehabilitation. The approach used different ASR methods to discriminate between the responses to the five English vowels. In this research, only the sustained response of the speech-evoked ABRs was considered, and not the transient response, as the previous study by Sadeghian at el. (2015) showed that the sustained response of the speech-evoked ABRs can be classified with substantially higher accuracy in comparison to the transient response.

1.3 Contributions

The main contributions of this work are:

1. **Demonstrated that speech-evoked ABRs to five English vowels contain sufficient information for automatic discrimination between the vowels.**

   We were able to discriminate between the five English vowels with good accuracy using speech-evoked ABRs. This result confirms the findings of the very recent study by Sadeghian et al., (2015) that speech-evoked ABRs contain useful information that can help in better understanding human auditory processing.

2. **Demonstrated that ASR methods can be used to discriminate between the speech-evoked ABRs.**

   We were able to use standard ASR methods to process the speech-evoked ABRs in order to discriminate between the vowels with good recognition result. This result supports the observation that there are high similarities between speech-evoked ABRs and speech, with a fundamental frequency and a harmonic structure. Therefore, the same methods that are used in ASR can be used for recognition of speech-evoked ABRs.

3. **Demonstrated that both envelope and formant of speech-evoked ABRs can be used to discriminate between different vowels.**

   We were able to use both envelope and formant of the speech-evoked ABRs in order to perform the recognition of the vowels. Although both envelope and formant feature sets provided good recognition results, we were able to confirm that envelope features had a higher recognition rate in comparison to the formant feature set.

To complete this work, the author has both developed and used readily available algorithms for all the feature extraction methods utilized in the artificial neural network system, implemented all the algorithms in Matlab, implemented various feature extraction and recognition methods on the speech-evoked auditory brainstem responses, and conducted the data analysis. In addition, the author made modifications to the hidden Markov model toolkit (HTK) and the Neural Network Toolbox in Matlab to optimize their performance for the speech-evoked ABR signals. The speech-evoked auditory brainstem responses used in this study were collected by Mr. Amir Sadeghian as part of his Master's thesis work.

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

Samimi H, Forouzanfar M, Dajani HR, "Automatic Recognition of Speech-Evoked Brainstem Responses to English Vowels", *Proceedings of the 6th IASTED International Conference on Computational Intelligence*, Innsbruck, Austria, 2015.

1.4 Organization of Thesis

This thesis consists of six chapters that are organized as follows:

- Chapter 2 briefly discusses an overview of the human auditory system, fundamentals of speech-evoked ABRs, and ASR systems.
- Chapter 3 describes the methodology that was used in this study. This chapter discusses the data collection protocol as well as details of the experimental procedure. In addition, data analysis methods such as feature extraction and pattern recognition are described in this chapter.
- Chapter 4 presents the obtained results using different methods of speech recognition.
- Chapter 5 discusses and interprets the results presented in Chapter 4.
- Chapter 6 presents a brief summary of the work and provides recommendations for future work.

2 Background

2.1 The Auditory System and Basic Acoustics

2.1.1 Mechanical System

The ear is an essential organ of the human auditory system. It receives acoustic waves and transmits the sound to the brain through a complicated process. Completion of this process makes us able to detect and interpret sound waves around us. The human ear is divided into three major parts: the outer ear, the middle ear, and the inner ear (Møller, 2006a).

As shown in Figure 2.1, the outer ear is the visible portion of the ear and is responsible for gathering sound waves and passing them to the middle ear and eventually to the inner ear where these waves are converted into electrical impulses that will be perceived by the brain. The outer ear consists of two parts: the pinna or visible part of the ear and the meatus, also known as the ear canal, that connects pinna to the Tympanic membrane. The pinna is angled to enable better hearing for the waves that are generated from the front compared to the ones from the side or the back. This helps in localization of sound sources (Alberti, 1995).

The middle ear is located between the outer ear and the inner ear. It consists of the Tympanic membrane and three bony structures; Malleus, Incus and Stapes. The Tympanic membrane or the eardrum is a very thin membrane which vibrates with the received fluctuating air pressure from acoustic waves and transforms the acoustic energy to mechanical energy by vibrating the above mentioned three bony structures, passing on frequency and amplitude information to them (Alberti, 1995).

After this stage, all the information from sound waves enters the inner ear. This is the innermost part of the ear and is responsible for converting mechanical waves into electrical nerve pulses which travel further to the brain. The inner ear consists of the Cochlea and the Auditory Nerve. The cochlea is a snail shell-shaped organ with two and a half turns and a volume of about 0.2 millilitres. It hosts around 3000 inner hair cells, each of which is innervated by numerous nerve fibers. Inner hair cells are responsible for converting vibration of the basilar membrane that is caused by acoustic waves into electrical pulses. The nerve fibers then carry the electrical pulses to higher centers in the brain, and in some cases receive neural activity from higher centers (Alberti, 1995). The basilar membrane is wider and more flexible at the apex compared to the base that is narrower and stiffer. The way that the basilar membrane responds to acoustic waves helps to clarify early speech encoding in the cochlea (discussed in detail in Section 2.1.4). A diagram of the three parts of the human ear is shown in Figure 2.1 below.

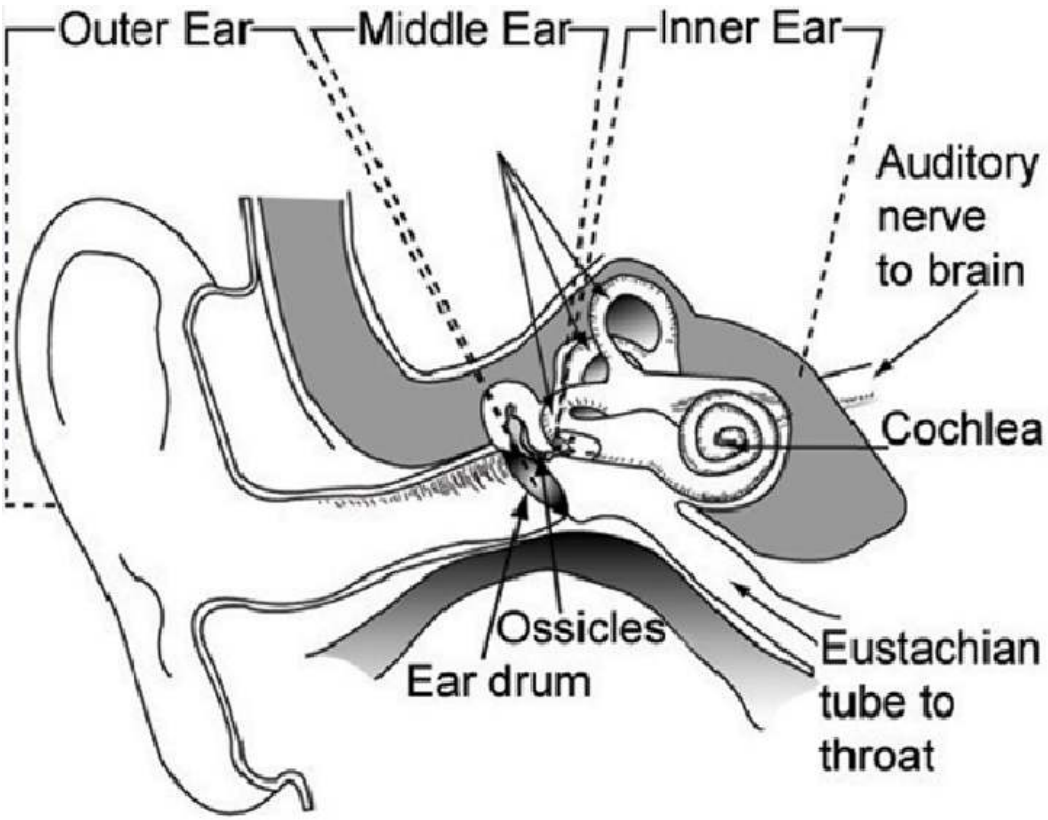

Figure 2-1. The three parts of the Ear (adapted from Wikimedia Commons at https://commons.wikimedia.org/wiki/File%3AOuter%2C_middle_and_inner_ear.jpg).

2.1.2 Nervous System

The auditory nervous system consists of ascending and descending pathways. Both these pathways work in parallel and transfer information between the cochlear hair cells and the brain. While the ascending pathway transfers sensory information from the ear to the brain, the descending pathway sends feedback from the brain to the ear. The ascending pathway itself is divided into two subcategories; classical and non-classical. The non-classical pathway is connected to many parts of the brain such as pain centers, but the function of this pathway remains largely unknown. In the ascending classical pathway, the neurons are arranged in a tonotopic manner. That is, that they are organized based on the different frequencies to which they are tuned (Møller, 2006b).

Auditory Brainstem Responses (ABRs) are a sequence of low amplitude positive and negative waves recorded from the scalp in response to acoustic stimuli, and are believed to be generated along the ascending pathway (Moore, 1987). The first five positive and negative waves in response to a tone burst and their suggested corresponding origin in the auditory pathway are illustrated in Figure 2.2 below.

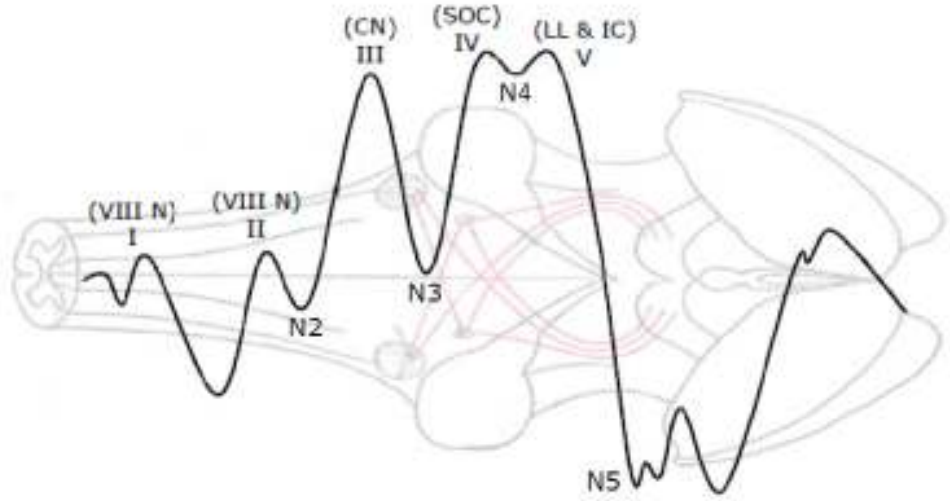

Figure 2-2. Positive and negative peaks of an ABR signal following a tone burst stimulus (adapted with modifications from the Wikimedia Commons at http://commons.wikimedia.org/wiki/File:Lateral_lemniscus.PNG).

There is an assumption that the positive waves are the result of axon activity in fiber tracks while the negative waves are the reflection of dendritic potentials in cell groups (Moore, 1987). Figure 2.2 indicates that Wave I is generated by the cochlear action potential (CAP) and the distal part of the eighth nerve; Wave II comes from the response of the proximal part of the eighth nerve; Wave III is the response of the cochlear nuclei; and Waves IV and V are the responses from superior olivary, inferior colliculus and lateral lemniscus (He et al., 2014). As for the negative waves, Wave N2 is believed to be the first central response in the cochlear nuclei; Wave N3 represents depolarization of the medial olivary nuclei; Wave N4 comes from depolarization in the dorsal lemniscal nuclei; and Wave N5 is the response to depolarization of

neurons in the inferior colliculus (Moore, 1987). However, the exact origin of some of the waves remains controversial.

The descending classical pathway is considered to be the reciprocal to ascending pathway and it consists of two distinct systems; the corticofugal system and the olivocochlear system. Out of these two descending systems, the olivocochlear is the one that is understood the best. The olivocochlear pathway starts from the superior olivary complex (SOC) and ends at the hair cells of the cochlea (Møller, 2006b). A simplified connection diagram of the classical auditory pathway is shown in Figure 2.3 below.

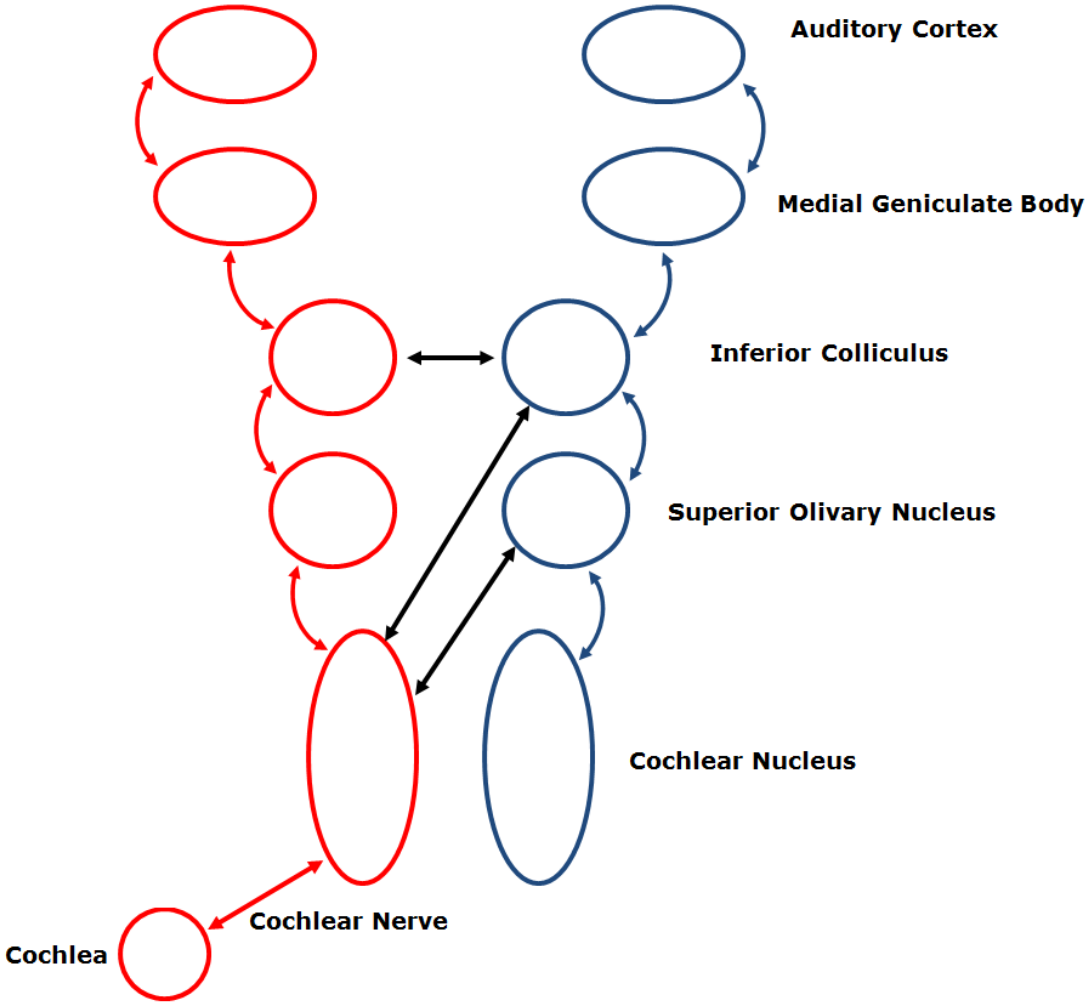

Figure 2-3. A simplified block diagram of the classical auditory pathway (adapted from Wikimedia Commons at http://commons.wikimedia.org/wiki/File:Aud_pathway.png). The speech-evoked ABR is thought to originate mainly from the brainstem, and in particular, from upper brainstem nuclei such as the inferior colliculus.

2.1.3 Sound Characteristics

When waves from any sound source reach the ear, they generate pressure waves by vibrating the air and the auditory receptors respond to these vibrations. Two important characteristics of sound are pitch that is usually based on the fundamental frequency of the sound and loudness that is based on the magnitude of the wave. In infants, the range of audible frequencies is normally between 20Hz and 20kHz. The upper limit of this range usually decreases significantly with age (Siegel and Sapru, 2006).

The range of magnitude of audible sound is large. It can range from 20 μPa (micro Pascals) rms for sounds that are barely audible to 200 Pa for painful sound levels. This wide range leads to limitations in representing the magnitude in absolute units. Therefore a more convenient way to represent sound level is to use a logarithmic scale. In the logarithmic scale, sound pressure is expressed in decibel sound pressure level (dB SPL) and it is calculated as $sound\ pressure\ (dB\ SPL) = 20\ log_{10}\ \frac{pressure\ of\ the\ test\ sound}{reference\ pressure\ of\ 20\ \mu Pa}$ (Siegel and Sapru, 2006). (2.1)

2.1.4 Speech Encoding

Classically, two different mechanisms were considered for speech encoding in the auditory system, rate-place coding and temporal coding. Each of these coding schemes has its own shortcomings - in terms of accounting for human processing of speech - which we will briefly discuss. However, physiological recordings of auditory nerve fibers (ANF) have demonstrated that they are both complimentary to each other and the speech encoding process is likely done by a combination of both mechanisms (Holmberg et al., 2007).

2.1.4.1 Rate-place coding theory

Rate-place coding theory was first introduced by von Helmholtz and his theory of hearing (Holmberg et al., 2007). This theory implies that particular groups of hair cells located at different places along the vibrating basilar membrane in the cochlea respond to different frequency components in the acoustic signal. As shown in Figure 2.4, while high frequency components stimulate one group of hair cells along the basilar membrane of the cochlea (closer to the base), low frequency components excite completely different sets of hair cells (closer to the apex). The separation of activity along such a tonotopic frequency map is thought to continue along different nuclei in the auditory pathway. This separation of neural activity leads to a sort of spatial frequency analysis, where the rate of firing of neurons would reflect the energy in different frequency regions.

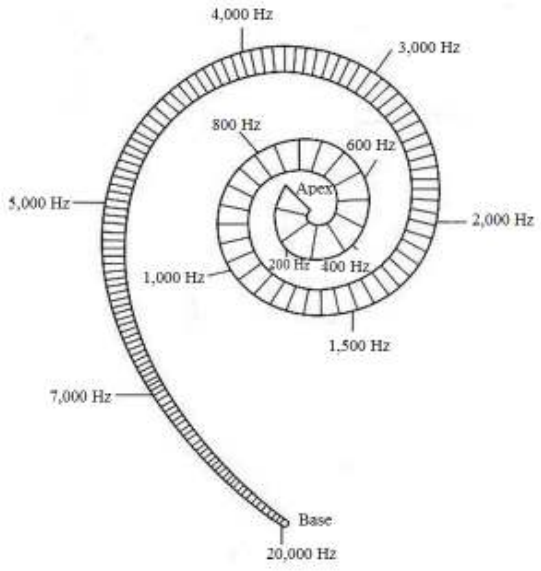

Figure 2-4. Layout of the tonotopic map of the basilar membrane in the cochlea (after Stuhlman, 1943).

The rate-place theory therefore supposes particular places of maximal excitation along the basilar membrane for each frequency and this could partially explain speech encoding in the auditory system. However, it fails to address some perceptual phenomena such as the perception of the pitch of harmonic complexes with a missing fundamental (this is where a few consecutive

harmonics of a signal give a perception of a low frequency signal at the fundamental frequency of those harmonics) (Evans et al., 1978).

2.1.4.2 Temporal coding theory

Temporal coding theory was first introduced by Wundt in 1880 (Eggermont, 2015). Based on this theory, the place of stimulation along basilar membrane is irrelevant to the encoding of a signal. Encoding happens based on the discharge pattern in the auditory nerve. In this mechanism, low frequency signals cause low frequency firing of the neurons based on the relatively slow motion that is initiated by the low frequency signal in the basilar membrane. On the other hand, high frequency signals cause fast motion in basilar membrane hence high frequency firing of the neurons in auditory nerve (Holmberg et al., 2007).

Based on temporal coding theory, the time interval between two subsequent action potentials (or closely grouped action potentials) could lead the brain to potentially recognise the frequency of a pure tone. This property is referred to as phase-locking or synchronous firing (Oxenham, 2013), which is believed to be strong with low frequencies. A number of research studies have shown that good phase-locking is observed in a range of 60-250 Hz, then it begins to decline at about 600 Hz and by the time that frequency exceeds 3.5 KHz phase-locking is not detectable (Palmer and Russell, 1986; Wallace et al., 2002). Since the measurement of firing in the auditory is an invasive procedure, there are no data available from humans on phase-locking and the information available to us is based on studies done on other mammals (Oxenham, 2013).With speech-evoked ABRs, the FFR is formed as a result of a phase-locked response to the harmonics of the stimulus near the region of the first formant F1, while the EFR is a phase-locked response to the envelope of the speech stimulus (Sadeghian et al., 2015); therefore phase locking could be viewed to be the basis for generating the speech-evoked ABR.

2.2 *The Speech Signal*

In humans, speech production can be described using the source-filter model (Rabiner and Schafer, 2010; Kraus and Nicole, 2005). As the name implies, the source-filter model consists of a two stage process: the sound source and a filter. The sound source is the result of the vibration of the vocal folds caused by airflow coming from the lungs, and in a non-tonal language like English, is responsible for the non-linguistic portion of speech. The fundamental frequency (F0) or vocal pitch is a characteristic of the sound source. On the other hand, the filter (transfer function) refers to all the processes that the sound source goes through after the vocal folds (by going through the vocal track, tongue, lips, etc.). The filter is the part that is believed to be responsible for linguistic content of English speech. Filtering of the sound source amplifies certain harmonics of the fundamental frequency at the resonant frequencies of the filter, which are called the formant frequencies (Kraus and Nicole, 2005). The linear model of speech production is the result of convolution between excitation signal which is the source signal and the vocal track impulse response that is referred to as filter (Rabiner and Schafer, 2010). In frequency domain, the convolution is corresponds to multiplication and the resulting speech waveform is generated as shown in Figure 2.5.

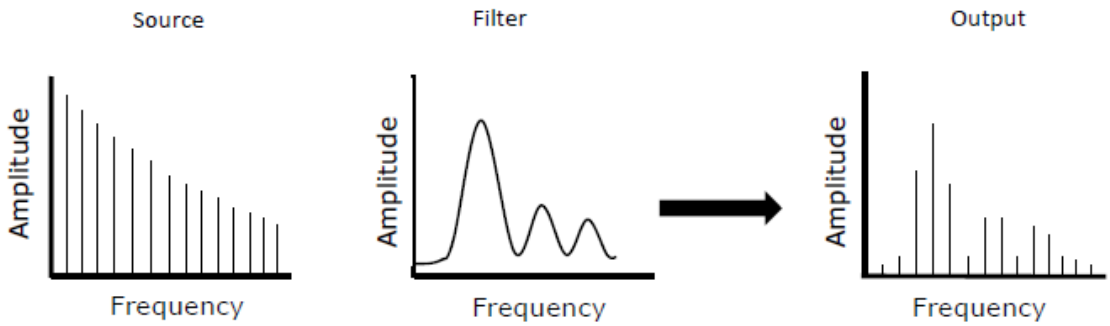

Figure 2-5. Demonstration of the source filter theory (adapted with modifications from Titze, 2008).

The speech production system also determines whether speech is voiced or unvoiced. If the generated speech signal is nearly periodic (such as in vowels) then it is referred to as voiced speech, and if the generated speech is random without periodic nature (such in fricative consonants like /s/ and /f/) it is referred to as unvoiced speech. The voiced speech is believed to be generated by the vibration of the vocal fold periodically interrupting the airflow coming out of the lungs. The periodic interruption does not occur in the case of unvoiced speech; however, along the vocal track partial or complete closure happens that causes the air flow to randomly vary generating the unvoiced speech (Qi and Hunt, 1993).

The fundamental frequency (F0) is influenced by a number of factors that include age, personality, body size, and emotional state (Tielen, 1989). It ranges between 85 and 155 Hz for men and between 165 to 255 Hz for women (Olguin-Olguin and Pentland, 2010). Similar variation also exists in regards to the formants. The frequencies of the formants also change with gender and age since the filter characteristics also depend on these factors (Busby and Plant, 1995). Previous studies have indicated that the formant frequencies decrease as age increases and also formant frequencies tend to be higher in females compared to males (Busby and Plant, 1995). The following four Figures are the time domain and the frequency domain demonstrations of two different vowels. Figure 2.6 and Figure 2.7 are the time domain representations of two synthetically generated vowels, /ae/ and /i/ respectively, spoken by a male voice. These figures show the first 100 ms of each vowel with the fundamental period (T0) of 10 ms. Figure 2.8 and Figure 2.9 are the frequency domain representations of the above-mentioned vowels up to 3.5 kHz. The fundamental frequency of the sound source is at 100 Hz (F0=1/T0), and the first three formants are also represented.

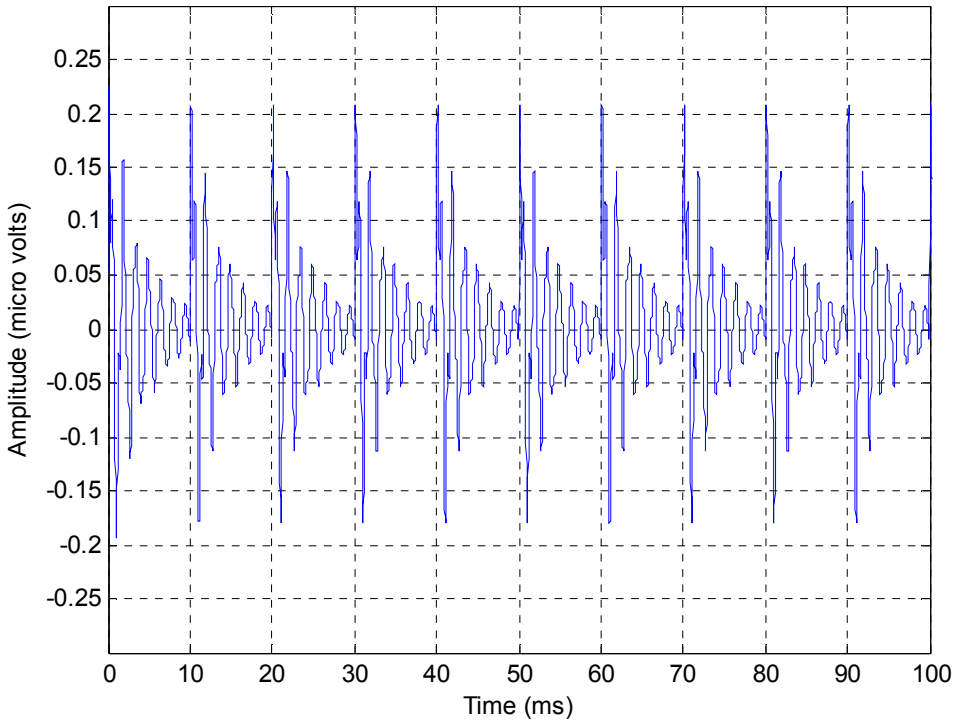

Figure 2-6. Time domain representation of synthetic vowels /ae/ spoken by a male with fundamental period of T0=10ms.

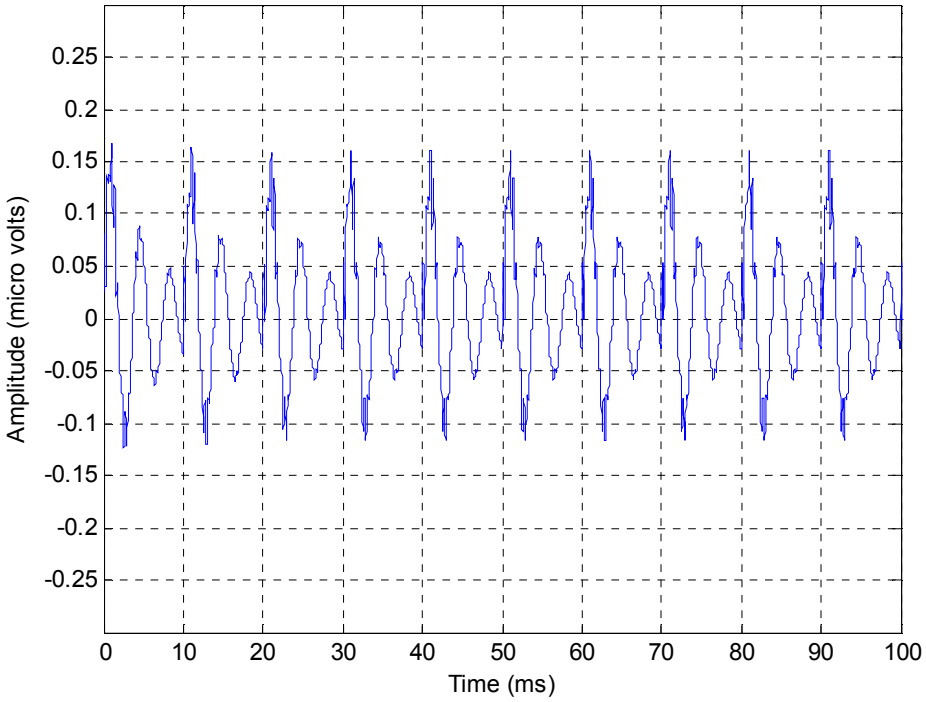

Figure 2-7. Time domain representation of synthetic vowels /i/ spoken by a male with fundamental period of T0=10 ms.

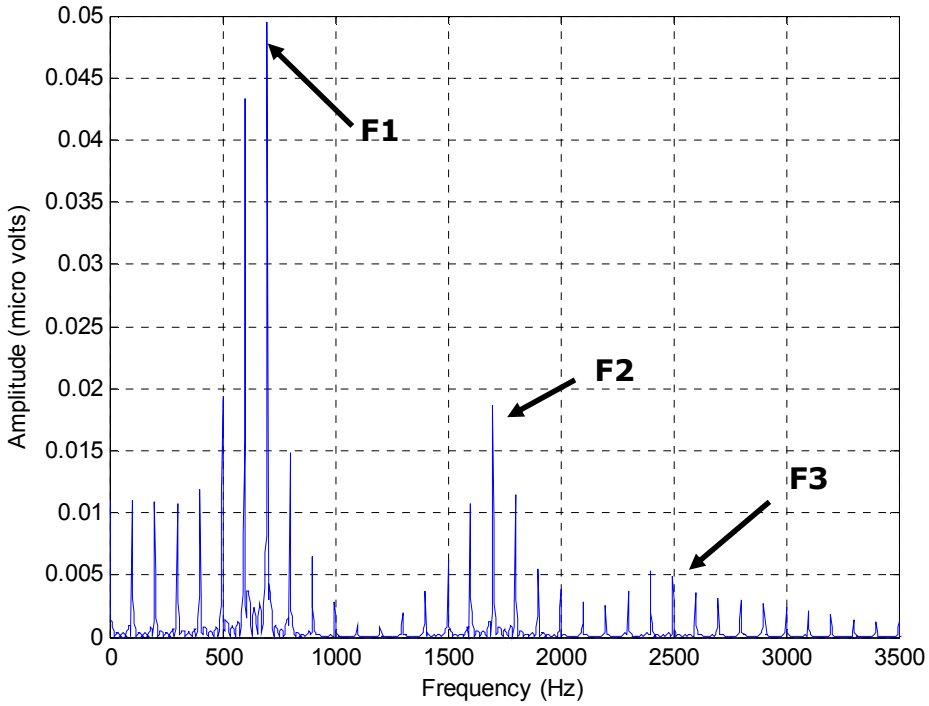

Figure 2-8. Frequency domain representation of synthetic vowels /ae/ spoken by a male with fundamental frequency F0=100 Hz. The first three formants (F1, F2 and F3) that are the result of the filter also are marked in this Figure (F1=660 Hz, F2=1720 Hz, F3=2410 Hz).

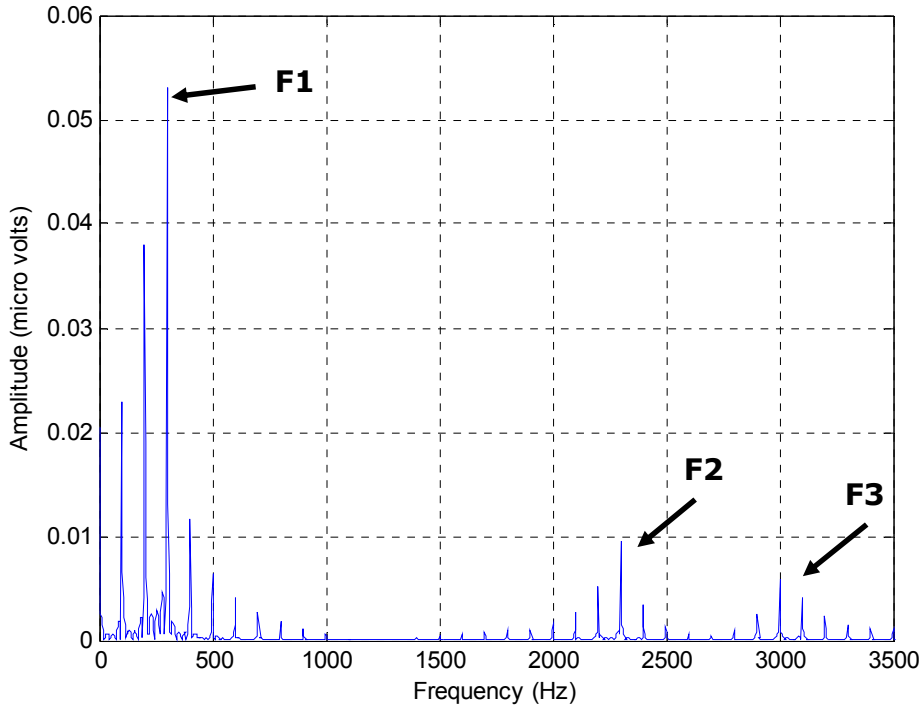

Figure 2-9. Frequency domain representation of synthetic vowels /i/ spoken by a male with fundamental frequency F0=100 Hz. The first three formants (F1, F2 and F3) that are the result of the filter also are marked in this Figure (F1=270 Hz, F2=2290 Hz, F3=3010 Hz).

2.3 The Speech-Evoked Auditory Brainstem Response

Speech-evoked auditory brainstem responses can be measured by placing surface electrodes on the scalp and measuring the electrical potential generated mainly by populations of neurons in the upper brainstem (Thiagarajan, 2015). These responses follow different components in the speech stimulus, and so are "speech-like", with a fundamental frequency, harmonics etc. They can even be understood as intelligible speech when converted into audio signals (Galbraith et al., 1995). The speech-evoked ABR to a steady-state vowel, the stimulus used for this study consists of an initial transient response followed by a sustained response. The transient response is short and consists of several waves (Sadeghian et al., 2015). On the other hand, the sustained response is longer in duration and periodic in the case of a steady-state synthetic vowel, and is quasi-periodic in the case of a natural vowel (Skoe and Kraus, 2010).

2.3.1 The Transient Response

The transient response takes place within 15 ms of the stimulus onset in the human adult. Based on Jewett and Williston's convention (Burkard et al., 2006), there are seven peaks in the transient response; however, in most cases only the amplitude of the first five peaks and their latencies are considered to evaluate the performance of central auditory processing or in an objective hearing test for infants. For example, one study showed that with increasing background noise level, ABR peak amplitudes decreased while their latencies increased (Burkard et al., 2006). It was also shown that the peaks are higher and latencies are shorter in a young individual in comparison with an older person (Burkard et al., 2006). Another study on infants showed that the latencies of all the peaks shorten within the first year of life. In a healthy infant, peak I achieves the same latency as an adult by 3 months of age. The latency of peak III reaches its maturity

between 8 and 16 months of age and the latency of peak V stabilizes at 18 to 36 months where it achieves an asymptotic value (Burkard et al., 2006).

2.3.2 The Sustained Response

With a steady-state vowel stimulus, the sustained response follows the transient response and corresponds to the steady-state portion of the stimulus. The peaks in the sustained response are synchronized to the stimulus fundamental frequency and its harmonics (Johnson, 2008). The sustained response is divided into two different types, the envelope following response (EFR) and the frequency following response (FFR). The EFR represents the neural response that follows the envelope of speech at its fundamental frequency (F0) and at its early harmonics. On the other hand, the FFR represents the neural response that directly follows the harmonics of speech and is usually most prominent in the region of the first formant (F1) (Sadeghian et al., 2015). Typically, the stimulus is presented to the subject in two polarities (original and inverted), and the EFR is obtained by averaging the responses to both stimulus polarities, while the FFR is obtained by averaging the response to one polarity and the negative of the response to the other polarity (Prévost et al., 2013; Aiken and Picton, 2008). A simplified model of how the EFR and FFR are thought to be generated is shown in Figure 2.10 below.

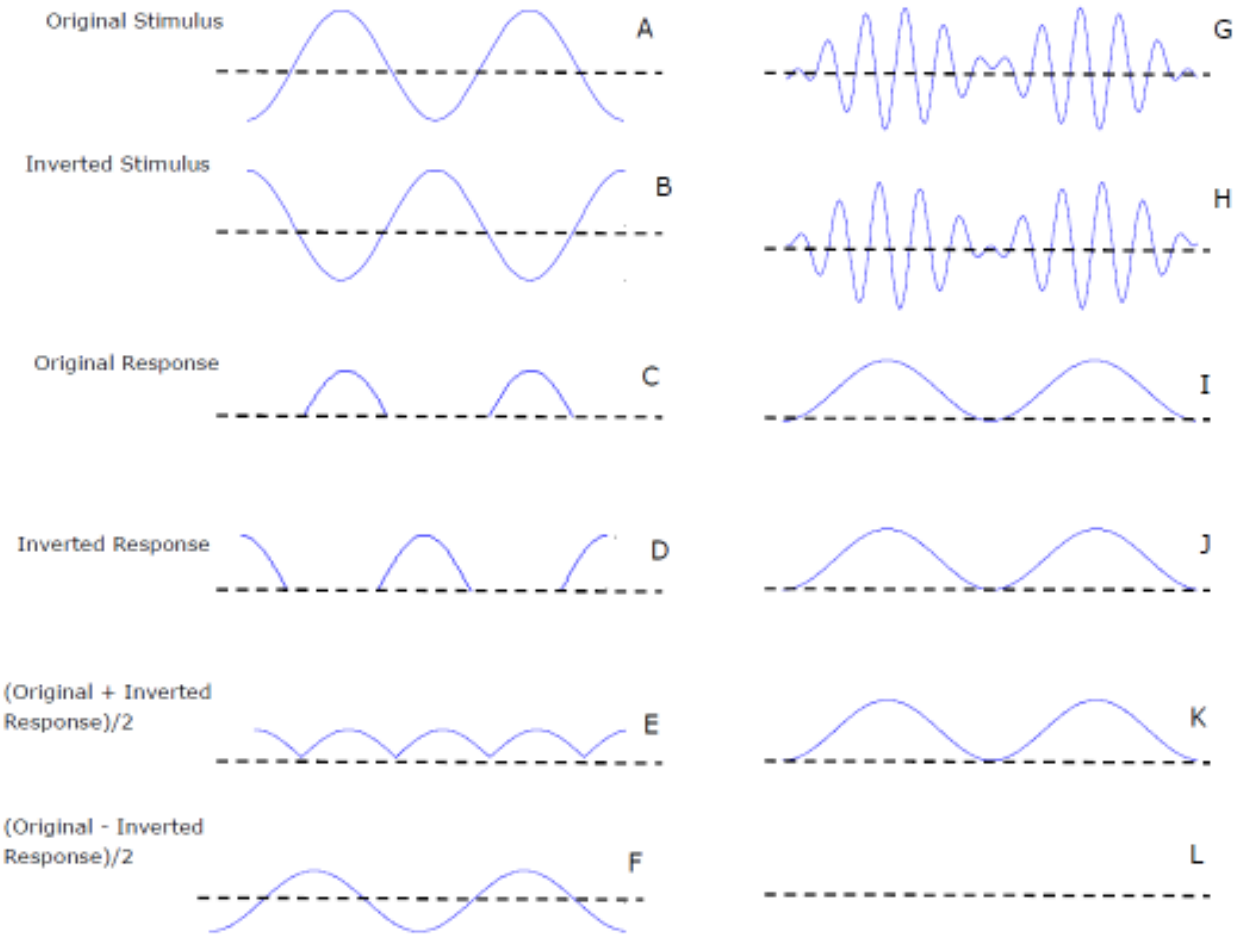

Figure 2-10. Simplified model of how the EFR and FFR are generated (after Aiken and Picton, 2008). See text for explanation.

In Figure 2.10, represents EFR and FFR signals. The stimuli for the FFR are original and inverted 200 Hz tones (A and B), while the stimuli for EFR are opposite polarities of a 2 kHz tone amplitude modulated at 200 Hz (G and H). As shown, the reversal of polarity of the amplitude modulated signal does not have any effect on the modulation envelope. Waves C, D and I, J are the corresponding responses to the above-mentioned stimuli. The rectification of the signals in C and D is due to properties of the hair cells in the inner ear. The resulting signals (E and K) represent the EFR that is the average of the responses to both stimulus polarities while the second set of resulting signals (F and L) represents FFR that is the average of the response to

one polarity and the negative of the response to the other polarity. As shown, the EFR follows the envelope of the amplitude modulated signal, while the FFR the stimulus frequency itself.

2.4 Automatic Speech Recognition

Automatic speech recognition (ASR) has been studied for over four decades. Despite all the research that has been done and all the advancements that have been made in this field, we are still far from achieving the ultimate goal of developing a system that can make machines understand spoken words on any subject by anyone and in any environment with the same accuracy and speed as a human (Rabiner and Juang, 1993; Anusuya and Katti, 2011). The difficulty is that the machine needs to follow in some way the human speech production and perception process and that is a very difficult task. A simplified diagram of the speech production/perception process in humans is presented in Figure 2.11 below.

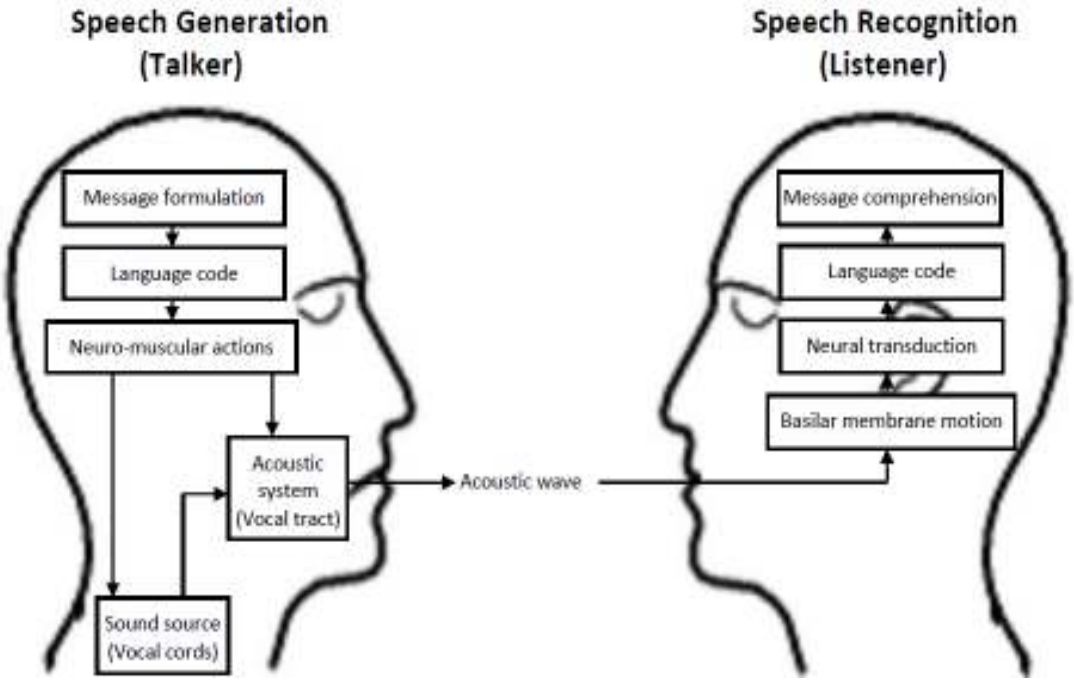

Figure 2-11. Human speech production/perception process (adapted with modifications from Rabiner and Juang, 1993).

There are different methods of speech recognition that have been used either academically for research purposes or commercially in practical systems. In general, speech recognition methods can be categorized in the following three classes: 1) the acoustic-phonetic approach, 2) the pattern recognition approach, and 3) the artificial intelligence approach (Rabiner and Juang, 1993; Anusuya et al., 2009).

The acoustic-phonetic approach relies on the theory of acoustic phonetics. This theory assumes that there are finite unique phonetic units in the speech signal and these units have distinct properties in their waveforms or their frequency spectra. The acoustic-phonetic approach is a two-step process. The first step is to segment and label the speech that needs to be recognized, and the second step is to conclude a valid word from the sequence of phonetic labels of the first step (Rabiner and Juang, 1993; Hemdal and Hughes 1967). The steps that are taken in the acoustic phonetic approach are represented in Figure 2.12 below.

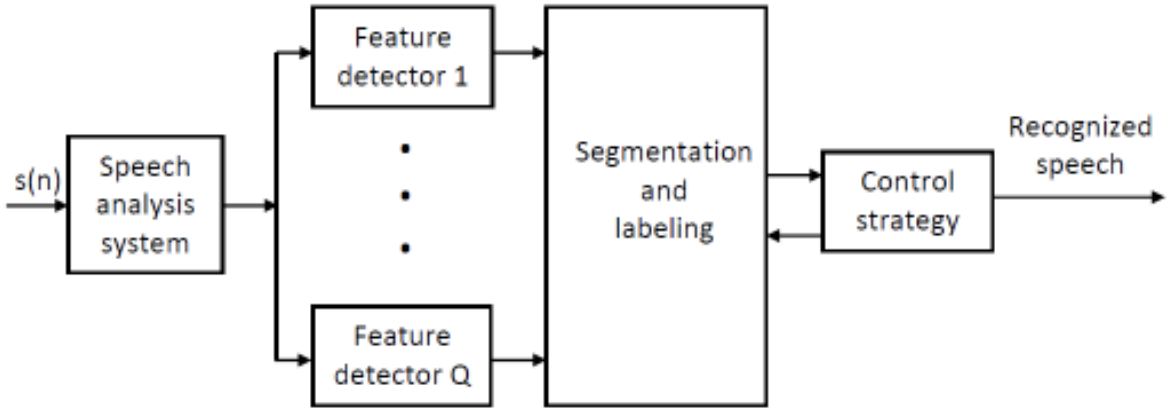

Figure 2-12. Block diagram of the acoustic-phonetic system for automatic speech recognition (from Rabiner and Juang, 1993).

The pattern recognition approach does not use segmentation of speech and instead it uses the speech pattern directly. This method is also a two-step process, training on the speech pattern and recognition of the pattern. The number of training data samples plays an important role in this approach as it provides information about the pattern of speech in a variety of contexts.

When the machine has enough information about the pattern, the pattern recognition component comes into play and makes a direct comparison between the unknown speech and the training library in an effort to recognize the speech (Rabiner and Juang, 1993; Itakura 1975).

Although different approaches have been used in speech recognition, the pattern recognition approach is the one that has received more attention due to its relative ease of implementation, simplicity, robustness, and proven high performance (Rabiner and Juang, 1993). The major steps in the recognition of speech using this approach are discussed in more detail later in the chapter, and are illustrated in Figure 2.13 below.

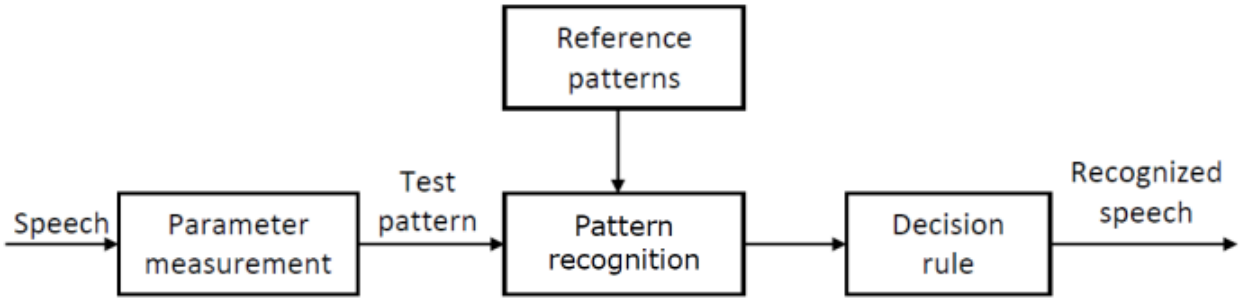

Figure 2-13. Block diagram of a pattern recognition system for automatic speech recognition (adapted from Rabiner and Juang, 1993).

The artificial intelligence approach is a combination of both the acoustic-phonetic approach and the pattern recognition approach. This method is used to automatically recognize speech by applying the same type of intelligence that human uses to make decisions based on a set of measured acoustic features (Rabiner and Juang, 1993; Gaikwad et al., 2010). In this method, knowledge from a variety of sources such as acoustic knowledge, syntactic knowledge etc. is considered to recognize a speech signal (Rabiner and Juang, 1993). The artificial intelligence approach could be implemented in several different ways. The bottom-up process that is the most standard approach is illustrated in Figure 2.14 below.

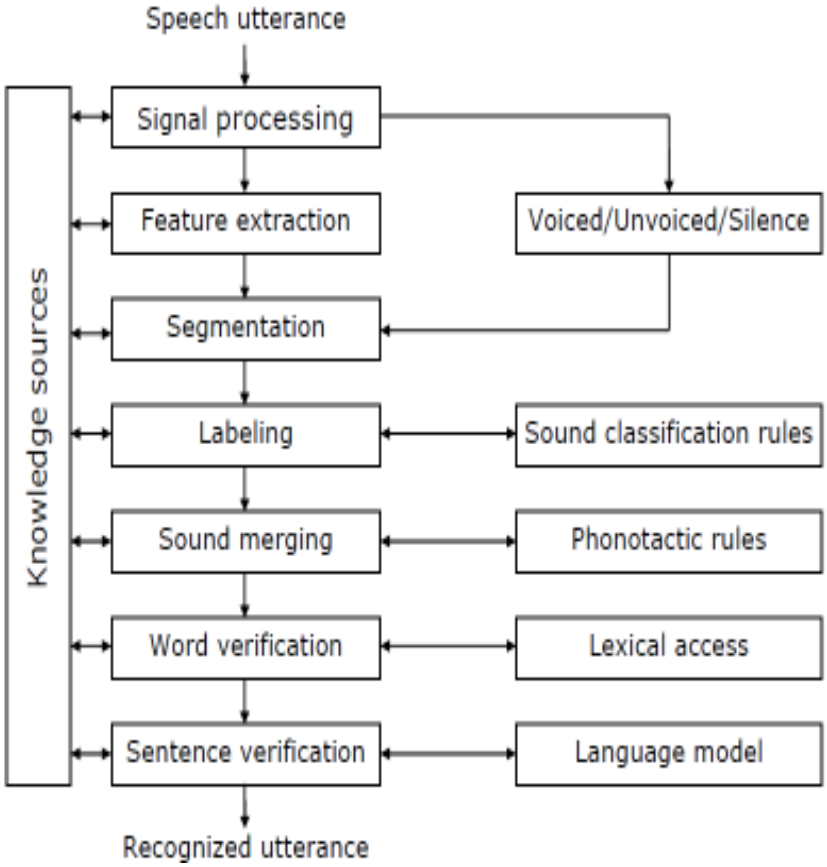

Figure 2-14. Block diagram of the bottom-up process in the artificial intelligence approach for automatic speech recognition (adapted from Rabiner and Juang, 1993).

Within the artificial intelligence approach class of recognition methods is the connectionist approach (Anusuya et al., 2009). The connectionist approach takes advantage of tools such as artificial neural networks (ANN) to learn the relationships between phonetic events. The ANN is a combination of a number of processing units called neurons that are trained by input-output datasets fed to the network. After training is done, the network can be tested with a similar dataset to recognize the pattern (Dede et al., 2010). The connectionist approach is the newest method used in automatic speech recognition and is the subject of controversies (Anusuya et al., 2009). In spite of these controversies, neural networks have performed better in a number of studies in comparison to other methods of speech recognition. For example, a study on isolated Malay digits has reported a recognition rate of 80.5% and 90.7% when dynamic time warping and Hidden Markov Model (HMM) techniques were used respectively (Al-Haddad et al., 2008),

whereas recognition rates of 98% (Azam et al., 2007), 99.5% (Alotaibi, 2005) and 98.125% to 100% (Dede et al., 2010) were reported in a similar application using neural networks.

A derivative of neural network that has recently been studied by many researchers for speech recognition and has shown progress in the field of acoustic modeling is the deep learning neural network (Yoshioka and Gales, 2015). Deep learning neural network is trained by back-propagation algorithm that allows better modeling of the non-linear data in combination with standard feed-forward neural network (Hinton et al., 2012; Deng et al., 2013). This method takes advantage of more hidden layers along with more nodes in each layer. More hidden layer means that each hidden unit is connected to many other hidden units below it hence deep learning consists of multiple layers of non-linear processing units. Deep learning neural network unlike the conventional neural network (that is used with supervised learning) can be trained in an unsupervised or supervised manner for both unsupervised and supervised learning tasks (Hinton and Salakhutdinov, 2006; Erhan et al., 2010). The combination of the above mentioned factors contributes to the success of this method. Deep learning neural networks have shown performance improvements especially in large vocabulary ASR tasks (Zhou et al., 2015).

2.4.1 Front-End Processing

Before the speech signal can be recognized, certain features have to be extracted from the signal to form the feature vectors of that signal (Anusuya and Katti, 2011). Because this stage precedes the speech recognition stage, it is referred to as front end processing. Using different feature sets can directly affect the performance of the ASR system, so appropriate selection of features is very important. Feature extraction methods are divided into two main categories: production-based methods (such as linear predictive coding – LPC) and perception-based methods (such as mel-frequency cepstral coefficients - MFCC - and perceptual linear prediction - PLP). In

production-based methods, a speech sample at any given time can be represented as a linear combination of number of previous samples (Anusuya and Katti, 2011), while in perception-based methods, the focus is to understand the way that the auditory system processes speech and try to use this information to collect features from speech that can be used in ASR (Trabelsi and Ban Ayed, 2012).

2.4.2 Pattern Recognition

Pattern recognition methods are used to map a sequence of feature vectors obtained from the feature extraction stage to the desired underlying speech signal. This mapping is done through a well formulated mathematical structure. One of the well-known methods for pattern recognition that is widely used in speech recognition is the Hidden Markov Model (HMM). Another approach for pattern recognition that has received attention in recent years is the artificial neural network (ANN) (Siniscalchi et al., 2013). Both of these powerful methods are used in this thesis and more details on them are provided in Chapter 3.

2.5 Related Work

There have been many studies done using ABR to investigate how the human auditory system processes speech signals in both the time and spectral domains (Kraus and Nicole, 2005; Burkard et al., 2006; Skoe and Kraus, 2010; Sadeghian et al., 2015). The studies included both transient and sustained responses that were discussed earlier in Section 2.3. For example, one of the studies concluded that gender makes a difference in how speech is processed in the auditory system. It was found that women on average have higher ABR peak amplitudes and shorter peak latencies in comparison to men (Burkard et al., 2006). A different study on transient responses concluded that to evaluate hearing loss in infants the best way is to monitor wave V of the ABR

to click stimuli (Picton and Durieux-Smith, 1978); however, a later study on children with learning problems and normal-learning abilities showed that while click evoked ABRs can be used to evaluate integrity of cochlea and the ascending auditory pathway, to gather information on the encoding process in the auditory system, speech-evoked ABRs need to be used (Song et al., 2006).

Other studies comparing transient and sustained responses in children have shown that these responses are independent of each other. One of these studies showed that an auditory stressor such as background noise negatively affects the transient response and degrades how it is encoded neurally; however, the sustained response remains almost unaltered (Russo et al., 2004). Based on this finding, this study concluded that in such a scenario (presence of background noise), the sustained response will be more useful for clinical and research applications because it would be expected to degrade differentially depending on the hearing condition of the subject under test. Another study reported that the amplitude of the response to the first formant of the /da/ stimulus was reduced in children with learning problems, while there was no difference in the activity in the range of the fundamental frequency of the stimulus /da/ between the children with learning problems and normal-learning abilities (Wible et al., 2004).

In this research, we investigate the automatic discrimination of speech-evoked ABRs to different vowels. Since the sustained part of these ABRs has many similarities to speech, with a fundamental frequency and a harmonic structure, we use standard automatic speech recognition techniques for this purpose, and we will not discuss the transient response any further. This part of the response corresponds to the stimulus fundamental frequency and its harmonics.

There are two main differences between our work and previous studies. The first difference is the purpose of this study. Our objective was to evaluate if speech-evoked ABRs to different

vowels could be discriminated with high accuracy using automatic speech recognition methods. The recognition result provides a quantitative measure for discriminating speech-evoked ABRs using the sustained response. One of the potential applications of this finding is in the field of hearing aid fitting. The outcome of this work could enable us to objectively adjust the hearing aids during the fitting process to maximize the recognition accuracy using ASR. This could result in optimal separation of the speech-evoked ABRs in a way that the brain of the hearing impaired person would readily discriminate between these signals. The second difference is in the stimuli used in this study. In the majority of the previous works, the stimuli used were either consonant-vowel syllables (Cunningham et al., 2001; Wible et al., 2005; Russo et al., 2004; Song et al., 2006; Wible et al., 2004) or tones and clicks (Schrode et al., 2014; Starr et al., 1996; Sininger, 1993), but in this study we used pure vowels as stimuli. The main reason for choosing vowels as stimuli in this study was that different vowels have distinct spectral features; and since our goal was to use automatic speech recognition methods for discriminating between speech-evoked ABRs, the vowels were chosen over other stimuli for an initial investigation of the proposed approach.

3 Methodology

3.1 Data Collection

The stimuli used for this work were five English vowels (/a/, /ae/, /ɔ/, /i/, /u/) generated using formant synthesis for a duration of 300ms with the fundamental frequency (F0) set to 100Hz for all vowels (Sadeghian et al., 2015). Considering the importance of dominant formants, only the first three formants (F1, F2 and F3) were used to generate the vowels, and their frequencies are shown in Table 3.1. Brainstem responses were collected from eight subjects (six males and two females between ages of 24 to 45) in a sound treated room with six trials for each vowel. In each trial, 500 repetitions of the stimulus were presented in alternating polarity at the repetition rate of 3.1/sec and responses were coherently averaged over the 500 repetitions. The reason for averaging is based on the assumption that the responses to stimuli are similar in all the repetitions, while the noise is random and uncorrelated with the responses (with zero mean). Therefore averaging over a number of trials will reduce the noise level and increase SNR. The responses were recorded using a BioMARK v.7.0.2 system (Biological Marker of Auditory Processing, Biologic Systems Corp., Mundelein, IL). Three gold-plated Grass electrodes were used to collect data. The recording electrode was placed at the vertex, the reference electrode on right earlobe and the ground electrode on the left earlobe. Prior to digitization, the responses were passed through a bandpass filter with high and low cut-off frequencies of 30 and 1000 Hz. Stimuli were presented with 16 bit resolution at 48KHz, and the speech-evoked ABRs were recorded at the sampling rate of 3202Hz (Sadeghian et al., 2015).

Vowels	F1 (Hz)	F2 (Hz)	F3 (Hz)
/a/	700	1220	2600
/ae/	660	1720	2410
/ɔ/	570	840	2410
/i/	270	2290	3010
/u/	300	870	2240

Table 3-1. Formant frequencies used to synthesize the 5 stimulus vowels.

In order to suppress the artefacts in the collected responses, several different techniques were used. The first technique was averaging responses of alternating stimulus polarities to eliminate the cochlear microphonic (CM) artefact since CMs from opposite polarities negate each other. The second technique was to use a plastic tube to connect earphones to the BioMARK stimulus generation system and also take advantage of foam insert earphones to reduce electromagnetic leakage from the BioMARK stimulus transducer to the recordings electrodes. The third technique was to present the stimuli at a rate so that the inter-presentation interval is different from multiple integers of the power line noise cycle. This ensures power line noise does not add coherently when multiple trials are averaged. Another technique to reduce noise was to provide subjects with a comfortable reclining seat to decrease body movements and artefacts due to muscle activity. The distance between any electronic devices in the room and the subjects were also increased to reduce electromagnetic artefacts. In addition, the responses that had higher than normal overall amplitudes (threshold = 23.8 μV) were rejected (Sadeghian et al., 2015). Figure 3.1 shows the time domain representation of the five synthetic vowels spoken by a male with fundamental period of T0=10 ms, corresponding to a fundamental frequency F0 = 100 Hz. This figure represents the first 100 ms of the signals that were used as stimuli in our work. The same signals are represented in the frequency domain up to 1000 Hz in Figure 3.2 below.

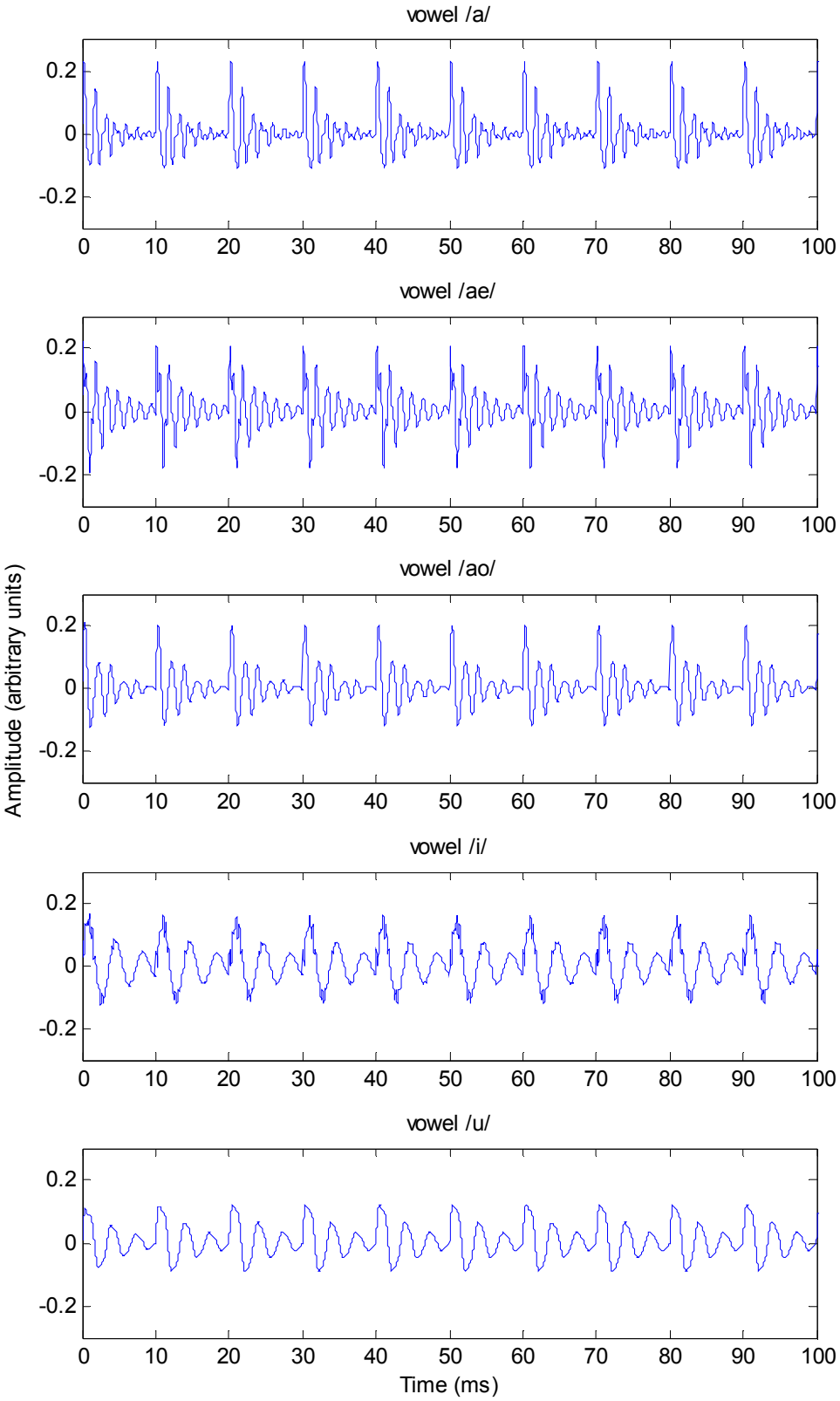

Figure 3-1. Time domain waveforms r (up to 100 ms) of the five synthetic English vowels as spoken by a male with fundamental period of T0=10ms. These signals are used as stimuli to generate speech-evoked ABRs that are studied in this work.

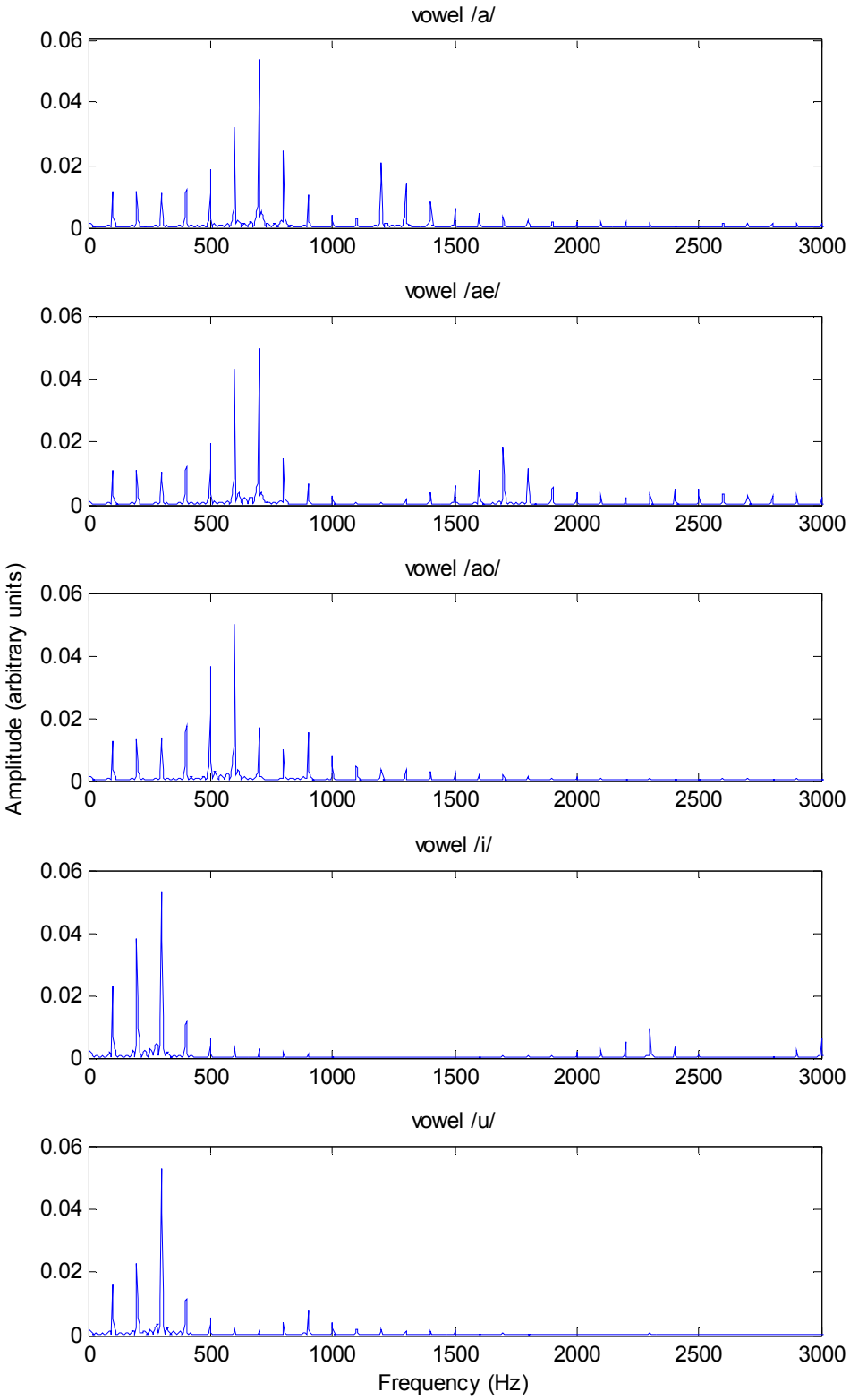

Figure 3-2. Amplitude spectra (up to 1000 Hz) of the five synthetic English vowels as spoken by a male with a fundamental frequency of F0=100Hz. These signals are used as stimuli to generate speech-evoked ABRs that are studied in this work.

Figures 3.3 and 3.4 are the time domain and frequency domain representations of the EFR signals averaged over all trials from all subjects (grand-average) in response to above mentioned stimuli, respectively, Figures 3.5 and 3.6 are the time domain and the frequency domain representations of the FFR signals averaged over all trials from all subjects (grand-average) in response to the same stimuli, respectively.

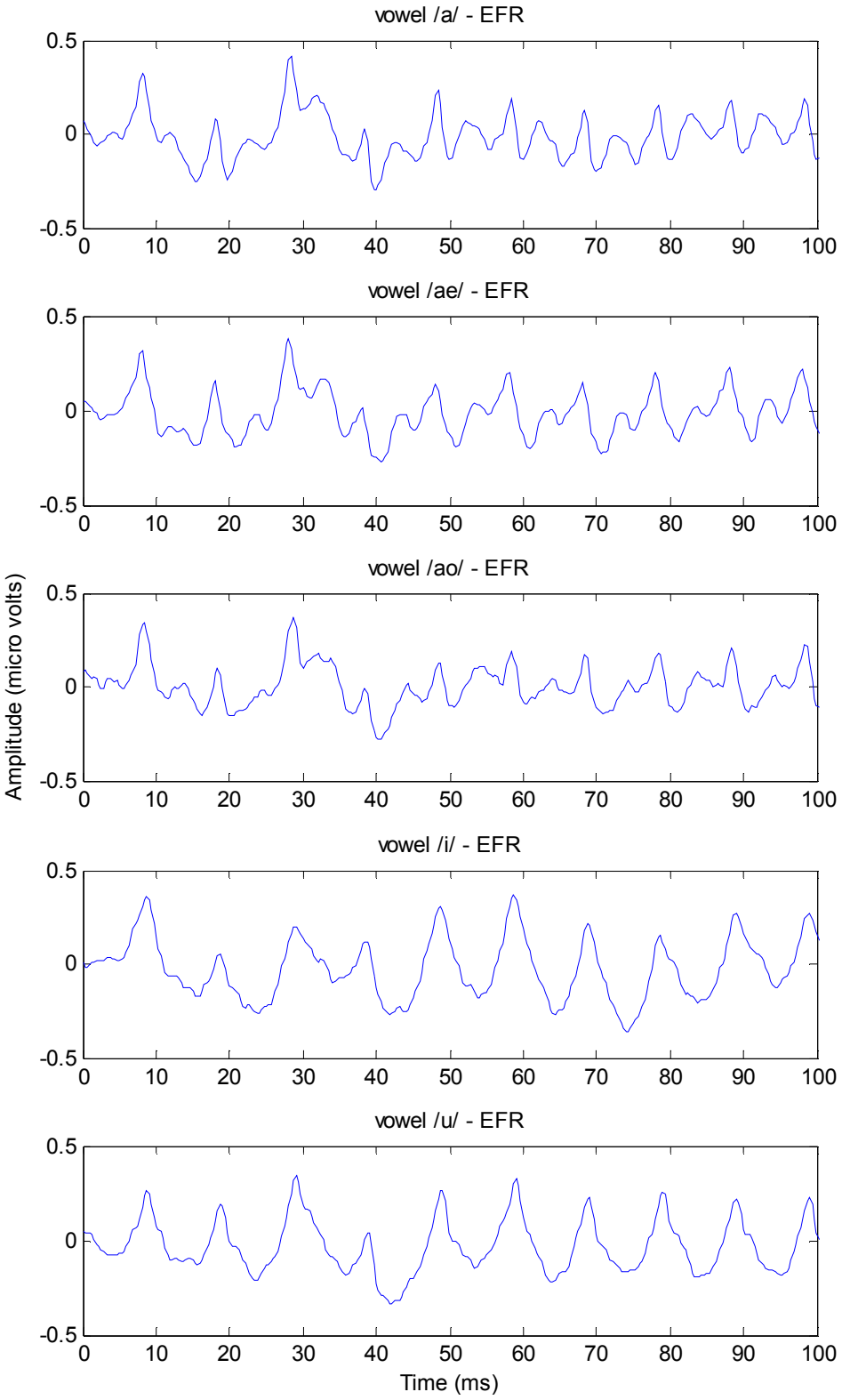

Figure 3-3. Time domain waveforms (first 100ms) of the Envelope Following Response (EFR) to five English vowels averaged over all trials and all subjects.

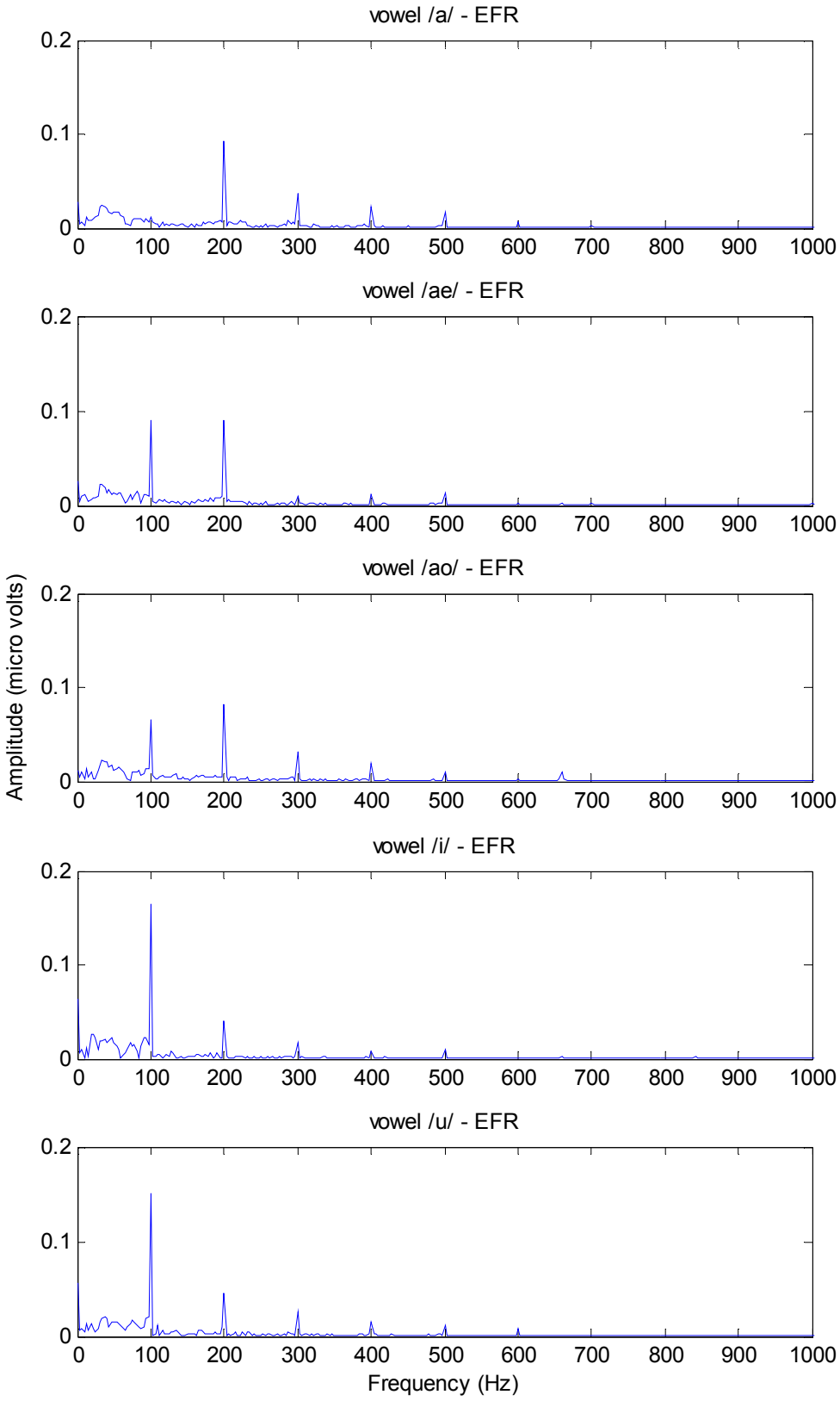

Figure 3-4. Amplitude spectra up to 1000Hz of the Envelope Following Response (EFR) to five English vowels averaged over all trials and all subjects.

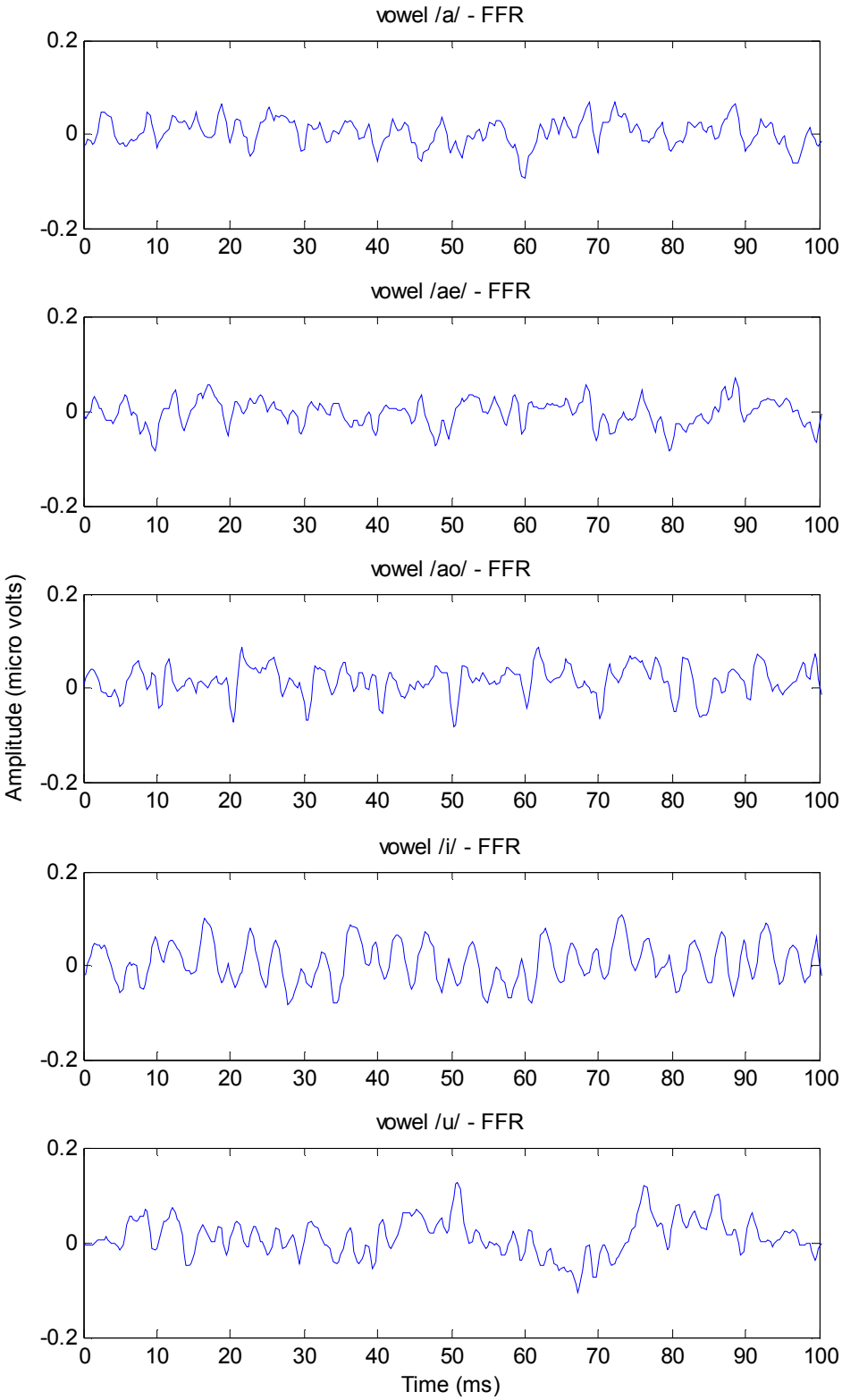

Figure 3-5. Time domain waveforms (first 100ms) of the Frequency Following Response (FFR) to five English vowels averaged over all trials and all subjects.

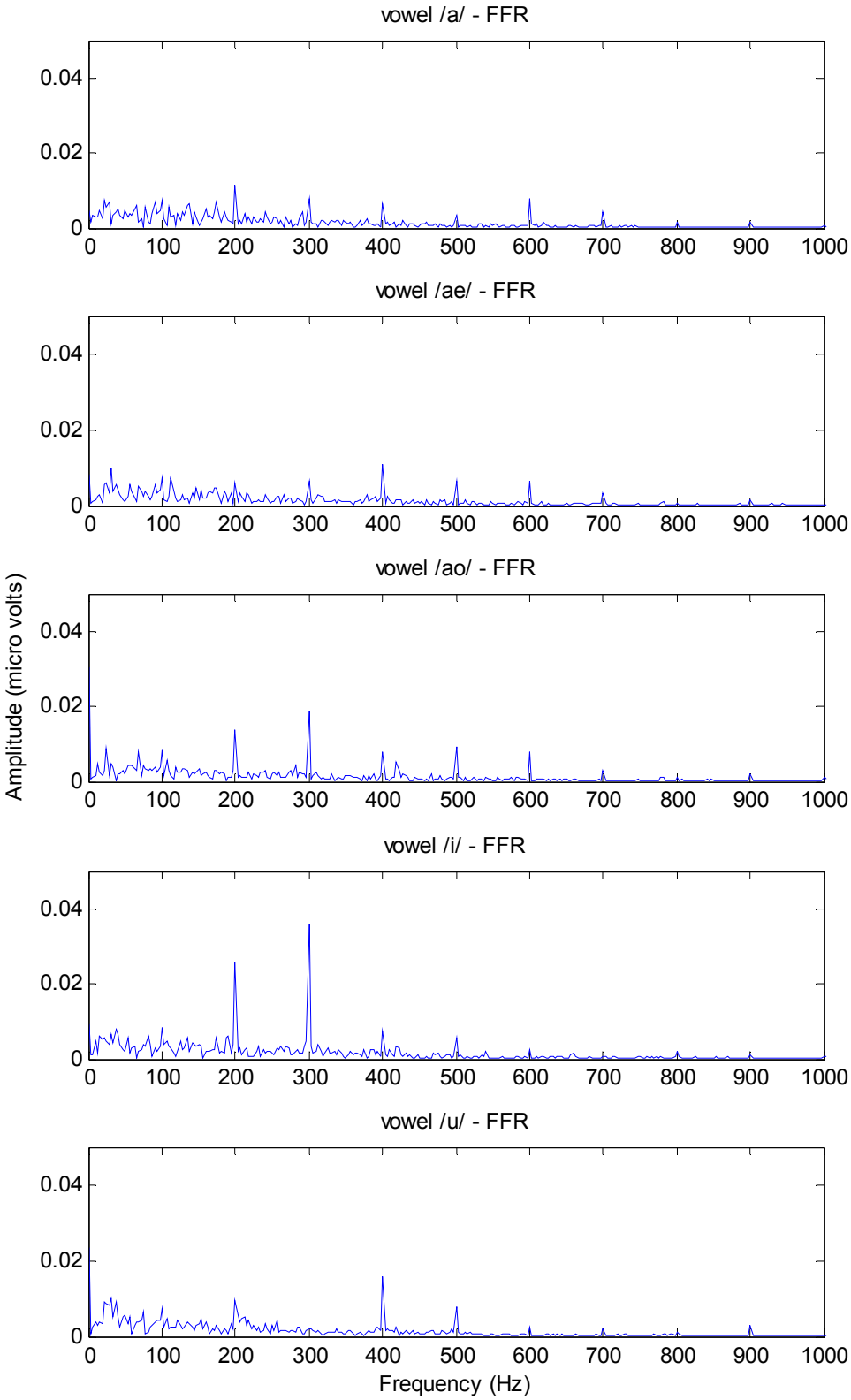

Figure 3-6. Amplitude spectra up to 1000Hz of the Frequency Following Response (FFR) to five English vowels averaged over all trials and all subjects.

3.2 Feature Extraction

The main idea of the front-end processing of an ASR system is to extract certain features of speech, which are then fed to the pattern recognition engine. The most popular and well-known feature extraction methods in ASR are linear predictive coding (LPC) and mel-frequency cepstral coefficients (MFCC). Another method that has been proven to work in noisy environments is perceptual linear prediction (PLP) (Namrata, 2013). In addition to LPC, MFCC and PLP, in this study, linear predictive cepstral coefficients (LPCC), power spectral analysis, principal component analysis (PCA), and independent component analysis (ICA) were used in the front-end processing for feature extraction. The basic concept behind transforming speech signal into vectors of parameters is illustrated in Figure 3.7 below.

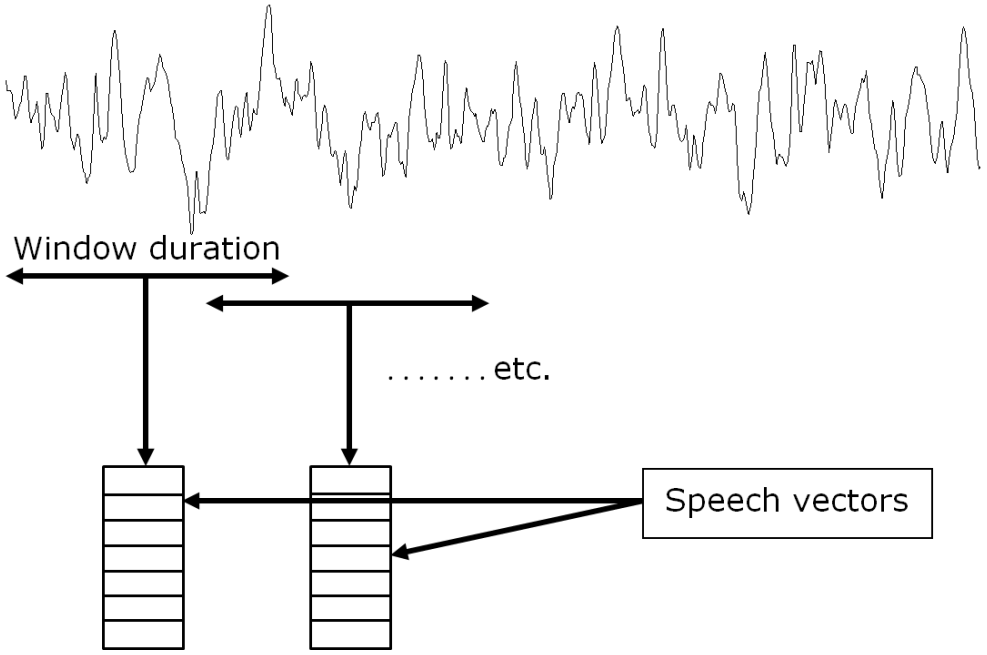

Figure 3-7. High level diagram of the feature extraction process (adapted with modification from Young et al., 2000).

As shown in the figure above, the speech signal is divided into a number of windows. Each window overlaps the previous window by a certain amount and outputs a parameter vector.

These parameter vectors then can be used as extracted features of the speech signal after going through specific analysis. LPC, MFCC, PLP and LPCC all use this windowing method to extract parameter vectors from the speech signal.

Some studies have shown that PCA and ICA, when used as part of feature extraction methods for ANN, result in improved performance for the recognition system (Mohamad-Saleh and Hoyle, 2008; Mendes et al., 2008). PCA and ICA are both statistical and computational methods used in feature extraction. While PCA de-correlates the data using an orthogonal transformation, ICA identifies statically independent components of non-Gaussian signals. ICA was applied on the raw waveform and PCA was applied both on the raw waveform and on the feature set obtained from the power spectral analysis.

In the case of the power spectral analysis, the amplitudes of the EFR and FFR spectra at the first 10 harmonics of F0 (i.e. at 100Hz, 200Hz,…, 1000Hz inclusive) were used as amplitude feature sets (Sadeghian et al., 2015).To perform the power spectral analysis, the discrete Fourier transform (DFT) was implemented using the fast Fourier transform (FFT) algorithm. Below we briefly discuss each of the feature extraction methods used in this study.

3.2.1 Linear Predictive Coding (LPC)

Linear predictive coding is one of the most powerful feature extraction methods for ASR. It models the human vocal track as an all pole system (Anusuya and Katti, 2011), but while it works well for voiced portions of the speech signal it does not perform as effectively for the unvoiced portions. The main idea behind LPC is that by using past speech signal samples one can predict the speech sample at the present time. This is done by finding a set of coefficients for a short section of speech signal that results in minimizing the mean square error between the original speech and the estimated speech. The resulting predictor coefficients are used as

parameters of the system function for speech production. LPC uses the following procedure to extract features from a given signal. First, the signal goes through a pre-emphasis stage where the speech signal goes through an FIR filter to flatten the spectrum of the signal. This step is usually done using the following transfer function,

$$H(z) = 1 - az^{-1} \qquad 0.9 < a < 1.0 \tag{3.1}$$

Next, the pre-emphasized speech is divided into blocks of N samples with an overlap between each two adjacent blocks. These blocks then go through a windowing process to minimize the discontinuities between the blocks. A typical window used in LPC is the Hamming window,

$$W(n) = 0.54 - 0.46\ \cos(\frac{2\pi n}{N-1}) \qquad 0 \le n \le N-1 \tag{3.2}$$

The result of windowing is,

$$X(n) = x(n)w(n) \qquad 0 \le n \le N-1 \tag{3.3}$$

The next step is to use autocorrelation to find fundamental frequency and any repeating pattern in the signal. The autocorrelation is done on each frame using,

$$n(m) = \sum_{k=0}^{N-1-m} X(k)\bar{X}(k+m) \qquad m = 0, 1, 2, \ldots, p \tag{3.4}$$

where p is the order of LPC and $\bar{X}$ represents the complex conjugate.

Using the LPC method for feature extraction, the vocal track is modeled as a digital all-pole filter (Bhattacharjee, 2013). The transfer function for the filter is,

$$V(z) = \frac{G}{1-\sum_{k=1}^{p} a_k z^{-k}} \tag{3.5}$$

where *G* is the gain of the filter and $\{a_k\}$ is a set of LPC coefficients.

The gain G of the filter can be calculated as,

$$G = \sqrt{n[0] - \sum_{k=1}^{p} a_k\, n[k]} \tag{3.6}$$

where p is the order of the all-pole filter. All p+1 frames of autocorrelations are converted to the LPC parameter set (Anusuya and Katti, 2011) using the Levinson-Durbin algorithm as described by Barnwell, (1980). The Levinson-Durbin algorithm solves the following equation by using the following set of equations,

$$\begin{bmatrix} n[0] & \cdots & n[p-1] \\ \vdots & \ddots & \vdots \\ n[p-1] & \cdots & n[0] \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} n[1] \\ \vdots \\ n[p] \end{bmatrix} \tag{3.7}$$

where n[p] is the autocorrelation function of a windowed speech signal.

$$E^{(0)} = n[0] \tag{3.8}$$

$$k_i = \frac{n[i] - \sum_{j=1}^{i-1} a_j^{(i-1)} n[i-j]}{E^{(i-1)}} \qquad 1 \leq i \leq p \tag{3.9}$$

$$a_i^{(i)} = k_i \tag{3.10}$$

$$a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)} \qquad 1 \leq j \leq i-1 \tag{3.11}$$

$$E^{(i)} = (1 - k_i^2)\, E^{(i-1)} \tag{3.12}$$

The above set of equation is solved recursively for $i = 1, 2, \ldots, p$. When the p^{th} iteration is achieved, a set of LPC coefficients is given by,

$$a_j = a_j^{(p)} \qquad 1 \leq j \leq p \tag{3.13}$$

And the filter gain is given by,

$$G = \sqrt{E^p} \tag{3.14}$$

While LPC provides a good model for the speech signal, it generates highly correlated features which are not very desirable in acoustic modeling (Shanthi and Chelpa, 2013). The results obtained from this method of feature extraction are discussed in detail in Chapter 4.

3.2.2 Linear Predictive Cepstral Coefficients (LPCC)

LPCC could be viewed as an extension to the LPC feature extraction (Anusuya and Katti, 2011). Cepstral analysis involves finding the cepstrum of a speech signal by taking an inverse Fourier transform of the logarithm of the spectrum of the signal. The equation below is the definition of signal cepstrum $\hat{s}[n]$,

$$\hat{s}[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln[S(\omega)]e^{j\omega n}d\omega \tag{3.15}$$

where $S(\omega)$ is the Fourier spectrum of a speech frame and $\hat{s}[n]$ is the cepstrum. If we represent the impulse response of the vocal track by $v[n]$ and the glottal excitation by $u[n]$ then the speech sequence in time domain and frequency domain can be modeled as,

$$s[n] = v[n] * u[n] \tag{3.16}$$

And,

$$S(\omega) = V(\omega)U(\omega) \tag{3.17}$$

Considering the logarithm inside the integral in the equation above, the cepstrum of the speech sequence can be presented as the summation of vocal track cepstrum and the glottal excitation cepstrum,

$$\hat{s}[n] = \hat{v}[n] + \hat{u}[n] \tag{3.18}$$

where $\hat{v}[n]$ is the vocal track cepstrum and $\hat{u}[n]$ is the glottal excitation cepstrum.

In the case of LPCC, a set of recursive operations is performed on LPC to generate the cepstral coefficients. This procedure is represented by the following equations,

$$\hat{v}[n] = \begin{cases} \ln(G) & for\ n = 0 \\ a_n + \sum_{k=1}^{n-1}\left(\frac{k}{n}\right)\hat{v}[k]a_{n-k} & for\ 1 \le n \le p \end{cases} \tag{3.19}$$

where $\hat{v}[n]$ is a set of LPCC for p number of coefficients.

This method can be compared to mel-frequency spectral coefficient (MFCC) since they are both based on cepstral analysis. However, some previous research in the field of speech recognition has shown that MFCC performs better in noisy environments compared to LPCC (Bhattacharjee, 2013). We have used LPCC to extract features that were used in our study and the results are provided in detail in Chapter 4.

3.2.3 Perceptual Linear Prediction (PLP)

PLP takes advantage of the psychoacoustic properties of the human auditory system (Anusuya and Katti, 2011). Although PLP also divides the speech signal into a number of windows (frames) with an overlap between each two adjacent windows, it is different from LPC and LPCC since it modifies the speech spectrum by a number of psychophysical transformations such as critical band spectral analysis, the equal loudness curve, and the intensity power law (Anusuya and Katti, 2011; Namrata, 2013). First, the signal has to be subdivided using a windowing techniques (same as previous methods) and the FFT is applied to obtain the spectrum of the signal. This spectrum then is processed further through following psychophysical transformations:

1) Critical band spectral analysis

$$X_m = \sum_{k=0}^{\frac{N}{2}-1} |S[k]|^2 |\psi_m[k]| \qquad 2 \leq m \leq M-1 \qquad (3.20)$$

where X_m is the filter output of the m^{th} filter, $|S[k]|^2$ is the N-point power spectrum of the windowed speech, and ψ_m is the filter weight.

$$\psi = \begin{cases} 0 & for \quad f_{Bark} - f_{c(Bark)} < -2.5 \\ 10^{(f_{Bark} - f_{c(Bark)} + 0.5)} & for \quad -2.5 \leq f_{Bark} - f_{c(Bark)} \leq -0.5 \\ 1 & for \quad -0.5 < f_{Bark} - f_{c(Bark)} < 0.5 \\ 10^{-2.5(f_{Bark} - f_{c(Bark)} - 0.5)} & for \quad 0.5 \leq f_{Bark} - f_{c(Bark)} \leq 1.3 \\ 0 & for \quad f_{Bark} - f_{c(Bark)} > 1.3 \end{cases} \tag{3.21}$$

where f_{Bark} is the frequency and $f_{c(Bark)}$ is the center frequency of one filter in Bark scale given by:

$$f_{Bark} = 6 \ln \left(\frac{f}{600} + \left(\left(\frac{f}{600} \right)^2 + 1 \right)^{0.5} \right) \tag{3.22}$$

where f is the frequency in Hz and f_{Bark} is the corresponding Bark frequency in Barks.

2) Equal loudness weighting

$$E = \frac{(\omega^2 + 56.8 \times 10^6)\omega^4}{(\omega^2 + 6.3 \times 10^6)^2 (\omega^2 + 0.38 \times 10^9)(\omega^6 + 9.58 \times 10^{26})} \tag{3.23}$$

where E is the equal loudness weight for each filter with ω angular frequency.

The output of the m^{th} filter $(X_{m(e)})$ after applying equal loudness weight is,

$$X_{m(e)} = E_m X_m \qquad 2 \leq m \leq M - 1 \tag{3.24}$$

Or,

$$X_{m(e)} = \sum_{k=0}^{\frac{N}{2}-1} |S[k]|^2 \left| \psi_{m(e)}[k] \right| \qquad 2 \leq m \leq M - 1 \tag{3.25}$$

where,

$$\psi_{m(e)}[k] = E_m \psi_m[k] \qquad 2 \leq m \leq M - 1$$

3) Intensity power law

The intensity is calculated using,

$$\varphi_m = \left(X_{m(e)}\right)^{0.33} \qquad 1 \leq m \leq M \tag{3.26}$$

where φ_m is the corresponding filter output after intensity-loudness comparison.

Using the result from the intensity calculation above, the entire spectrum can be expressed as the following vector,

$$\Phi = [\varphi_1 \;\; \varphi_2 \;\; \varphi_3 \;\; \dots\dots \;\; \varphi_{M-1} \;\; \varphi_M \;\; \varphi_{M+1} \;\; \dots\dots \varphi_3 \;\; \varphi_2] \tag{3.27}$$

To obtain the PLP coefficients, the IDFT is applied on the vector Φ above and the first p+1 coefficients are considered as the autocorrelation function R[n] for $0 \leq n \leq p$ where p is the order of PLP. The autocorrelation and cepstral analysis are used in the same manner as in LPC and LPCC to compute the PLP coefficients. Our test results for this method of feature extraction are provided in detail in Chapter 4.

3.2.4 Mel-frequency Cepstral Coefficients (MFCC)

When using feature extraction methods, one of the objectives is to generate features that have the least amount of correlation between the components and MFCC attempts to satisfy this objective (Anusuya and Katti, 2011). Since MFCC provides good discrimination and a low correlation between the components, this method is one of the more popular methods in feature extraction for automatic speech recognition (Anusuya and Katti, 2011). MFCC takes advantage of a Mel-frequency filterbank that is based on a non-linear frequency scale (mel-scale). The mel-scale is approximately linear below 1000 Hz and changes to logarithmic for frequencies above 1000 Hz (Bhattacharjee, 2013). The mathematical representation of the Mel scale based on a linear frequency scale is,

$$f_{Mel} = 1127.01 \ln\left(\frac{f}{700} + 1\right) \tag{3.28}$$

where f_{Mel} is the Mel-frequency in mels and f is the frequency in Hz.

This scaling follows human hearing scaling that has higher resolution (equivalent to narrower filters) for frequencies below approximately 1000 Hz, and lower resolution as the frequency increases.

The following equation can be used to evaluate the centre frequency of the m^{th} filter in filter bank:

$$f_{cm(Mel)} = f_{L(Mel)} + \frac{m(f_{H(Mel)} - f_{L(Mel)})}{M+1} \qquad 1 \leq m \leq M \qquad (3.29)$$

where $f_{c(Mel)}$ is the centre frequency and $f_{H(Mel)}$ and $f_{L(Mel)}$ are the upper and lower ends of the frequency range in mels, respectively. M is the number of filters between the upper and lower frequencies.

The filter output can be computed using,

$$X_m = \sum_{k=0}^{\frac{N}{2}-1} |S[k]|^2 |H_m[k]| \qquad 1 \leq m \leq M \qquad (3.30)$$

where X_m is the filter output of the m^{th} filter, $|H_m[k]|$ represents the frequency magnitude response for the k^{th} discrete frequency index and $|S[k]|^2$ is the N-point power spectrum of the windowed speech. The output of the filter is compressed using equation 3.31 to model the perceived loudness.

$$X_{m(ln)} = \ln X_m \qquad 1 \leq m \leq M \qquad (3.31)$$

where $X_{m(ln)}$ is the logarithmically compressed output of the m^{th} filter.

The result is then de-correlated using the Discrete Cosine Transform (DCT), and the first few coefficients are used as an MFCC feature vector. The following equation is the representation of the k^{th} MFCC coefficient for a feature vector of length p:

$$MFCC_k = \sqrt{\frac{2}{M}} \sum_{m=1}^{M} X_{m(ln)} \cos\left(\frac{\pi k(m-0.5)}{M}\right) \qquad 1 \leq k \leq p \qquad (3.32)$$

The results using MFCC features are discussed in detail in Chapter 4.

3.2.5 Spectral Analysis (Amplitude Feature Set)

One of the common methods in studying the speech signal is to use spectral analysis (Shrawankar and Thakare, 2010). To compute the spectrum, the discrete Fourier transform is performed on windows of the speech signal during which the signal is assumed to be quasi-stationary. The DFT converts the time domain signal into frequency domain using,

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{\frac{-j2\pi kn}{N}} \qquad (3.33)$$

and the resulting frequency spectrum contains signal information at different frequencies over time. Usually the result of such analysis contains the magnitude and phase information of the speech signal. However, previous studies have shown that phase information is not essential for speech recognition (Kwon and Lee, 2004). A very recent study by Sadeghian et al. (2015) has shown that this is also true when it comes to speech-evoked ABRs where the contribution of phase information was very minimal. In addition, delays in auditory processing at various frequencies may differ between individuals based on different ear canal lengths and different basilar membrane delays; therefore, the phase might reflect differences between subjects rather than differences between the vowels. As a result, the magnitudes of the power spectrum were used as feature vectors in our study and phase information was ignored. In this work, the extracted features obtained with this method are referred to as the amplitude feature set and the results obtained using this feature set are discussed in detail in Chapter 4.

3.2.6 Independent Component Analysis (ICA)

ICA transforms a non-Gaussian signal into statistically independent components (Xiang et al., 2014). The main reason behind generating feature vectors from a speech signal in ASR is to obtain a set of uncorrelated components that can be used in recognition process; therefore ICA could be considered as a good candidate since it provides coefficients that are statistically as independent as possible. The goal in using ICA is to obtain independent components s and a mixing matrix of A for a given data vector of x such that,

$$x = As \tag{3.34}$$

where x is an Nx1 column vector of a speech sample, A is an NxN mixing matrix with column vectors of basis functions, and s is an Nx1 column vector of the source signals.

At a high level description of the approach, to compute ICA components the following steps need to be followed:

- Finding w_j that maximizes non-Gaussianity of ${w_j}^T x$. This is done by centering and whitening the data (Hoyer and Hyvarinen, 2000).
- Computing independent components s using,

  $$s = Wx \tag{3.35}$$

  where $W = A^{-1} = [w_1 \ \ w_2 \ \ w_3 \ \dots \ w_m]$

Results from Kwon and Lee's (2004) study show that the above statement could be correct if there is a small amount of training data, but when it comes to a large training database, ICA does not perform as well as other feature extraction methods (Kwon and Lee, 2004). ICA is sensitive to phase change and provides different coefficients with different phase; however, the phase information is not essential for speech recognition as human perception is not sensitive to phase

change. In this study we used ICA to extract features from raw speech-evoked ABRs and the results from our experiment are discussed in Chapter 4.

3.2.7 Principal Component Analysis (PCA)

PCA is another method of data transformation similar to ICA with the difference that PCA de-correlates the data using an orthogonal transformation (Na et al., 2007; Cao et al., 2008). This results in finding the principal components of the dataset that account for the greatest amount of variance in the data, and that can be used as the new coefficients set. In addition to speech recognition, PCA has been applied in other domains such as face recognition and has provided good results. The disadvantage of the PCA technique is the size of the covariance matrix. The dimension of the covariance matrix is proportional to the size of the data set; therefore this method becomes computationally expensive for high dimensional data sets (Anusuya and Katti, 2011). The mathematical representation for PCA is as follows (Lima et al., 2004),

Assuming that the data contains M centered observations $x_k \epsilon R^n$, $k = 1, 2, \dots, M$ and $\sum_{k=1}^{M} x_k = 0$, the corresponding covariance matrix for this dataset can be represented as

$$C = \frac{1}{M} \sum_{j=1}^{M} x_j {x_j}^T \quad (3.36)$$

where C is the covariance matrix and x represents observations of the variables. The principal components then are obtained by solving,

$$\lambda v = Cv \quad (3.37)$$

where λ and v represent the eigenvalue and eigenvector respectively. Also there exist coefficients α_j such that,

$$v = \sum_{j=1}^{M} \alpha_j \, x_j \quad (3.38)$$

Therefore,

$$\lambda(x_k.v) = (x_k.Cv) \qquad k = 1, 2, \ldots, M \tag{3.39}$$

In our study, we used PCA on the raw speech-evoked ABRs as a feature extraction method and on the amplitude feature sets that were obtained from spectral analysis. In both cases, we also tried to eliminate features with low variance, which resulted in smaller resultant feature vectors. The results of our analysis are discussed in detail in Chapter 4.

3.3 Pattern Recognition

3.3.1 Hidden Markov Model (HMM)

HMM is a statistical model that offers a simple structure for the classification and recognition of signals with time-varying spectra. Almost all of today's continuous speech recognition systems are based on HMM (Gales and Young, 2007). HMM contains a number of states that are connected by a transition probability and an output probability. Depending on the speech rate, the transition between states could be a self-loop (slow rate), transition to the next state (normal rate), or transition skipping consecutive states (fast rate). HMM uses the Viterbi algorithm to compare the probability of the presented word to the probabilities of the other words in the dictionary (Li and Liu, 1999). The hidden Markov model toolkit (HTK) (Young et al, 2000) was used to perform all the front end processing as well as recognition on speech-evoked ABRs for HMM, as it has been used worldwide by many researchers (Elgarrai et al., 2014; Tripathy et al., 2013; Maqqor et al., 2014; Young et al., 1994; Woodland et al., 1994; Woodland et al., 1997; Nguyen et al., 2010). Young et al. (2000) describes the steps that are taken in the HTK toolkit for recognition, based on HMM including fundamental and mathematical concepts behind this method.

Our data consisted of ABRs to five vowels with 48 samples for each vowel. Due to our small dataset, we used the leave-one-out method to train and test. The training was done on all the subjects except the one that was set aside to test. This procedure was repeated until the all the sample sets (48 samples for 5 vowels) were covered and then the result was averaged over all the samples. Figure 3.8 below shows a block diagram of the recognition procedure using HMM as the pattern recognition engine.

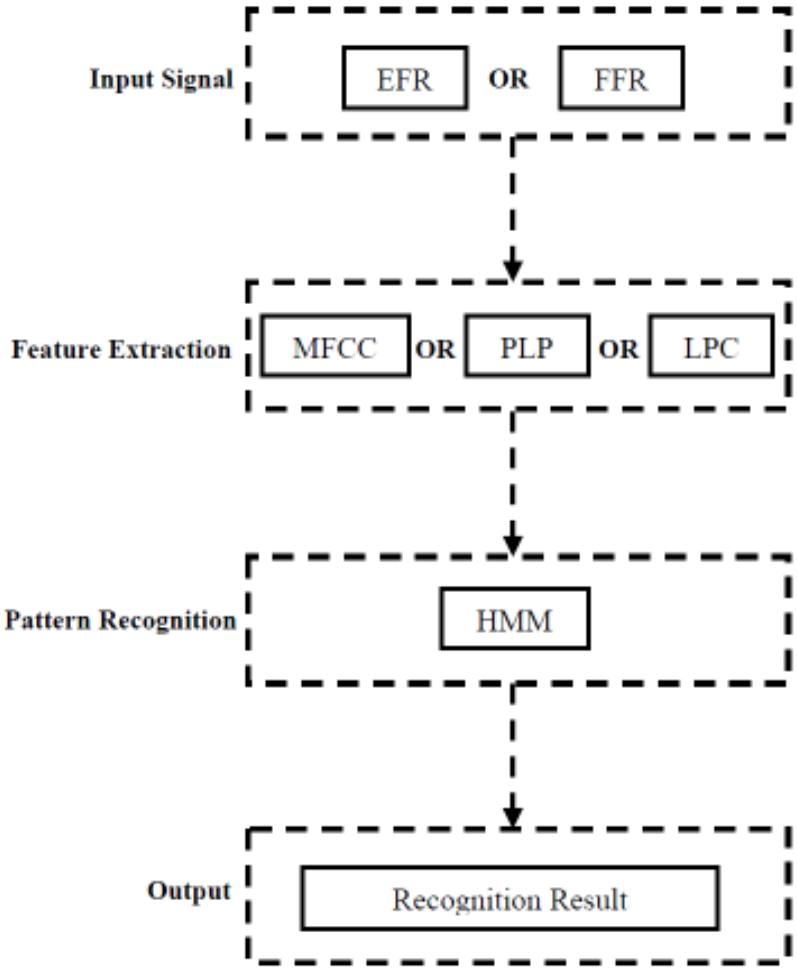

Figure 3-8. Block diagram of recognition steps using HMM

3.3.2 Artificial Neural Network (ANN)

The other pattern recognition method that was used in this study was the ANN, which is inspired by the function of the biological nervous system. The ANN contains many neurons (processing

elements) that work together to solve different types of pattern classification problems, and hence it could be a good fit for solving difficult speech recognition tasks (Lippmann, 1988). Hudson Beale et al., (2015) describes artificial neural networks in detail, including mathematical concepts and design architectures. We have taken advantage of the Neural Network Toolbox in Matlab to perform ANN pattern recognition and for front-end processing with ANN. Figure 3.9 below shows a block diagram of the recognition procedure using ANN as the pattern recognition engine.

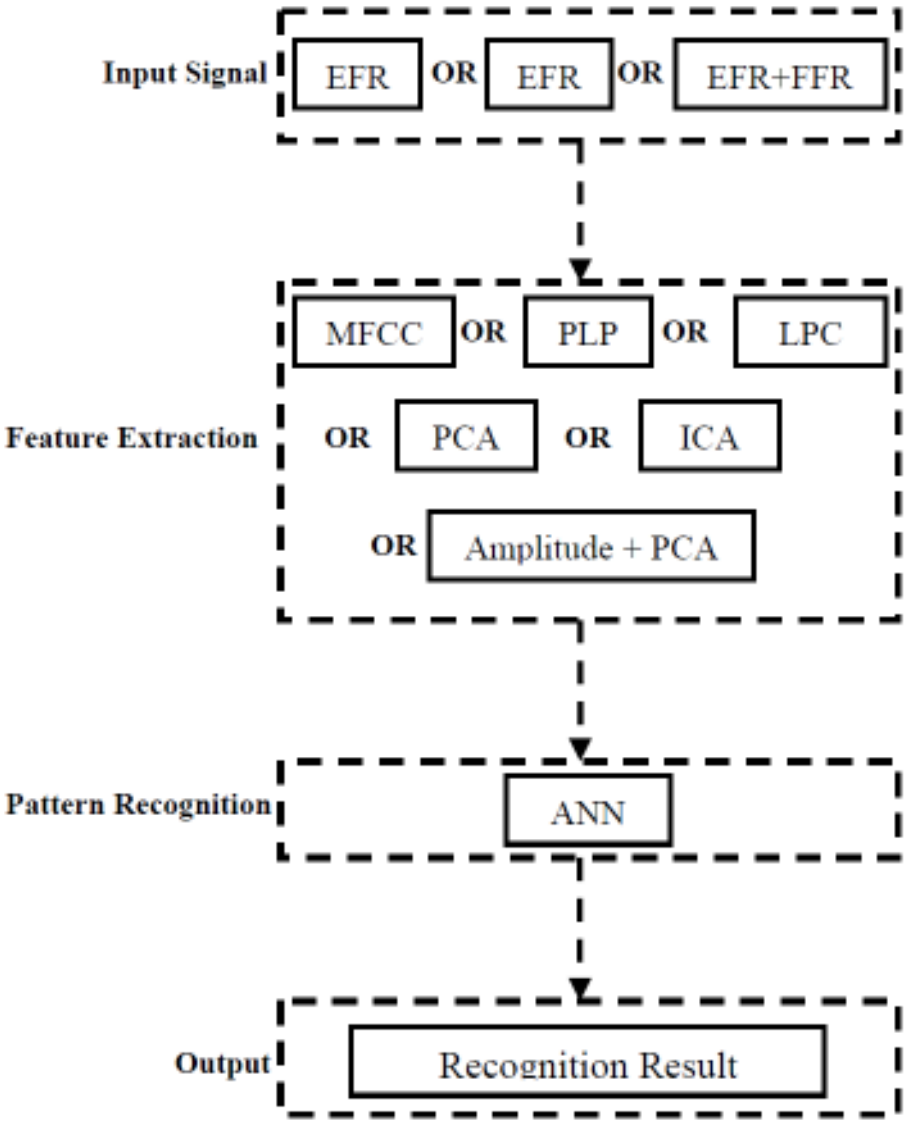

Figure 3-9. Block diagram of recognition steps using ANN

In the case of ANN, we used the same method for training and testing as the one for HMM (leave-one-out method) with the difference in that for the ANN, 15% of the training data were allocated for validation. The validation data was used to avoid over-fitting using the early stopping technique (Yao et al., 2007). The ANN used was based on a feed-forward network with one sigmoid hidden layer and one output layer. Different numbers of neurons were examined to determine the best structure. Since we were using the early stopping technique, it was essential to use as many neurons in the hidden layer as possible to avoid inaccurate local optima (Sarle, 1995). With this in mind, we based the optimization of the number of neurons in the hidden layer on the raw data. The raw data had the most number of features (input layer) and potentially required the most number of neurons (Blum, 1992). We started with a low number of neurons (one) and trained the neural network based on that, then using the validation step we acquired the validation error. Next we increased the number of neurons by one and computed the validation error. We repeated this procedure as long as the validation error decreased. When the validation error started increasing we stopped the process and chose the number of neurons that gave the lowest validation error. In our case, the best value among the different number of neurons for the hidden layer that were examined was 19. This gave us a mean square error (the error between the actual output and the ideal output) of 0.05 for EFR signals and 0.11 for FFR signals. The network was trained with the scaled conjugate gradient back propagation algorithm, and the structure of the ANN was fixed prior to applying the test dataset. Chapter 4 shows in detail the results that were achieved by using HMM and ANN as pattern recognition engines in combination with different feature extraction methods.

4 Results

Different combinations of feature extraction and pattern recognition approaches from Chapter 3 were used for recognition of the speech-evoked ABR of the five English vowels. The ASR process was done on EFR as well as FFR data; however, the accuracy with FFR data was substantially lower in comparison to the accuracy obtained with the EFR regardless of the feature extraction or pattern recognition method that was used. In this chapter, we discuss the set-up for each method and the obtained result in more detail. One point to keep in mind for the recognition performance is that since there are five vowels, chance level is 20%.

4.1 HMM with LPC

In using LPC in front end processing for HMM, we adjusted a few parameters in order to achieve the best recognition result. These optimizations were done on the training sets and then used for the testing data.

One of the adjusted parameters was the order of LPC. The sampling frequency in kHz is a good place to start for estimating the order of LPC and adjusting the order by trial and error (Vallabha et al., 2004). In our case, we tried values between 3 and 20, but the recognition results for all the trials were very similar. The value that was chosen for LPC filter order was 12 since this was the default value used by HTK and we did not observe any noticeable improvement or degradation by modifying this value.

The other elements that could make a difference for the end result are the duration of the windows over which the coefficients are calculated and also the amount of overlap between two subsequent windows. Although modifying these two variables (window length and overlap)

showed improvements with some of the collected data, there were degradations with others. Overall, the changes in window length and overlap did not provide any consistent change in the results based on our dataset for both EFR and FFR data. The obtained recognition result in the case of LPC with HMM on EFR signals was 32.1% with window length of 25 ms and an overlap of 10 ms.

We repeated the same procedure for FFR signals. There was also no noticeable difference in the recognition results using different numbers of LPC coefficients ranging from 10 to 20. We chose 12 LPC coefficients to be in line with the EFR test. The recognition result for FFR signals using a window length 25 ms and an overlap of 10 ms was 20.4%.

4.2 HMM with LPCC

When LPCC is used as a feature extraction method for speech recognition, the number of cepstral coefficients should be considered in addition to filter order. We set the filter order to 14 in our experiments, which is the default value set by HTK. In many of previous works in speech recognition the value of 12 has been used for the number of coefficients; however, some studies recommend using 20 coefficients for better recognition results. Using our data, we tried different coefficient numbers ranging from 10 to 20, and the recognition result remained the very similar for all the tests. Based on these results, we chose 12 coefficients which is in line with most of the previous works.

We also tried different window lengths and overlap sizes for two adjacent windows in order to optimize the recognition result. However, as was mentioned in Section 4.1, these did not produce consistent changes in the result. A recognition accuracy of 46.3% was achieved using a window length of 25 ms and an overlap of 10ms between two subsequent windows for EFR signals.

The same method was used in computing recognition results for FFR signals. With the same filter order and number of cepstral coefficients as used in the EFR test, the best recognition result of 37.5% was achieved with the window length of 25 ms and an overlap of 10 ms. All the optimisations were done on training data for both EFR and FFR signals and then the obtained values of the parameters were used with the test data.

4.3 HMM with MFCC

The next method used was the combination of MFCC and HMM. Different numbers of coefficients were generated along with different window and overlap lengths to achieve the highest possible recognition result for this method. With this method, as with the other two previously described methods, modification in the length of windows and the overlap between them did not produce consistent improved results and therefore we used window length of 25 ms and overlap of 10 ms to maintain consistency with the other methods. However, the number of coefficients had a reliable effect on recognition rate. All the optimisations were done on training data for both EFR and FFR signals and were then used with the test data. Figure 4.1 and Figure 4.2 show the recognition results with the different number of coefficients for EFR and FFR signals, respectively. For EFR signals, the best recognition result of 48.3% was achieved with 12 coefficients, while in the case of FFR signals, the best result of 31.7% was also obtained with 12 coefficients.

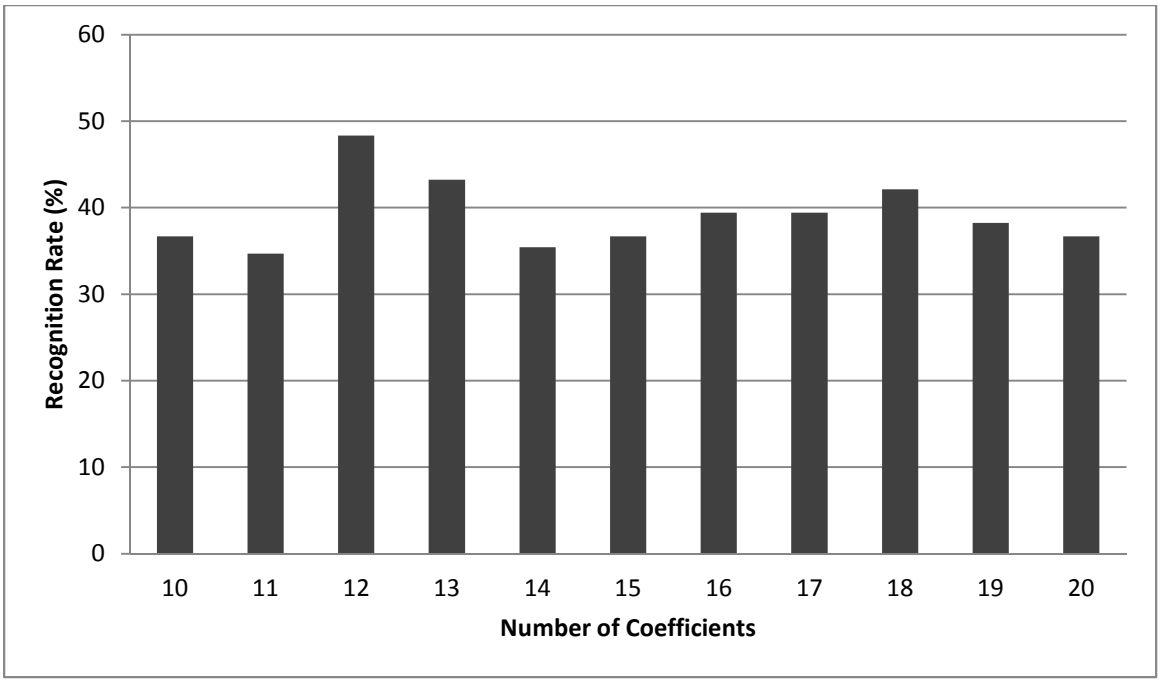

Figure 4-1. EFR recognition results using various numbers of MFCC coefficients with window length of 25 ms and overlap of 10 ms.

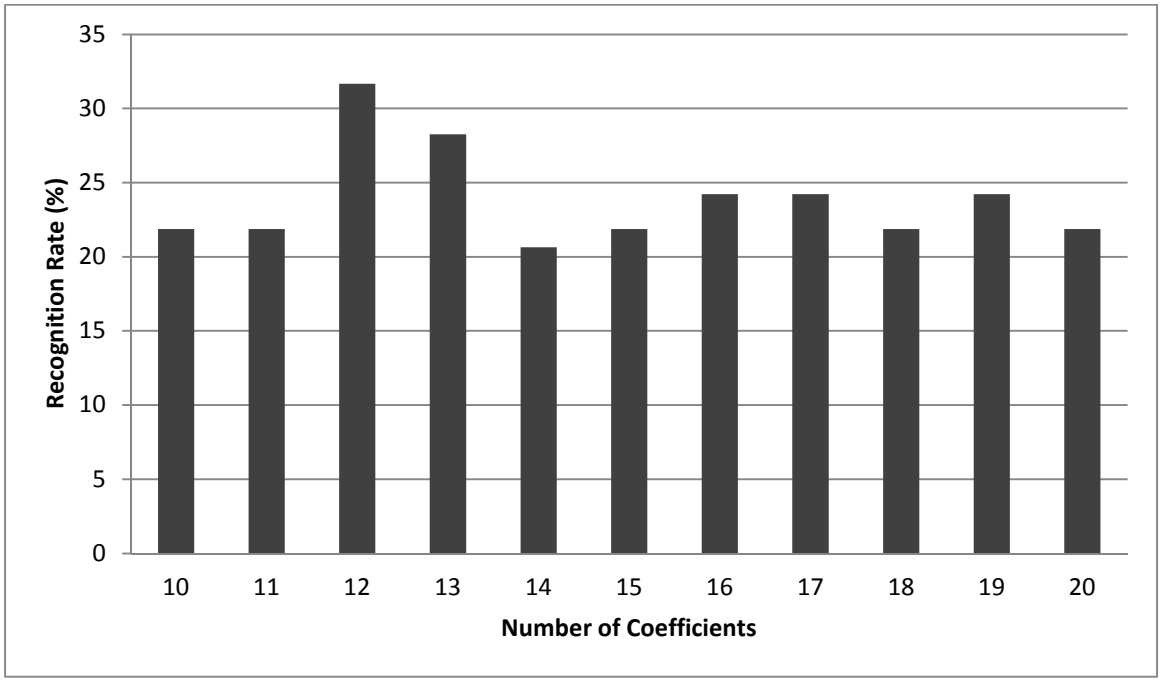

Figure 4-2. FFR recognition results using various number of MFCC coefficients with window length of 25 ms and overlap of 10 ms.

4.4 HMM with PLP

The last feature extraction method used with HMM was PLP. In this case, we also tried to find highest recognition result by modifying window length of the input signal and the size of overlap between two consecutive windows. However, as mentioned with the other methods, these modifications did not produce any reliable improvements, and therefore for consistency with those methods, a window length of 25 ms and overlap of 10 ms were chosen. The recognition result achieved using PLP and HMM with EFR signals was 49.6%. The same steps were followed to find the highest recognition result with FFR signals. With these signals, we also used a window length of 25 ms and an overlap of 10 ms and obtained a 35.8% recognition rate.

4.5 ANN with raw data

In addition to using HMM for pattern recognition, we investigated the use of ANNs for this purpose. Using the same EFR dataset on ANN without any front end processing provided us with a recognition accuracy of 50.8%, which was higher than all the results that were obtained using HMM. The improvement in recognition rate was also observed with FFR signals. Using ANNs on FFR signals with no front end processing provided a recognition rate of 40% which was also higher than what was obtained with any of the previously described methods used with this signal. This result motivated us to look into this recognition method further and try to combine ANN with different front end processing approaches to achieve even higher recognition rate.

4.6 ANN with PCA on raw data

As a next step, we considered re-mapping our data and choosing smaller and more relevant feature sets for recognition. PCA was used to remap the EFR dataset and eliminate components

with lower variance (information). When PCA was performed with our raw dataset of 1024 samples, for both EFR and FFR, 160 samples accounted for 99% of the variance. Based on this information we kept these 160 samples as our new feature sets for each trial. Recognition accuracy obtained using PCA on raw data as mentioned above with ANN was 33.3% for EFR signals and 27.1% for FFR signals.

4.7 ANN with ICA on raw data

Another remapping method that was used for recognition of the speech-evoked ABR signals was ICA. Our raw data contains 1024 samples for each trial, and performing ICA on this dataset for both EFR and FFR signals showed that 120 samples of the remapped data accounted for 99% of the variance. These 120 samples of each trial were considered as the new feature sets. The recognition accuracy obtained using this method was 46.7% for EFR and 39.2% for FFR signals.

4.8 ANN with LPC

Since the results from remapping techniques did not provide any improvement over just using the raw data, we decided to use some of the same feature extraction methods that were used in conjunction with HMM. One of the methods used for this purpose was LPC. We used a filter order of 12 for this test on both EFR and FFR signals. Window length and overlap were also matched to the ones used for HMM (window length of 25 ms with an overlap of 10ms). The recognition accuracy achieved with this method was 31.7% for EFR and 20.4% for FFR signals.

4.9 ANN with MFCC

Another feature extraction method used with ANN was MFCC since it is the most common method used for feature extraction in ASR applications. With MFCC, we based our test on both

EFR and FFR signals, and we used the same window length and overlap sizes as we used previously in the case of MFCC and HMM (i.e. 25 ms of window length and 10 ms of overlap). We also kept the extracted number of cepstral coefficients the same at 12. The recognition accuracy obtained using this method was 30.8% for EFR and 27.5% for FFR signals.

Table 4.1 summarizes all the results from all the different recognition methods that have been described so far.

	Feature Extraction	Pattern Recognition	Recognition Accuracy For EFR Signals	Recognition Accuracy For FFR Signals
1	No feature extraction	ANN	50.8%	40.0%
2	PLP	HMM	49.6%	35.8%
3	MFCC	HMM	48.3%	31.7%
4	ICA (on raw data)	ANN	46.7%	39.2%
5	LPCC	HMM	46.3%	37.5%
6	PCA (on raw data)	ANN	33.3%	27.1%
7	LPC	HMM	32.1%	20.4%
8	LPC	ANN	31.7%	20.4%
9	MFCC	ANN	30.8%	27.5%

Table 4-1. Recognition accuracy for EFR and FFR signals with various combinations of feature sets and recognition engines.

4.10ANN with Amplitude Feature Set

We also used the amplitude feature set in combination with ANN for pattern recognition. Initially, we considered the first seven harmonics of F0 (i.e. at 100Hz, 200Hz,…,700Hz inclusively) from EFR spectra as features. The recognition result with this subset was 64.3%. We then repeated the same test with first 10 harmonics of F0 and extended the feature set to first 15 harmonics. The recognition result with the first 10 harmonics was 75.8% and with the first 15 harmonics was 67.1%. We also tried the test with different numbers of harmonics between 10 and 15, but the best results were achieved using the first 10 harmonics. We repeated the same procedure for FFR spectra and the same pattern of behavior was observed with the first 10 harmonics providing the best recognition result of 63.7%. We also repeated the same test using a

combination of 20 features (10 from EFR and 10 from FFR) and the recognition result in this case increased to 80.4%. Table 4.2 shows the recognition accuracy on the test dataset with the amplitude feature sets corresponding to EFR, FFR and EFR+FFR (i.e. combination of 10 features from each of EFR and FFR), and using the ANN for pattern recognition.

Data	Recognition Accuracy
EFR+FFR	80.4%
EFR	75.8%
FFR	63.7%

Table 4-2. Recognition accuracy with the amplitude features of EFR+FFR, EFR and FFR, with ANN as the recognition engine.

As a next step, we tried to remap these features using PCA. Obtaining 10 features using PCA on the first 10 harmonics of F0 (no feature elimination) degraded the results for the EFR, FFR and EFR+FFR feature sets. However, eliminating one component that accounted for the lowest variance improved the recognition results for EFR, while it lowered the result for FFR. We continued eliminating more features based on the variance, but the result degraded even further. The best recognition result was achieved by eliminating only one feature from the new PCA feature sets in the case of EFR signals. For FFR signals, the new PCA feature set had lower recognition rate compared to the amplitude feature set without PCA, but within the PCA features, eliminating two features gave the best result. Figure 4.3 and Figure 4.4 show the recognition results obtained for EFR and FFR amplitude feature sets respectively after PCA was applied and one feature was eliminated at a time.

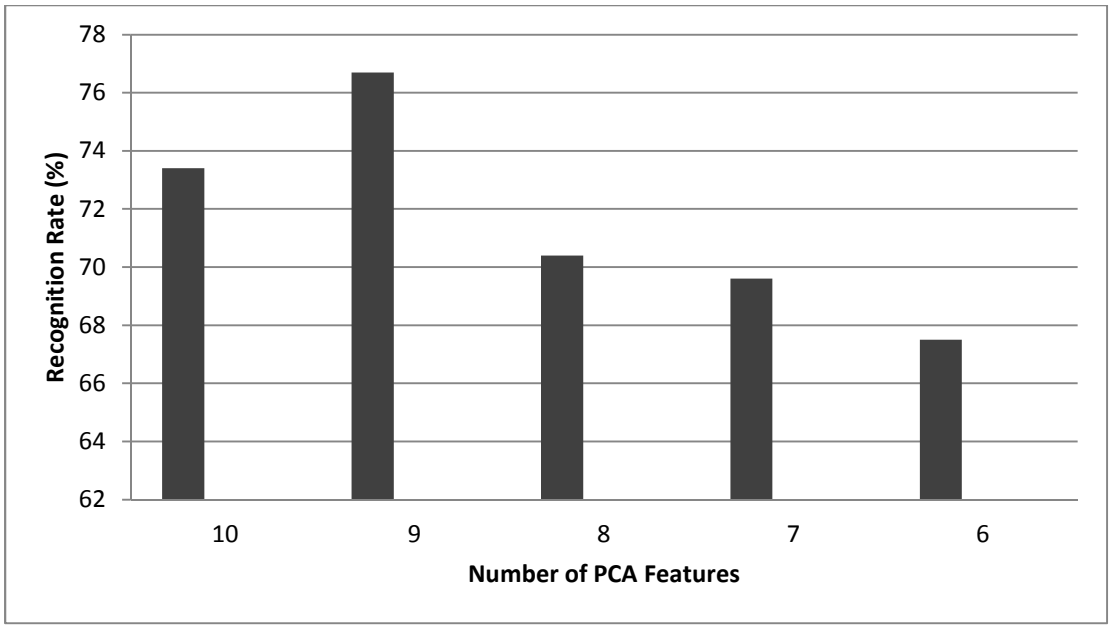

Figure 4-3. Recognition results using PCA on EFR amplitude feature set, and eliminating one feature at a time based on the variance.

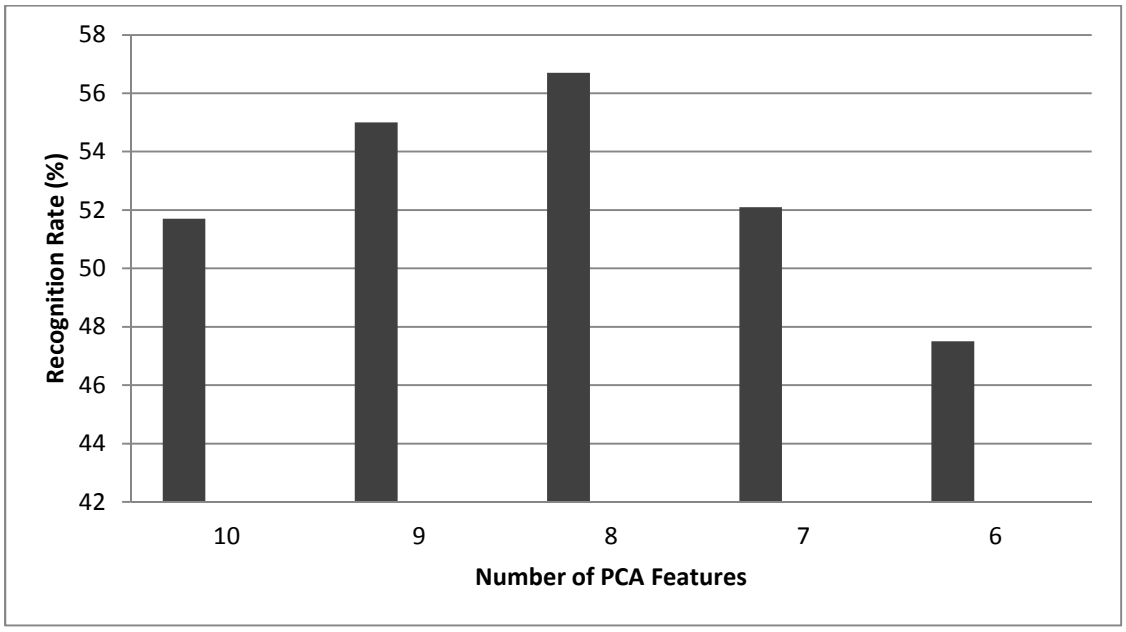

Figure 4-4. Recognition results using PCA on FFR amplitude feature set, and eliminating one feature at a time based on the variance.

In the case of the harmonic amplitude feature set from EFR+FFR, we followed two different approaches. One approach was to add the two feature sets then perform PCA on the new set (20 features) and eliminate features that accounted for the lower variance. Figure 4.5 shows the

recognition results obtained by applying PCA to the combination of EFR and FFR amplitude feature set and then eliminating features with lower variance one at a time.

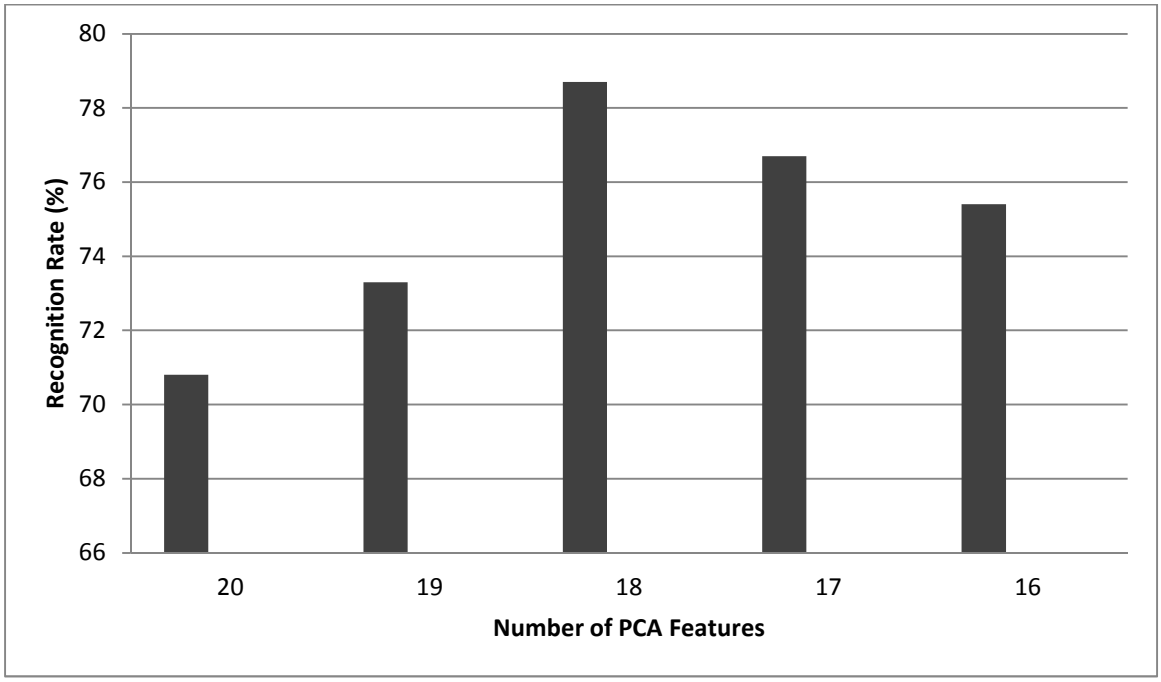

Figure 4-5. Recognition results using PCA on the combination of features from EFR+FFR feature sets, and eliminating one feature at a time based on the variance.

The other approach was to remap the EFR and FFR features using PCA separately, eliminate the features with lower variance in each set, and then combine the remaining features to construct a new feature set for ANN. Figure 4.6 illustrates the results obtained from elimination of features (one at a time) for EFR and FFR amplitude feature sets separately based on the variance that the feature accounted for, and then combining them together to construct the new feature set.

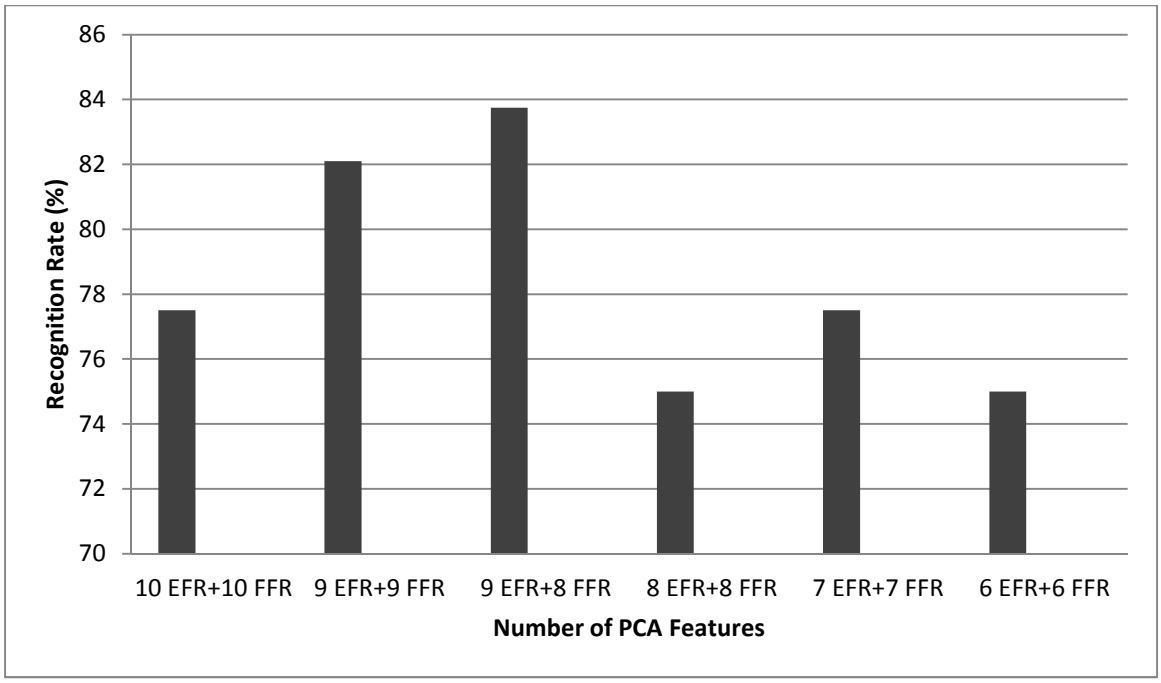

Figure 4-6. Recognition results using PCA on combination of features from EFR and FFR feature sets separately. One feature is eliminated at a time from EFR and FFR based on variance and the features are combined after the eliminations.

While the first approach did not produce any improvement in the accuracy, the second method resulted in improvements. The best recognition result was obtained by eliminating one feature in EFR and two features in FFR with the lowest variance in each set, and using the resulting 17 features (combination of 9 features from EFR and 8 features from FFR) as the new feature set. Table 4.3 shows the recognition results on the test dataset after applying the PCA to the amplitude feature sets corresponding to the EFR, FFR and EFR+FFR using ANN for pattern recognition.

Data	Recognition Accuracy
EFR+FFR (combine features, perform PCA and then eliminate)	78.7%
EFR+FFR (perform PCA, eliminate features and then combine)	83.8%
EFR	76.7%
FFR	56.7%

Table 4-3. Best recognition accuracy with amplitude features of EFR+FFR, EFR and FFR using PCA applied on the amplitude feature set and after retaining 9 components for the EFR, 8 components for the FFR, and 18 and 17 components for the EFR+FFR. The 18 components for EFR+FFR (row 1) were obtained by combining 10 features from EFR and 10 features from FFR, performing PCA on the complete 20 components and eliminating two features that had the lowest variance . The 17 components for EFR+FFR (row 2) were obtained by performing PCA on EFR and FFR separately, eliminating one feature for EFR and 2 features for FFR with lowest variance and adding the two sets. For EFR (row 3), PCA was performed on the amplitude feature set of EFR signals and one feature was eliminated based on the variance. For FFR (row 4), PCA was performed on the amplitude feature set of FFR signals and two features were eliminated based on their contributions to the variance.

Table 4.4 provides more details on the recognition results from the above mentioned methods (amplitude feature set with PCA applied and number of features eliminated based on their contributions to the variance in combination with ANN) by presenting confusion matrices for three feature sets. The confusion matrix named EFR+FFR (4.4(a)) corresponds to 17 features (combination of 9 features from EFR and 8 features from FFR amplitude feature set after applying PCA on each sets separately). The confusion matrix named EFR (4.4(b)) corresponds to 9 features obtained from PCA being applied to EFR amplitude feature set and eliminating one feature based on the variance. The confusion matrix named FFR (4.4(c)) is the result of applying PCA to FFR amplitude feature set and eliminating two features based on the variance. Each entry in the confusion matrix represents the percentage of recognition for the speech-evoked ABR corresponding to a vowel in relation to a particular vowel. For example, the entry in row 1 column 2 of Table 4.4(a) shows a value of 2% which means that 2% of the test data for vowel /ae/ were recognized incorrectly as /a/. At the same time, the entry in row 2 column 2 of the same

table shows a value of 78%, which indicates that 78% of the test data for vowel /ae/ were correctly recognized as /ae/.

EFR+FFR		Actual Vowels				
		/a/	/ae/	/ɔ/	/i/	/u/
Recognised Vowels	/a/	84%	2 %	4 %	0%	4 %
	/ae/	8%	78 %	6%	0%	2%
	/ɔ/	4%	8 %	78%	2%	6%
	/i/	0%	0%	2%	96%	2%
	/u/	4 %	12%	10%	2 %	86%

(a)

EFR		Actual Vowels				
		/a/	/ae/	/ɔ/	/i/	/u/
Recognised Vowels	/a/	82%	6%	17%	0%	0%
	/ae/	8%	69%	27%	4%	2%
	/ɔ/	10%	17%	52%	0%	6%
	/i/	0%	4%	0%	92%	2%
	/u/	0%	4%	4%	4%	90%

(b)

FFR		Actual Vowels				
		/a/	/ae/	/ɔ/	/i/	/u/
Recognised Vowels	/a/	63%	31%	8%	2%	8%
	/ae/	19	35%	15%	2%	12%
	/ɔ/	0%	19%	44%	0%	4%
	/i/	0%	2%	19%	79%	13%
	/u/	18%	13%	14%	17%	63%

(c)

Table 4-4. Confusion matrices for a) EFR+FFR, b) EFR, and c) FFR obtained by applying PCA on the amplitude feature sets (corresponding to the three bottom rows of Table 4.3).

5 Discussion

The results of this study show that pattern recognition methods can be used to classify speech-evoked ABRs with reasonably high accuracy. The results from standard speech recognition methods (shown in Table 4.1) generally provided accuracy above the chance level of 20% (100% / 5 vowels). The types of signals used in our test could be the main reason behind the limited performance of the HMM. Initially, when we tested HMM at the beginning of our work on speech signals (not the speech-evoked ABRs) to validate the use of this method and also confirm our set-up, we achieved a recognition rate of 97.3%. However, changing the data from speech to speech-evoked ABRs caused substantial degradation in recognition rate using HMM. The power of HMM in recognition comes from the way it deals with transitions between states as the speech signal evolves over time. In our case, the signals used are the steady-state responses to synthetic vowels, hence there is no transition between the states where the HMM can be most useful. Even with these steady-state responses, HMM sometimes still provided results well above the chance level.

On the other hand, the use of feature sets based only on the amplitude of the first 10 harmonics of F0, along with ANN as the recognition engine led to relatively high recognition accuracies compared to the best case scenario with a standard speech recognition method (PLP+HMM) (Tables 4.1, 4.2, and 4.3). We used ANN as a substitute for HMM since our preliminary results on the use of ANN with raw data (without any feature extraction) provided higher recognition rate in comparison to our results from any previously-tried HMM-based methods. The initial results were in line with the recent resurgence in the use of neural network in automatic speech recognition that shows it can outperform HMM recognition based methods for short time-sequences of speech signals (Trentin and Gori, 2001; Alotaibi, 2012). The best

results were obtained when the EFR and FFR features were combined and when PCA was applied to select the best features. The main reason for the superior performance of the amplitude features is probably because the useful information in the responses to synthetic vowels is concentrated at the harmonic frequencies, whereas other frequency bins contain only noise. It is therefore possible that other information at the harmonic frequencies, such as absolute or relative phase, and bispectral features, could further improve the accuracy of recognition. Additional improvement may be possible if the trials that are used as input to the recognition system are obtained based on the coherent averaging of a higher number of responses so that they contain less noise. This, however, requires a longer recording time and may not always be possible in a practical application.

The highest accuracy of 83.8% was achieved by combining EFR and FFR, and applying PCA to eliminate one of the features from EFR and two of the features from FFR. The elimination of the features in both cases (EFR and FFR) was based on their contributions to the variance of data. The best classification accuracy of 83.3% in a previous study that used Linear Discriminant Analysis (LDA) was also obtained when EFR and FFR features were combined (Sadeghian et al., 2015). In the study done by Sadeghian et al., (2015) classification was performed on the complete dataset whereas the approach that we took was to divide the dataset to training and testing sets and obtained our results based on this method. Therefore, compared to the study done by Sadeghian et al., (2015) our approach faced a more challenging problem; however, we still were able to obtain a slightly higher accuracy. This shows that both the EFR and FFR contain useful and non-redundant information for recognition of speech-evoked ABR. The higher rate of recognition for EFR in comparison to FFR confirms the results obtained by Sadeghain et al. (2015). The relatively high recognition accuracies with EFR show that envelope information

could contribute to vowel discrimination, even though it is usually thought that formant information (particularly at F1 and F2) is more important. The lower accuracies obtained with the FFR may be because F2 is not well represented in the evoked response with some of the vowels whose F2 frequencies are higher than 1000 Hz. We have also tried to include responses between 1000Hz and 1500Hz, but the overall recognition results dropped as these higher frequencies did not contribute positively to the recognition. The reason behind this could be because of very weak or absent of responses over 1000Hz due to weaker phase-locking in auditory neurons above this frequency.

To indicate the importance of each feature in our test case we eliminated one feature at a time for both EFR and FFR signals and obtained recognition results using the new feature set. This was done using the amplitude feature set and ANN as the pattern recognition method (without using PCA). Although some features appeared to contribute more to the recognition, the differences in performance were not very large, varying between 3 to 4 percent. In addition, reducing the number of features using the above mentioned method resulted in overall lower recognition rates.

We also followed the same procedure (eliminating one feature at a time) after applying PCA to the amplitude feature sets and that was where we got the best recognition rate performance as shown in Table 4.3. In terms of confusion between specific vowels, /i/ and /u/ had the best overall recognition scores using all 3 amplitude feature sets (Table 4.4). This high recognition accuracy was obtained even though these two vowels have F1 frequencies separated by only 30 Hz (Table 3.1). On the other hand, in some cases, incorrect recognition occurred despite having well-separated F1 frequencies. For example, /a/ was recognized as /u/ 18% of the time with FFR features despite having an F1 separation of 400 Hz.

One reason for the confusion between the vowels could be related to the articulatory configuration required to produce the sound. A classification of articulatory configurations of the vowels can be made based on tongue hump position and its height as has been described by Rabiner and Juang (1993). Based on this classification, vowels /i/ and /ae/ are in the category of front vowels, vowels /a/ and /ɔ/ are considered as mid vowels, and the vowel /u/ is part of the group that is called back vowels. This classification could explain why there is a high recognition rate for /u/ in all the cases (EFR, FFR and EFR+FFR) since it is the only vowel in our test case that belongs to the category of back vowels. This categorization also could be the reason why 17% of /ɔ/ was recognized as /a/ (for EFR signals) since both of these vowels belong to the same category. However, some of the results do not follow the same pattern. For example, based on this classification we cannot explain the result related to the recognition of /ae/ and /i/. While /ae/ and /i/ both belong to the category of front vowels, the recognition rate for either of them to be recognized as the other vowel was 0% in the case of EFR+FFR signals. Similarly/ae/ (a front vowel) was recognized 12% of the time in our test as /u/ which belongs to the back vowels (with EFR+FFR signals).

6 Conclusions, Limitations and Future work

6.1 Conclusions

The main purpose of this study was to investigate if speech-evoked auditory brainstem responses (speech-evoked ABRs) to five English vowels could be discriminated using methods used in automatic speech recognition. The motivation of using these approaches is that the speech-evoked ABRs to vowels are "speech-like" in that they are composed of a fundamental frequency component and its harmonics. Applying PCA on the harmonic amplitude features of both EFR and FFR and eliminating one feature (the one with least variation) from EFR and two features (the two with least variations) from FFR, resulted in a relatively high recognition accuracy of 83.8% when ANN was used as the pattern recognition engine. This accuracy is slightly higher than the 83.3% obtained in a previous study that used LDA (Sadeghian et al., 2015) which confirms that pattern recognition methods such as ANNs are promising approaches for the classification of speech-evoked ABRs.

One of the problems associated with hearing aid use is that of obtaining a best fit. This is a labour-intensive task that involves adjusting several device settings and which often yields less than optimum results, particularly in infants and newborns. Speech-evoked ABRs may help to objectively tune hearing aids by adjusting the device settings so that the accuracy of automatic recognition of the evoked responses is maximized.

Another unexplored potential application of the knowledge gained from the recording and classification of speech-evoked ABRs is to the design of automatic speech recognition (ASR) systems. Despite much progress that has been made in improving these systems, ASR systems remain very susceptible to challenging acoustic environments (Morgan et al., 2004). The overall

performance of ASR is limited by the quality of the extracted features, and today, many of the best performing ASR systems employ front ends that are inspired by auditory processing (e.g., MFCC and PLP). Therefore, there is the expectation that incorporating additional auditory-inspired processing will improve performance (Morgan et al., 2004). Since they represent transformations in the auditory periphery and brainstem, speech-evoked ABRs recorded with various speech stimuli would provide important information about human "front end" processing of speech. As a result, speech-evoked ABRs could guide the development of novel noise-robust front ends for ASR systems, given the robustness of the normal human auditory system in difficult acoustic environments.

6.2 Limitations and Future Work

There were number of limitations in our study that can be addressed by extending this work in the future to cover the following issues:

- **Collecting data in a noisy environment**

  This study was done based on the data that was collected in a quiet environment and the factor of noise was not considered. Since we do not have a clear picture of how the noisy environment may affect the recognition result on speech-evoked ABRs, it would be useful to investigate this matter.

- **Using larger database for recognition**

  In this study, we have used 48 samples from each vowel as our training and test data. Using a larger database from more subjects for the training, would allow the recognition system to generalize to a larger variety of responses from different individuals. Using more samples to form a larger database could also eliminate the possibility of overtraining the ASR system. Another advantage of the larger database is that it allows bigger more diverse population to be covered .

- **Averaging over more repetitions**

  Averaging over more repetitions could improve the performance. The coherent averages of the higher number of responses would have reduced noise and so may provide higher accuracy when used as the inputs to the recognition system.

- **Using natural stimuli**

  In this study, responses from five synthetic vowels were considered for recognition purposes. Our work has shown that the responses that are generated using synthetic vowels carry valuable information in discriminating between the vowels; however, for future work it would be beneficial to use responses to natural vowels for automatic recognition. This could provide better understanding of how the human auditory system processes speech, and would more useful in an application to hearing aid fitting.

- **Using different feature extraction and pattern recognition methods**

  In this study we wanted to investigate the feasibility of using standard speech recognition methods to automatically recognize speech-evoked ABRs. Our findings have shown that the standard methods of speech recognitions can be used for this purpose. However, since this was a first attempt, we only investigated a limited number of feature extraction and pattern recognition methods. There are other methods available either in terms of feature extraction such as relative spectral filtering of log domain coefficients (Shrawankar and Thakare, 2010), or in terms of pattern recognition methods such as support vector machines (Solera-Urena et al., 2007) or hybrid models of HMM and ANN (Trentin and Gori, 2001) that can be examined to identify the best possible method for achieving a higher recognition rate for speech-evoked ABR signals. Also since speech-evoked ABRs are typically the responses to the single stimuli (there is no transition of states between stimuli), it could be more beneficial to use the Gaussian Mixture Model (GMM) that is a parametric probability density function, characterized as a weighted sum of Gaussian

component densities (Vyas, 2013), instead of HMM that models the data as a sequence of states.

- **Using other information at the harmonics**

  In this study, we used amplitude features of the signals and focused on the harmonic frequencies since the useful information in the responses to synthetic vowels is concentrated at these harmonic frequencies. However, there is a possibility that other information at the harmonic frequencies, such as absolute or relative phase, or bispectral features, could improve the accuracy of the recognition even further.

- **Evaluating the proposed method for objective tuning of the hearing aids**

  In this study we investigated the feasibility of using speech-evoked ABR signals with ASR systems. Our hope was that the outcome of this work will eventually help with the process of hearing aid setting adjustment. Our results have shown that the speech-evoked ABRs carry the type of information that ASR systems require to discriminate between different stimuli; however, the goal of being able to maximize recognition accuracy by the hearing aid user by maximizing recognition accuracy by the ASR remains a hypothesis.

  In this work, we did not attempt to adjust hearing aid parameters based on maximizing ASR accuracy. This approach for hearing aid tuning can be investigated in future work.

References

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

Alberti, P.W., 1995, The Anatomy and Physiology of the Ear and Hearing, University of Toronto Press, Canada, pp. 53-62.

Al-Haddad, S.A.R., Samad, S.A., Hussain, A., Ishak, K.A., 2008, 'Isolated Malay digit recognition using pattern recognition fusion of dynamic time warping and hidden Markov models', *American Journal of Applied Sciences*, vol. 5, no. 6, pp. 714-720.

Alotaibi, Y.A., 2005, 'Investigating spoken Arabic digits in speech recognition setting', *Information Science,* vol. 173, pp. 115-139.

Alotaibi, Y.A., 2012, 'Comparing ANN to HMM in implementing limited Arabic vocabulary ASR systems', *International Journal of Speech Technology*, vol. 15, no. 1, pp. 25-32.

Anderson, S., Parbery-Clark, A., White-Schwoch, T., Kraus, N., 2013, 'Auditory brainstem response to complex sounds predicts self-reported speech-in-noise performance', *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 1, pp. 31-43.

Anderson, S., Kraus, N., 2013, 'The potential role of the cABR in assessment and management of hearing impairment', *International Journal of Otolaryngology*, vol. 2013, pp. 1-10.

Anusuya, M.A., Katti, S.K., 2011, 'Front end analysis of speech recognition: a review', *International Journal of Speech Technology*, vol. 14, pp. 99–145.

Anusuya, M.A., Katti, S.K., 2009, 'Speech recognition by machine: a review', *International Journal of Computer Science and Information Security*, vol. 6, no. 3, pp. 181-205.

Azam, S.M., Mansoor, Z.A., Mughal, M.S., Mohsin, S., 2007, 'Urdu spoken digits recognition using classified MFCC and backpropagation neural network', *IEEE Computer Graphics, Imaging and Visualisation*, Bangkok, Thailand, pp. 414-418.

Barnwell, T.P., III, 1980, 'Windowless techniques for LPC analysis', *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 421-427.

Bhattacharjee, U., 2013, 'A comparative study of LPCC and MFCC features for the recognition of assamese phonemes', *International Journal of Engineering Research and Technology*, vol. 2, no. 1, pp. 1-6.

Blum, A., 1992, Neural Networks in C++: An Object-Oriented Framework for Building Connectionist Systems, 1st ed. Wiley, New York, pp. 1-60.

Burkard, R.F., Don, M., Eggermont, J.J., 2006, Auditory Evoked Potentials: Basic Principles and Clinical Application, 1st ed. Lippincott Williams & Wilkins, Philadelphia, pp. 7-20; 229-252.

Busby, P.A., Plant, G.L., 1995, 'Formant frequency values produced by preadolescent boys and girls', *Journal of Acoustical Society of America*, vol. 97, no. 4, pp. 2603-2606.

Cao, X.J., Pan, B.C., Zheng, S.L., Zhang, C.Y., 2008, 'Motion object detection method based on piecemeal principal component analysis of dynamic background updating', *Proceedings of the Seventh International Conference on Machine Learning and Cybernetics*, Kunming, China, pp. 2932-2937.

Cunningham, J., Nicol, T., Zecker, S.G., Bradlow, A., Kraus, N., 2001, 'Neurobiologic responses to speech in noise in children with learning problems: deficits and strategies for improvement', *Clinical Neurophysiology*, vol. 112, pp. 758-767.

Dajani, H.R., Purcell, D., Wong, W., Kunov, H., 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.

Dajani, H.R., Heffernan, B., Giguère, C., 2013, 'Improving hearing aid fitting using the speech-evoked auditory brainstem response', *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Osaka, Japan, pp. 2812–2815.

Dede, G., Sazli, M.H., 2010, 'Speech recognition with artificial neural network', *Digital Signal Precessing*, vol. 20, pp. 763–768.

Deng, L., Hinton, G., Kingsbury, B., 2013, 'New types of deep neural network learning for speech recognition and related applications: An overview', *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing,* Vancouver, Canada, pp. 8599-8603.

Eggermont, J. J., 2015, Auditory Temporal Processing and its Disorders, 1st ed. Oxford University Press, pp. 1-22.

Elgarrai, Z., El Meslouhi, O., Allali, H., Kardouchi, M., 2014, 'Face recognition system using Gabor features and GTK toolkit, *Proceedings of Tenth International Conference on Signal-Image Technology and Internet-Based Systems (SITIS)*, Marrakech, Morocco, pp. 32-36.

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.A., Vincent, P., Bengio, S., 2010, 'Why does unsupervised pre-training help deep learning?', *Journal of Machine Learning Research*, vol. 11, pp. 625-660.

Evans, E.F., 1978, 'Place and time coding of frequency in peripheral auditory system: some physiological pros and cons', *Audiology*, vol. 17, pp. 369-420.

Gaikwad, S.K., Gawali, B.W., Yannawar, P., 2010, 'A Review on Speech recognition technique', *International Journal of Computer Applications*, Vol. 10, no.3, pp. 16-24.

Galbraith, G.C., Arbagey, P.W., Branski, R., Comerci, N., Rector, P.M., 1995, 'Intelligible speech encoded in the human brain stem frequency following response', *NeuroReport*, vol. 6, no. 17, pp. 2363-2367.
Gales, M., Young, S., 2007, 'The application of hidden Markov models in speech recognition', *Foundations and Trends in Signal Processing*, vol. 1, no. 3, pp. 195-304.

He, W., Ding, X., Zhang, R., Chen, J., Zhang, X., Wu, X., 2014, 'Electrically-evoked frequency-following response (EFFR) in the auditory brainstem of guinea pigs', *PLos ONE*, vol. 9, no. 9, pp. 1-10.

Hemdal, J.F., Hughes, G.W., 1967, 'A feature based computer recognition program for the modeling of vowel perception. In W. Wathen-Dunn, Ed.', *Models for the Perception of Speech and Visual Form*, MIT Press, Cambridge, USA.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., Kingsbury, B., 2012, 'Deep neural networks for acoustic modeling in speech recognition', *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97.

Hinton, G.E., Salakhutdinov, R.R., 2006, 'Reducing the dimensionality of data with neural networks', *Science*, vol. 313, pp. 504-507.

Holmberg, M., Gelbart, D., Hemmer, W., 2007, 'Speech encoding in a model of peripheral auditory processing: Quantitative assessment by means of automatic speech recognition', *Speech Communication,* vol. 49, pp. 917–932.

Hoyer, P.O., Hyvarinen, A., 2000, 'Independent component analysis applied to feature extraction from colour and stereo images', *Network: Computation in Neural Systems*, vol. 11, pp. 191-210.

Hudson Beale, M., Hagan, M.T., Demuth, H.B., 2015, Neural Network Toolbox User's Guide Matlab, The Math Works, Inc., Natick.

Itakura, F., 1975, 'Minimum prediction residual principle applied to speech recognition', *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, pp. 67-72.

Johnson, K.L., Nicol, T.G., Kraus, N., 2005, 'Brainstem response to speech: a biological marker of auditory processing', *Ear and Hearing*, vol. 26, no. 5, pp. 424-434.

Johnson, K.L., 2008, Human auditory brainstem representation of transient and sustained acoustic cues for speech, PHD thesis, Department of Communication Sciences and Disorders, Northwestern University.

Kraus, N., Nicol, T., 2005, 'Brainstem origins for cortical 'what' and 'where' pathways in the auditory system', *TRENDS in Neurosciences*, vol. 28, No.4, pp. 176-181.

Kwona, O.W., Lee, T.W., 2004, 'Phoneme recognition using ICA-based feature extraction and transformation', *Signal Processing*, vol. 84, pp. 1005–1019.

Li, B.N.L., Liu, J.N.K., 1999, 'A comparative study of speech segmentation and feature extraction on the recognition of different dialects', *Proceedings of IEEE International Conference on System, Man, and Cybernetics*, vol. 1, Tokyo, Japan, pp. 538–542.

Lima, A., Zen, H., Nankaku, Y., Miyajima, C., Tokuda, K., Kitamura, T., 2004, 'On the use of kernel PCA for feature extraction in speech recognition', *IEICE Transactions on Information and Systems*, vol. E87-D, no. 12, pp. 2802-2811.

Lippmann, R.P., 1988, 'Neural network classifiers for speech recognition', *The Lincoln Laboratory Journal*, vol. 1, no. 1, pp. 107-124.

Maqqor, A., Halli, A., Satori, K., Tairi, H., 2014, 'Using HMM toolkit (HTK) for recognition of Arabic manuscripts characters', *Proceedings of International Conference on Multimedia Computing and Systems (ICMCS)*, Marrakech, Morocco, pp. 475-479.

Mendes, J.A.G., Robson, R.R., Labidi, S., Barros, A.K., 2008, 'Subvocal speech recognition based on EMG signal using independent component analysis and neural network MLP', *Proceedings of Congress on Image and Signal Processing*, Hainan, China, pp. 221-224.

Mohamad-Saleh, J., Hoyle, B.S., 2008, 'Improved neural network performance using principal component analysis on Matlab', *International Journal of the Computer, the Internet and Management*, vol. 16, no. 2, pp. 1-8.

Møller, A.R. 2006a, Section I: Hearing: Anatomy, Physiology, and Disorders of the Auditory System, 2nd ed. Elsevier Science, London, pp. 1-68.

Møller, A.R. 2006b. Section II: The Auditory Nervous System, Hearing: Anatomy, Physiology, and Disorders of the Auditory System, 2nded. Elsevier Science, London, pp. 75-192.

Moore, J.K., 1987, 'The human auditory brain stem as a generator of auditory evoked potentials', *Hearing Research*, vol. 29, pp. 33-43.

Morgan, N., Bourlard, H., Hermansky, H., 2004, 'Automatic speech recognition: an auditory perspective', in Speech Processing in the Auditory System, S. Greenberg,W.Ainsworth, R. Fay, (Eds.), Springer Verlag, New York, pp. 309-338.

Na, J.H., Park, M.S., Choi, J.Y., 2007, 'Pre-clustered principal component analysis for fast training of new face databases', *International Conference on Control, Automation and Systems*, Seoul, Korea, pp. 1144-1149.

Namrata, D., 2013, 'Feature extraction methods LPC, PLP and MFCC in speech recognition', *International Journal for Advance Research in Engineering and Technology*, vol. 1, no. 6, pp. 1-5.

Nguyen, H.Q., Trinh, V.L., Le, T.D., 2010, 'Automatic speech recognition for Vietnamese using HTK system', *Proceedings of IEEE International Conference on Computing and*

Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), Hanoi, Vietnam, pp. 1-4.

Olguin-Olguin, D., Pentland, A., 2010, 'Sensor-based organisational design and engineering', *International Journal of Organisational Design and Engineering*, vol. 1, pp. 69-97.

Oxenham, A.J., 2013, 'Revisiting place and temporal theories of pitch', *Acoustical Science and Technology*, vol. 34, no. 6, pp. 388-396.

Palmer, A.R., Russell, I.J., 1986, 'Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells', *Hearing Research*, vol. 24, pp. 1-15.

Picton, T. W., & Durieux-Smith, A.,1978,'The practice of evoked potential audiometry',*Otolaryngology Clinics of North America*, vol. 11, pp. 263-282.

Prévost, F., Laroche, M., Marcoux, A., Dajani, H.R., 2013, 'Objective measurement of physiological signal-to-noise gain in the brainstem response to a synthetic vowel', *Clinical Neurophysiology*, vol. 124, no. 1, pp. 52-60.

Qi, Y., Hunt, B., 1993, 'Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier', *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 2, pp. 250-255.

Rabiner, L. R., Juang, B.H., 1993, Fundamentals of Speech Recognition, Prentice Hall PTR, Upper Saddle River, pp. 1-390.

Rabiner, L. R., Schafer, R. W., 2010, Theory and Applications of Digital Speech Processing, Prentice Hall, Upper Saddle River, pp. 67-123.

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

Sadeghian, A., Dajani, H.R., Chan, A.D.C., 2015, 'Classification of speech-evoked brainstem responses to English vowels', *Speech Communication*, vol. 68, pp. 69-84.

Sarle, W.S., 1995, 'Stopped training and other remedies for overfitting', *Proceedings of the 27th Symposium on the Interface of Computing Science and Statistics*, pp. 352-360.

Schrode, K.M., Buerkle, N.P., Brittan-Powell, E.F., Bee, M.A., 2014, 'Auditory brainstem responses in Cope's gray treefrog (Hyla chrysoscelis): effects of frequency, level, sex and size', *Journal of Comparative Physiology A – Springer*, vol. 200, pp. 221-238.

Shanthi, T., Chelpa, L., 2013, 'Review of feature extraction techniques in automatic speech recognition', *International Journal of Scientific Engineering and Technology*, vol. 2, no.6, pp. 479-484.

Shrawankar, U., Thakare, V.M., 2010, 'Feature extraction for a speech recognition system in noisy environment: a study', *Second International Conference on Computer Engineering and Applications,* pp. 358-361.
Siegel, A. and Sapru, H.N., 2006, Essential Neuroscience, Lippincott Williams & Wilkins, pp. 292-310.

Sininger, Y.S., 1993, 'Auditory brain stem response for objective measures of hearing', *Ear Hear*, vol. 14, no. 1, pp. 23-30.

Siniscalchi, S.M., Yu, D., Deng, L., Lee, C.H., 2013, 'Speech recognition using long-span temporal patterns in a deep network model', *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 201-204.

Skoe, E., Kraus, N., 2010, 'Auditory brain stem response to complex sounds: a tutorial', *Ear & Hearing*, vol. 31, pp. 302-324.

Solera-Urena, R., Padella-Sendra, J., Martin-Iglesias, D., Gallardo-Antolin, A., Pelaez-Moreno, C., Diaz-De-Maria, F., 2007, 'SVMs for automatic speech recognition: a survey', Progress in Nonlinear Speech Processing, Springer-Verlag, Berlin/Heidelberg, Germany, vol. 4391, pp. 190-216.

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

Starr, A., Picton, T.W., Sininger, Y., Hood, L.J., Berlin, C.I., 1996, 'Auditory Neuropathy', *Brain*, vol. 119, pp. 741-753.

Stuhlman, O., 1943, An Introduction to Biophysics, John Wiley & Sons Inc., New York, pp. 253-310.

Thiagarajan, B., 2015, 'Brain stem evoked response audiometry A review', *Otolaryngology Online Journal*, vol. 5, no.1, pp. 1-7.

Tielen, M.T.J., 1989, 'Fundamental frequency characteristics of middle aged men and women', *Proceedings of the Institute of Phonetic Sciences, University of Amesterdam*, vol. 13, pp. 49-58.

Titze, I.R., 2008, 'Nonlinear source-filter coupling in phonation: Theory', *The Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 2733-2749.

Trabelsi, I., Ben Ayed, D., 2012, 'On the use of different feature extraction methods for linear and nonlinear kernels', *Proceedings of 6th IEEE International Conference on Sciences of Electronics, Technologies of Information and Telecommunications*, Sousse, Tunisia, pp. 797-802.

Trentin, E., Gori, M., 2001, 'A survey of hybrid ANN/HMM models for automatic speech recognition', *Neurocomputing,* vol. 37, pp. 91-126.

Tripathy, S., Baranwal, N., Nandi, G.C., 2013, 'A MFCC based Hindi speech recognition technique using HTK toolkit', *Proceedings of IEEE Second International Conference on Image Information Processing (ICIIP)*, Shimla, India, pp. 539-544.

Vallabha, G., Tuller, B., 2004, 'Choice of filter order in LPC analysis of vowels', *From Sound to Sense*, pp. C203-C208.

Vyas, M., 2013, 'A gaussian mixture model based speech recognition system using Matlab', *Signal and Image Processing : An International Journal (SIPIJ)*, vol.4, no.4, pp. 109-118.

Wallace, M.N., Rutkowski, R.G., Shackleton, T.M., Palmer, A.R., 2002, 'Phase-locked responses to pure tones in the primary auditory cortex', *Hearing Research*, vol. 172, pp. 160-171.

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

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

Woodland, P.C., Gales, M.J.F., Young, S.J., 1997, 'Broadcast news transcription using HTK', *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, vol.2 , pp. 719-722.

Woodland, P.C., Odell, J.J., Valtchev, V., Young, S.J., 1994, 'Large vocabulary continuous speech recognition using HTK', *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Adelaide, Australia, pp. II/125-II/128.

Xiang, Y., Peng, D., Yang, Z., 2014, Blind Source Separation: Dependent Component Analysis, 2015th ed. Springer, New York, pp. 1-14.

Yao, Y., Rosasco, L., Caponnetto, A., 2007, 'On early stopping in gradient descent learning', *Constructive Approximation*, vol. 26, no. 2, pp. 289–315.

Yoshioka, T., Gales, M. J. F., 2015, 'Environmentally robust ASR front-end for deep neural network acoustic models', *Computer Speech and Language,* vol. 31, pp. 65-86.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., 2000, The HTK Book (for HTK Version 3.1).

Young, S.J., Woodland, P.C., Byrne, W.J., 1994, 'Spontaneous speech recognition for the credit card corpus using the HTK toolkit', *Proceedings of IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 615-621.

Zhou, P., Jiang, H., Dai, L., Hu, Y., Liu, Q., 2015, 'State-clustering based multiple deep neural networks modeling approach for speech recognition', *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 23, no, 4, pp. 631-642.