Perceptual and Acoustic Impacts of Aberrant Properties of Electrolaryngeal Speech

by

Geoffrey Seth Meltzner

M.S. Electrical Engineering and Computer Science
Massachusetts Institute of Technology, 1999

Submitted to the Harvard-MIT Division of Health Sciences and Technology Speech and Hearing Biosciences and Technology Program in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Speech and Hearing Biosciences and Technology
at the Massachusetts Institute of Technology, August 2003

Signature of Author____

Harvard-MIT Division of Health Sciences and Technology
Speech and Hearing Biosciences and Technology Program
August 14, 2003

Certified by____

Robert E. Hillman, Ph.D., CCC-SLP
Associate Professor of Otology and Laryngology
Harvard Medical School
Thesis co-supervisor

____

Kenneth N. Stevens, Sc.D.
Clarence J. Lebel Professor of Electrical Engineering, MIT
Thesis co-supervisor

Accepted by____

Martha L. Gray, Ph.D.
Edward Hood Taplin Professor of Medical and Electrical Engineering, HST
Professor of Electrical Engineering and Computer Science, MIT
Co-Director, Division of Health Sciences and Technology

Perceptual and Acoustic Impacts of Aberrant Properties of Electrolaryngeal Speech

by
Geoffrey Seth Meltzner

Submitted to the Harvard-MIT Division of Health Sciences and Technology Speech and Hearing Biosciences and Technology Program on August 14, 2003 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Speech and Hearing Biosciences and Technology

ABSTRACT

Advanced laryngeal cancer is often treated by surgical removal of the larynx (laryngectomy) thus rendering patients unable to produce normal voice and speech. Laryngectomy patients must rely on an alternative means of producing voice and speech, with the most common method being the use of an electrolarynx (EL). The EL is a small, hand-held, electromechanical device that acoustically excites the vocal tract when held against the neck or at the lips. While the EL provides a serviceable means of communication, the resulting speech has several shortcomings in terms of both intelligibility and speech quality.

Previous studies have identified and tried to correct different single selected acoustic properties associated with the abnormal quality of EL speech, but with only limited success. There remains uncertainty about: 1) which components of the EL speech acoustic signal are contributing most to its abnormal quality and 2) what kinds of acoustic enhancements would be most effective in improving the quality of EL speech. Using a combination of listening experiments, acoustic analysis and acoustic modeling, this thesis investigated the perceptual and acoustic impacts of several aberrant properties of EL speech, with the overall goal of using the results to direct future EL speech improvement efforts.

Perceptual experiments conducted by having 10 listeners judge the naturalness of differently enhanced versions of EL speech demonstrated that adding pitch information would produce the most benefit. Removing the EL self-noise and correcting for a lack of low frequency energy would also improve EL speech, but to a lesser extent. However, this study also demonstrated that monotonous, normal speech was found to be more natural than any version of EL speech, indicating that there are other abnormal properties of EL speech contributing to its unnatural quality. An acoustic analysis of a corpus of pre- and post-laryngectomy speech revealed that changes in vocal tract anatomy produce narrower formant bandwidths and spectral zeros that alter the spectral properties of EL speech. Vocal tract modeling confirmed that these spectral zeros are a function of EL placement and thus their effects will vary from user to user.

Even though the addition of pitch information was associated with the greatest improvement in EL speech quality, its implementation is not currently possible because it would require access to underlying linguistic and/or neural processes. Based on these findings it was concluded that an enhancement algorithm that corrects for the low frequency deficit, the interference of the EL self-noise, the narrower formant bandwidths, and the effect of the source location, should produce EL speech whose quality surpasses what is currently available.

Thesis co-supervisor: Robert E. Hillman, Ph.D.
Title: Associate Professor of Otology and Laryngology, Harvard Medical School

Thesis co-supervisor: Kenneth N. Stevens, Sc.D.
Title: Clarence J. Lebel Professor of Electrical Engineering

Table of Contents

1. Introduction

1.1. Motivation

The electrolarynx (EL) is a small, hand-held, electromechanical device that acoustically excites the vocal tract when held against the neck or at the lips. This device is employed primarily by laryngectomy patients who, because they no longer have a larynx, need an alternative voicing source in order to speak. While the electrolarynx generally provides a serviceable means of communication, the resulting speech has several shortcomings in terms of both intelligibility and speech quality.

Since its invention in 1959 (Barney *et al*., 1959) there has been little change in basic EL technology with only a few attempts to improve the quality of EL speech. Some efforts sought to develop a new EL device (Norton and Bernstein 1993) while others employed post-processing schemes to enhance the speech itself (Qi and Weinberg 1991, Cole *et al.* 1997, Espy-Wilson *et al.* 1998). While each of these studies reported success in improving speech quality, the actual magnitude of the improvement and which method was the most effective remain unclear. Furthermore, given that EL speech is inherently monotonous (due to a lack of viable pitch control), one could claim that the best speech any enhancement algorithm could hope to produce would sound like monotonous natural speech.[1] Yet, even when multiple improvement methods are applied simultaneously to EL speech, the resulting speech still retains its artificial quality, sounding significantly less natural than monotonous EL speech. This demonstrates that there remain as yet unaddressed properties of EL speech that also contribute to its unnaturalness. These properties have not yet been adequately studied.

Thus, in an effort to improve the quality of electrolarynx speech, the Voice Project group in the W.M. Keck Neural Prosthesis Research Center in Boston is taking a comprehensive approach to developing an improved EL communication system that seeks to address several problem areas of EL speech (see Section 2.3). As a precursor to successfully developing such a system, it is useful to understand what properties of EL contribute most to its artificial quality and what the underlying causes of these properties are. With this knowledge in hand, research efforts can be focused on altering these properties to make EL speech sound more natural.

[1] It should be noted that this statement is only true insofar as that while some devices do provide some means of pitch control, it is cumbersome and rarely used. For example, the Servox EL provides two buttons that allow the user to drive the device at two different fundamental frequencies, while the frequencies at which the TruTone EL vibrates is proportional to the pressure applied to its activation button. Additionally, the Ultravoice incorporates a fixed pitch contour into its driving signal to provide pitch variation in the EL speech (although the pitch changes cannot be controlled to coincide with the user's intended intonation).

Throughout the rest of this document, the quality of EL speech will be discussed. In this case, the quality of EL speech is defined as how normal or human sounding the speech is. While the quality of speech is affected by its intelligibility (the ability of the speech to be understood), intelligibility was treated as a separate attribute and not addressed in this study.

1.2. Goals

The ultimate goal of the improved EL communication system is to make an EL user's speech sound as close as possible to the way his/her normal speech sounded prior to being laryngectomized. However, given the large gap in the naturalness between EL and normal speech, and the potential complexity of implementing certain improvements, this goal may not be attainable in the short term. Therefore, this study sought to attain basic new knowledge that will provide a solid basis for developing ways to improve EL speech.

Previous studies (Weiss *et al.*1979, Qi and Weinberg 1991, Norton and Bernstein 1993, Espy-Wilson *et al.*1998, Ma *et al.*1999) have indicated that there are three major problems with EL speech: (1) a low frequency energy deficit, (2) interference from the direct sound produced by the EL, and (3) lack of pitch modulation. However, those studies that sought to improve EL speech only dealt with one of the three EL speech issues and only demonstrated some improvement with respect to raw EL Speech. This means that neither the relative effectiveness of each enhancement method nor their combined effectiveness is known. Therefore, the first goal of this thesis was to determine the relative contributions of these three deficits to the artificial quality of EL speech and to formally establish that even if all three of these deficits are adequately addressed, some measure of unnaturalness remains. The ultimate result of this work will be a rank ordering of the relative effectiveness of these three enhancement methods in improving the naturalness of EL speech.

The second goal of this research was to identify and investigate potential causes for the artificial sound quality of EL speech that have not yet been explored. This objective was divided into two parts. The first sub-goal was to characterize the effects of source location on the acoustics of the EL speech. Because the EL voicing source is no longer located at the terminal end of the vocal tract, the vocal tract acoustics have been altered and this change in acoustics may have important effects on the quality of EL speech. The second sub-goal was to investigate the differences between the acoustic properties of normal and EL speech within the same individuals. The availability of a database of pre and post laryngectomy speech recordings of the same subjects provided a unique opportunity to meet this aim.

1.3. Contributions of this research

Both parts of this thesis constitute important steps in achieving the ultimate goal of improving the quality of EL speech quality. Establishing the relative effectiveness of different forms of EL speech enhancement (i.e. a rank ordering) provides a useful guide for future efforts to improve EL speech quality. Such a guide is valuable because it is crucial to know how much benefit one can expect to receive from implementing a certain combination of enhancements because some improvements (such as adding pitch control) are far more difficult to implement than others. In short, it would make little sense to pursue a complicated enhancement scheme to correct for one deficit if it only provides a minimal improvement in EL speech quality.

As this research will demonstrate, correcting for the three major deficits of EL speech still does not result in a close approximation to normal speech. It then follows that there are other deficits in EL speech that have not been explored. Identifying other properties of EL speech that contribute to its unnatural quality helps fill this gap in knowledge and could also be useful in directing future attempts to produce more natural sounding EL speech. The most likely result would be the development of DSP-based enhancement algorithms.

Improving the quality (and possibly the intelligibility) of EL speech would contribute to improving the quality of life for current and future EL users. EL users complain that the artificial nature of EL speech draws unwanted attention to them. A particular problem involves phone use, as EL users often find themselves being mistaken for computers and being hung up on by people with whom they are speaking. Therefore, making EL speech sound more human would vastly improve EL users' experiences when using the phone. This is especially vital in today's world where mobile phones are ubiquitous and important in daily life. Moreover, digital phones, which are increasing in popularity, are the perfect platform for a post-processing enhancement algorithm because the speech is already decomposed before transmission, thus facilitating alteration as needed prior to resynthesis.

2. Background[2]

Each year thousands of people lose the ability to speak normally because they are laryngectomized or suffer laryngeal trauma. As a result, they no longer possess the means to produce normal phonation and therefore must rely on an alternative voicing source to produce alaryngeal speech.

There are three major forms of alaryngeal speech: esophageal speech, tracheo-esophageal (T-E) speech, and electrolarynx speech. Esophageal speech involves inflating the esophagus by an oral injection of air and then expelling it, forcing the upper esophageal sphincter (pharyngoesophageal segment) to vibrate and act as a new voicing source. T-E speech relies on a T-E prosthesis to shunt air from the trachea to the esophagus to inflate the esophagus, which is again expelled to dive the upper esophageal sphincter to serve as a voicing source. Electrolarynx (EL) speech is produced by using an electrically powered device that generates a sound (or buzz) that can be used to acoustically excite the vocal tract, thereby acting as a substitute voicing source.

There is a wide variation in the reported usage of the EL's among alaryngeal speakers. Some studies report that a minority of total laryngectomy patients uses EL speech as their primary means of communication, with estimates of EL use ranging from 11% to 34% (Diedrich & Youngstrom 1977; Gates, *et al*. 1982a; Gates *et al*. 1982b; King, *et al.* 1968; Kommers & Sullivan 1979; Richardson & Bourque 1985; Webster & Duguay 1990). Conversely, other studies have shown that a majority of total laryngectomy patients use some type of EL to communicate, with estimates of EL use ranging from 50% to 66% (Gray & Konrad 1976; Hillman *et al*. 1998; Morris *et al*. 1992). Even though the prevalence of EL speech may vary among specific sub-populations of laryngectomized individuals, it is clear that EL devices continue to represent an important option for speech rehabilitation. Even in cases where esophageal or TEP speech is ultimately developed, EL devices may serve early on to provide a viable and relatively rapid method of post-laryngectomy oral communication (Hillman *et al*. 1998). It is also not uncommon for the EL device to continue to serve as a reliable back-up in instances where individuals experience difficulties with use of esophageal or TEP speech.

2.1. Description of the Electrolarynx

There are two main forms of commercially available EL's: the *neck-type* (transcervical or transcutaneous) and *mouth-type* (transoral or intraoral). Both types of EL devices function on the same principles used in a standard loudspeaker. That is, when activated,

[2] Some of the material in section can also be found in Meltzner *et al.* "Electrolarynx speech: The state-of-the-art and future directions for development" in *Contemporary Considerations in the Treatment and Rehabilitation of Head and Neck Cancer.* Ed. By P.C. Doyle and R.L. Keith

an electromechanical driver within the EL device causes a rigid membrane (or diaphragm) to vibrate, and hence, produces a sound source. The primary difference between the two types of ELs centers on where and how the EL acoustically excites the vocal tract, with one type being placed on the neck (neck-type) and the other at the lips (mouth-type). Because the research in this dissertation is concerned primarily with neck-type EL devices, unless otherwise specified, the terms electrolarynx and EL used herein will refer to neck-type devices only.

The most commonly used EL devices are probably those that are placed against the neck. All transcervical ELs transmit sound energy through neck tissue to provide acoustic excitation of the vocal tract. The optimal location of EL placement on the neck can be highly individualized and is often determined using trial and error to find the point of maximum energy transfer, or that location on the neck that produces the loudest speech output (sometimes referred to as the "sweet spot"). Factors such as the nature of surgical reconstruction and the extent to which post-surgical radiation treatment was used may contribute to the variability in the location and transmission capacity of the "sweet spot" across different laryngectomy patients. There are also a small percentage of laryngectomy patients who, because of post-surgical and/or post-radiation related changes to their neck tissue (e.g., scarring, fibrosis, etc.), cannot transmit usable levels of sound energy into their vocal tracts with a neck-placed EL.

Early forms of the neck-type ELs employed an electromechanical driver, much like a standard loudspeaker, to generate the sound source. In fact, the most successful of the original neck-type ELs used a modified telephone receiver as the driver (Barney *et al.* 1959; Bell Laboratories 1959). The driver was modified by placing a small rigid disk in the center of the diaphragm that was then used to serve as the focal point for transmitting vibrations into the vocal tract. This early device used transistors to generate an electrical pulse train that was used as the driving signal for the modified telephone receiver (speaker). The pulse interval could be adjusted to approximate the average fundamental frequencies of normal adult male or female voices. This EL was marketed by the Western Electric Company (Weiss & Basili 1985) as the "Western Electric Models 5A and 5B". The difference between the two models was in the respective fundamental frequency ranges. The 5A device was designed to be the "male" version with a lower pitch range, while the 5B was the "female" version with a higher pitch range. Both models allowed for some pitch modulation via real time manual adjustment of the voicing activation button, but there was no method for adjusting the loudness of these devices.

Since the introduction of the Western Electric EL's in the late 1950's, other companies have introduced different models of neck-type EL devices. Instead of having the electromechanical transducer drive neck tissue directly, these newer models use a mechanism that operates like a piston hitting a drumhead. When the electromechanical driver is activated, it forces a small cylindrical head mounted on a diaphragm (like a piston) to strike against a rigid plastic disk (like a drumhead), thus, producing a series of impulse-like excitations. This type of system is capable of producing a larger (louder) amplitude signal for vocal tract excitation, but it is essentially a non-linear transducer, thus limiting the extent to which other characteristics of the excitation waveform can be controlled (e.g., wave shape, spectral properties, etc.).

Examples of neck-type EL devices that use non-linear transducers include the *Neovox* by Aurex (Chicago, Illinois)*,* the *Speech-Aid* by Romet (Honolulu, Hawaii), the *Optivox* by Bivona (Gary, Indiana), the *Nu Vois* by Mountain Precision Manufacturing (Boise, Idaho)*,* the *SPKR* by UNI Manufacturing Company (Ontario, Oregon), the *TruTone* and *SolaTone* by Griffin Laboratories (Temecula, California), and the *Servox Inton* by Siemens (Munich, Germany). Examples of neck-type EL devices are shown in Figure 2.1. The Servox Inton is currently one of the most widely used neck-type EL devices. Its features include an internal adjustment screw for modifying the fundamental frequency of vibration to accommodate male and female users, two externally-placed control buttons that provide dual pitch variation, an externally-placed dial for volume adjustments, and rechargeable batteries (see Figure 2.1). Similar features are can be found on the other models of neck-type ELs, while the specifications vary to some extent. Although there is little objective information concerning how the different models of neck-type EL devices compare to each other in terms of performance criteria such as sound quality or ease of use, it has been demonstrated that the intelligibility of EL speech produced by the older Western Electric devices and the newer Servox EL are similar (Weiss & Basili, 1985). Future studies are needed to establish whether particular EL features such as dynamic pitch modulation offered by the *TruTone* or the dual pitch modulation capabilities of the *Servox Inton* improve EL speech quality or intelligibility.

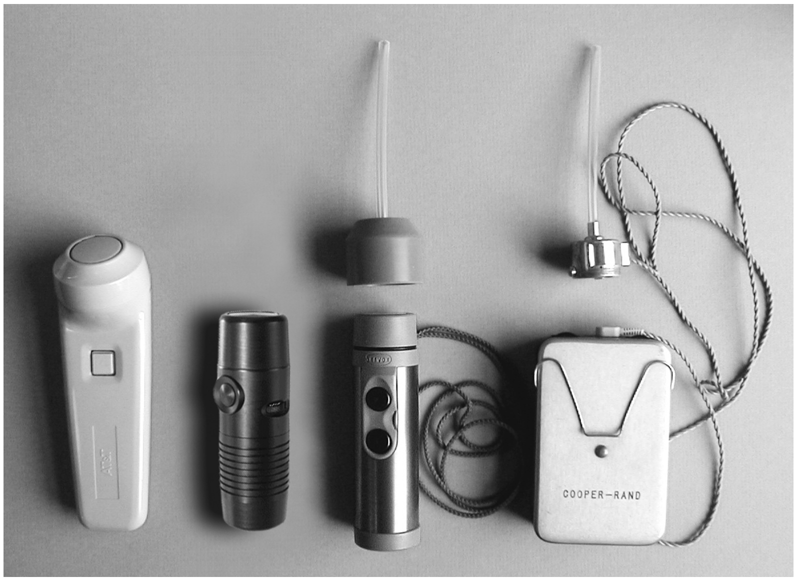


Figure 2.1. Examples of several different electrolarynxes. From left to right: the Western Electric neck-type, the TruTone neck-type , the Siemens Servox neck-type with oral adapter, and the Cooper-Rand mouth-type.

One shortcoming common to all neck-type ELs is that in addition to providing acoustic excitation to the vocal tract, these devices also directly radiate sound energy into the surrounding air. The resulting airborne "buzzing" sound competes with, or masks, the

EL speech that is being produced via vocal tract excitation. This phenomenon, which occurs to a greater or lesser degree depending on how well a particular device can be coupled to the neck of a given individual, clearly has a negative impact on both the intelligibility and quality of EL speech and the overall communicative effectiveness when using such a device (see below).

2.2. Deficiencies of EL Speech

While today's commercially available neck-type and mouth-type ELs generally provide a serviceable means of communication for the laryngectomized patients who depend on them, there are a number of persistent deficits in EL speech communication. The most problematic of these deficits were highlighted in a needs assessment that was recently conducted as part of an effort to establish a research program that focuses on developing an improved EL communication system (VA Rehabilitation Research and Development Grant C1996DA). Seventeen total laryngectomy EL users and seven speech-language pathologists (experienced in laryngectomy speech rehabilitation) were asked to rank order a randomized list of major deficits in EL speech communication that have been cited in the literature, as well as to add and rank any additional factors that they felt were problems with the use of currently available EL devices. The top five deficits identified by both groups were the same with a slightly different rank ordering by each group. These deficits include the following and the corresponding statements used in the needs assessment are shown in parentheses: 1) reduced intelligibility ("EL speech is hard to understand"), 2) lack of fine control over pitch and loudness variation, and voice onset and offset ("EL speech is monotonous"), 3) unnatural, non-human sound quality ("EL speech sounds mechanical"), 4) reduced loudness ("EL speech is too quiet"), and 5) inconveniences related to EL use ("EL is inconvenient to use"). Each of these five areas of deficit is discussed briefly below.

Several studies have demonstrated that EL speech has reduced intelligibility, with the amount of reduction related to the type of speech material that is used. When closed-set response paradigms are employed (i.e., listeners have to identify the target word from a limited set of options), intelligibility for EL speech has been reported to range from 80.5% to 90% (Hillman et al., 1998; Weiss, *et al.*, 1979). However, when listeners have been asked to transcribe running speech produced with an EL, intelligibility drops to a range of 36% to 57% (Weiss & Basili, 1985; Weiss *et al.*, 1979). Studies that have examined the types of intelligibility errors that listeners make in evaluating EL speech have reported that the greatest source of confusion is in discriminating between voiced and unvoiced stop consonants, with more of these errors occurring when consonants are in the word-initial position as compared to the word-final position (Weiss & Basili, 1985; Weiss *et al.*, 1979). Weiss *et al.* (1979) postulated that voicing feature confusions occur more frequently for word-initial consonants because EL users are unable to exercise the fine control over voice onset time that is necessary for producing these voiced-voiceless distinctions. Furthermore, the lower incidence of voiced-voiceless confusions for word-final consonants is attributed to the additional cues for this distinction that are provided by the length of the vowel preceding the consonant (i.e., vowels preceding unvoiced consonants are of significantly shorter duration than vowels preceding voiced consonants, at least in utterance-final positions) (Weiss & Basili, 1985).

There is evidence that the intelligibility of EL speech also varies depending on characteristics of the listener and the listening environment. Clark (1985) used two groups of judges, one comprised of normal hearing young adults, and the other made up of older adults with high-frequency hearing loss. Judges evaluated the intelligibility of normal, esophageal, TEP, and EL speech in quiet and with competing speech in the background at different signal-to-noise ratios. Overall, the young normally-hearing judges did better in evaluating intelligibility than the older hearing-impaired group; however, the hearing impaired group always found artificial laryngeal speech to be more intelligible than the other modes of alaryngeal communication. In terms of performance in the presence of competing speech noise, EL speech was more intelligible than the other modes of alaryngeal communication (e.g., esophageal and TEP speech) across the different signal-to-noise conditions. Furthermore, it has been reported that over telephone lines, EL speech is more intelligible than esophageal speech (Damste, 1975). However, it should be pointed out that many EL users complain that they cannot be adequately heard in a noisy environment.

In addition to the difficulties with voiced/voiceless distinctions for EL speech associated with poor on/off control, EL devices also lack the capability to produce finely controlled dynamic changes in pitch and loudness. The lack of such control appears to contribute to the impression that EL speech is monotonous-sounding, as well as probably contributing to the negative perceptions of EL speech as sounding non-human, mechanical, robotic, etc. (Bennett & Weinberg, 1973). Many EL users describe how the unnatural sound quality of their speech draws unwanted attention, and can even spawn barriers to communication, such as the oft-heard tale of EL users being hung-up on during attempts to use the telephone. In attempting to compensate for these deficits, some ELs include a finger-controlled button or switch for altering pitch or loudness. Unfortunately, finger-based control appears too cumbersome to adequately mimic the natural variation of these parameters in normal speech. The lack of adequate pitch control has been shown to be even more detrimental to the intelligibility of EL users who speak tone-based languages such as Thai and Cantonese (Gandour, *et al.* 1988; Ng *et al.* 1998). EL speakers also often complain that EL use is inconvenient because it occupies the use of one hand. In addition, the most commonly used devices are very conspicuous because they must be held to the neck or mouth (Goode, 1969), thus, attracting unwanted attention to this method of alaryngeal communication.

While the lack of normal pitch and loudness variation appears to contribute to the unnatural sound quality of EL speech, there is evidence that additional acoustic characteristics of the EL sound source may also play a role. Several investigators have noted that there is significantly less sound energy below 500 Hz in EL speech as compared to normal, laryngeal speech (Qi & Weinberg, 1991; Weiss *et al.*, 1979). Figure 2.2 illustrates the lack of low frequency energy in EL speech by comparing the spectra of the same vowel produced by the same speaker using both his normal voice and a Servox EL. One can see that in the EL speech spectrum that the energy below 500 Hz (highlighted in gray) is far less that that found in the spectrum of the normal vowel. Compensating for this “low frequency deficit” via a second order filter improves the quality of EL speech (Qi & Weinberg, 1991). Further, it is possible that the lack of random period-to-period fluctuations in both the frequency (jitter) and amplitude (shimmer) of typical EL sound sources may also contribute the unnatural sound quality of these devices. Supporting this possibility is evidence that a constant pitch in the voicing source of synthesized speech produces a mechanical sound quality (Klatt & Klatt, 1990). To date, however, there has been no systematic study of the effect on EL speech quality of adding such random fluctuations in pitch and amplitude to EL sound sources. Finally, the already mentioned shortcoming of neck-type ELs to directly radiate sound energy (the electronic "buzz") into the surrounding air, also likely contributes to the unnatural quality of speech produced with these types of devices.

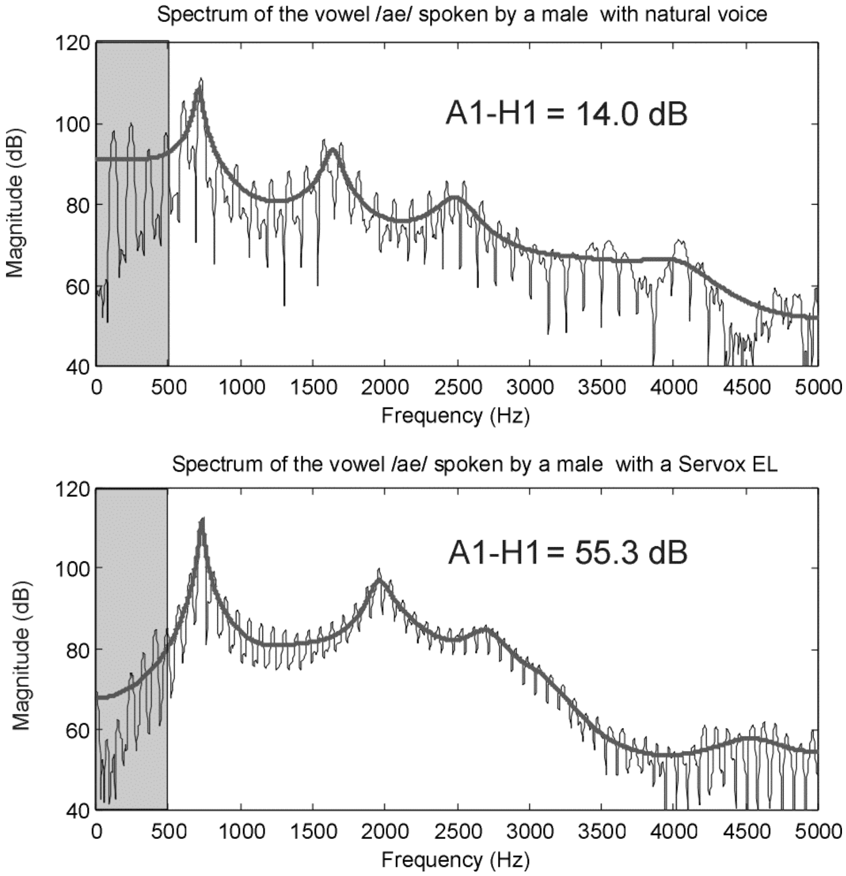


Figure 2.2. The spectral content of both normal (top) and electrolaryngeal (bottom) speech. The thick solid line representing the linear predictive (LP) smooth spectrum is displayed to emphasize the overall spectral shape. The spectrum below 500 Hz has been highlighted in gray to emphasize the low frequency deficit inherent in EL speech. The difference between the amplitude of the first formant and the amplitude of the first harmonic (A1-H1) is also shown for each case. Notice that in EL speech, this difference is much greater than that found in normal speech, indicating that there is little energy at low frequencies. These data were obtained from a normal male subject recorded in an acoustic chamber with a microphone placed at a distance of 2 cm from the lips.

2.3. Previous attempts at improving EL speech

It is clear there is much room for improving EL speech communication. However, until recently, there has been a little effort to remedy the primary deficits associated with EL speech production since EL technology was introduced over 40 years ago (Barney et al., 1959). Moreover, these recent attempts to improve EL speech have produced few, if any, clinically viable improvements. The lack of successful innovation can be at least partly attributed to the fact that there are relatively few EL users, that is, the potential commercial market is too small for mainstream industry to justify investing in EL research and design. The subsequent section will describe some recent and ongoing efforts to improve EL speech communication and indicate future directions for work in this area.

An early attempt to improve the intelligibility of EL speech produced with a mouth-type device employed a simple amplification system developed by an EL user and called the *Voice Volume Aid* (Verdolini, *et al.* 1985). The amplification system, which consisted of a microphone placed close to the user's lips and attached to a powered speaker worn in a shirt pocket, sought to improve the intelligibility of EL speech by amplifying the sound produced at the lips. It was believed that since the signal to noise ratio at the lips is greater than at a distance away from the EL user, amplifying the speech at the lips would improve intelligibility. It was found that the Voice Volume Aid enhanced EL speech intelligibility in quiet rooms or in rooms with moderate background noise (66 and 72 dB SPL, respectively), but was less effective in relatively high levels of background noise (76 dB SPL).

Norton and Bernstein (1993) tested a new design for an EL sound source based on an attempt to measure the sound transmission properties of neck tissue. They also attempted to minimize the sound that is directly radiated from the neck-type EL by encasing the EL in sound shielding. These proposed improvements to the EL source were implemented on a large, heavy, bench-top mini-shaker, making their prototype impractical for routine use. In addition, there is some question about whether their estimates of the neck transfer function were confounded by vocal tract formant artifact (Meltzner *et al.* 2003). However, the speech produced with the newly configured sound source was subjectively judged to sound better, thereby indicating that such alterations to the EL sound source could potentially improve the quality of EL speech.

In an endeavor to give EL users some degree of improved dynamic pitch control, Uemi *et al*. (1994) designed a device that used air pressure measurements obtained from a resistive component placed over the stoma to control the fundamental frequency of an EL. Unfortunately, only 2 of the 16 study subjects studied were able to master the control of the device and thereby produce pitch contours that resembled those in normal speech. Their results demonstrate how a pitch control device must not be too difficult for the user to employ in order to be clinically practical.

A different approach to adding pitch information to EL speech is taken by the latest version of the Ultravoice EL, which alters the fundamental frequency at which it vibrates in a fixed fashion, providing the user with a fixed pitch contour. Theoretically, having at least some degree of pitch change should make EL speech sound more natural, although this fixed pitch contour approach has yet to be formally tested. It remains to be seen whether a pitch contour that is independent of the speaker's intended intonation is better than no pitch change at all.

Some investigators have applied signal-processing techniques to post-process recorded EL speech in order to remove the effects of the directly radiated EL noise (i.e. sound not transmitted through the neck wall, or "self-noise"). Cole et al. (1997) demonstrated that a combination of noise reduction algorithms (spectral subtraction and root cepstral subtraction) originally developed for the removal of noise corruption in speech signals could be used to effectively remove the EL self-noise for the recordings of EL speakers. Nevertheless, the perceptual improvement afforded by this noise reduction algorithm was modest at best. The improved speech produced a mean quality rating of 2.8 (on a 1 to 5 scale) while the unaltered EL speech produced a mean rating of 2.5. Espy-Wilson et al. (1998) used a somewhat different approach to remove the EL self-noise. They simultaneously recorded the output at both the lips and at the EL itself, and then employed both signals in an adaptive filtering algorithm to remove the directly radiated EL noise. Spectral analysis of the filtered speech demonstrated that the enhancement algorithm effectively removed the directly radiated EL sound during non-sonorant speech intervals but with no significant impact on overall intelligibility. Perceptual experiments revealed that listeners generally preferred the post-processed enhanced speech as compared to the unfiltered speech.

There have also been efforts aimed at using post-processing techniques to compensate for deficits in the EL sound source. Qi and Weinberg (1991) attempted to improve the quality of EL speech by enhancing its low frequency content. Hypothesizing that the low frequency roll-off of EL speech first noted by Weiss *et al.* (1979) was at least partially responsible for the poor quality of EL speech, Qi and Weinberg developed an optimal second order low pass filter to compensate for this "low frequency deficit." Briefly, this filter was designed to emphasize spectral energy below 500 Hz without significantly altering the level of energy at higher frequencies. Perceptual experiments showed that almost all listeners preferred the EL speech with the low frequency enhancement. In an even more ambitious approach, Ma et al. (1999) used cepstral analysis of speech to replace the EL excitation signal with a normal speech excitation signal, while keeping the vocal tract information constant. Not only did the normal excitation signal contain the proper frequency content (i.e., no low frequency deficit), but it also contained a natural pitch contour to help eliminate the monotone quality of EL speech. In formal listening experiments, most judges preferred the post-processed speech to the original EL speech. The practical application of this enhancement technique is limited since it would require having a natural speech version of the utterances being spoken that could then be used as a basis for enhancing the EL speech. However, both reports demonstrate improvements in EL speech quality gained by recognizing and compensating for the differences between conventional EL sound sources and the normal laryngeal voicing source. Specifically, these post-processing strategies demonstrate the potential for substantial improvements in EL speech quality over the telephone and in broader contexts if these strategies can be implemented in a truly portable system that is capable of real-time processing.

Nevertheless, the fact remains that despite these reported improvements, EL speech still contains flaws that give it its obviously unnatural sound quality. Anecdotal evidence suggests that even combining multiple enhancement techniques still leaves EL speech sounding mechanical. This indicates that either these studies did not adequately address the properties of EL speech that are responsible for its unnatural quality and/or there remain other properties of EL speech that contribute to the unnatural sound that have not yet been adequately examined.

3. Perceptual impacts of aberrant properties of EL speech[3]

3.1. Introduction

The basics of current EL technology were introduced over 40 years ago (Barney *et al.* 1959) but until relatively recently there has been a little effort to remedy the primary deficits associated with EL speech. As mentioned in the previous chapter, Qi and Weinberg (1991) attempted to improve the quality of EL speech by enhancing its low frequency content. They developed an optimal second order low pass filter to compensate for the "low frequency deficit" in EL speech and found that the resulting speech was preferred over raw EL speech.

Cole *et al*. (1997) demonstrated that a combination of noise reduction algorithms (spectral subtraction and root cepstral subtraction) originally developed for the removal of noise corruption in speech signals could be used to effectively remove the EL self-noise from audio recordings of EL speakers. Espy-Wilson *et al.* (1998) used a somewhat different approach to remove the EL self-noise. They simultaneously recorded the output at both the lips and at the EL, and then employed both signals in an adaptive filtering algorithm to remove the directly radiated EL noise.

1. Uemi *et al.* (1994) designed a device that used air pressure measurements obtained from a resistive component placed over the stoma to control the fundamental frequency of an EL. In an even more ambitious approach, Ma *et al.* (1999) used cepstral analysis of speech to replace the EL excitation signal with a normal speech excitation signal, while keeping the vocal tract information constant. Not only did the normal excitation signal contain the proper frequency content (i.e., no low frequency deficit), but it also contained a natural pitch contour to help eliminate the monotone quality of EL speech.

The success of these studies indicates that EL users could gain some benefit from an EL communication system that improves the quality of the speech in one of these ways. However, each of these enhancements has been only tried in isolation and some are more difficult to implement than others. Thus, knowing the relative contribution that these different enhancements make (both alone and in combination) to improve the perceived quality of EL speech is critical in determining which approaches should be given priority in future attempts to actually implement such enhancements in a device that patients can use. Moreover, formally assessing how closely the perceived quality of the best enhanced EL speech approximates normal natural speech would indicate the limits of current enhancement approaches, and serve to estimate how much more room there is for

[3] An abridged version of this chapter was submitted to and accepted by the VOQUAL '03 conference in Geneva, Switzerland.

further improving EL speech. The goals of this investigation were to better quantify the sources and perceptual impact of abnormal acoustic properties typically found in EL speech by: 1) quantifying the relative contribution that acoustic enhancements make, both individually and in combination, to improving the perceived quality of EL speech and 2) determine how closely the best enhanced EL speech approximates normal-natural speech quality.

3.2. Methods

3.2.1. Data Recording

Two normal (i.e. non-laryngectomized) speakers, one male and one female, produced two sentences using both their natural voices and a neck-placed Servox electrolarynx (Siemens Corp.). The speakers were instructed to hold their breaths and maintain a closed glottis while talking with the Servox, in order to approximate the anatomical condition of laryngectomy patients in which the lower airway is disconnected from the upper airway. Recordings were made under two conditions: (1) inside an acoustic chamber and (2) with the subject's face sealed in a specially constructed port in the door of a sound isolated booth (see Appendix B). This was done to essentially eliminate the self-noise of the neck placed EL from the audio recording of the speech. All recordings were made with a Sennheiser (Model K3-U) microphone placed 15 cm. from the lips.

The subjects were asked to say two sentences: (1) "*We were away a year ago when I had no money*" and (2) "*She tried the cap and fleece so she could pet the puck*." The lengths of both sentences were chosen so that they could be easily spoken in a single breath (Crystal and House 1982, Mitchell *et al.* 1996) to prevent the speakers from inserting pauses in the speech. Because EL speech typically does not contain any pauses, any pauses in normal speech could provide listeners with another cue to distinguish between normal and EL speech (both raw and enhanced). The two sentences differ in their phonemic makeup: the first sentence is comprised entirely of voiced phonemes while the second contains both voiced and unvoiced phonemes. The speech signals were low pass filtered at 20 kHz by a 4 pole Bessel Filter (Axon Instruments Cyberamp) prior to being digitized at 100 kHz (Axon Instruments Digidata acquisition board and accompanying Axoscope software). The signals were then appropriately low pass filtered and downsampled to 8 kHz in MATLAB because this is the bandwidth at which the vocoder used in this study operates (See section 3.2.2).

3.2.2. Generation of sentence stimulus material

For each speaker, a total of ten versions of each sentence were generated: a normal version, a normal version with a fixed/mono pitch, raw EL speech, and EL speech with either one of the enhancements, all possible combinations of two enhancements, or all three enhancements. The following enhancements were implemented: low frequency enhancement (L), self-noise reduction (N), and added pitch information (P). Throughout the rest of this thesis, enhanced versions of EL speech will be denoted by placing an L, N, or P or some combination thereof. For example, so that low frequency enhanced, noise reduced EL speech become EL-LN. A description of the sentence version associated with each acronym is presented in Table 3.1.

Table 3.1: Notation and Description of Sentence Stimuli

Sentence Version	Sentence Desciption
EL-raw	Unprocessed EL speech
EL-L	EL speech with low frequency enhancement
EL-N	EL speech with noise reduction
EL-P	EL speech with pitch modulation
EL-LN	EL speech with low frequency enhancement & noise reduction
EL-LP	EL speech with low frequency enhancement & pitch modulation
EL-NP	EL speech with noise reduction & pitch modulation
EL-LNP	EL speech with all three enhancements
norm-mono	Monotonous (fixed pitch) normal speech
Normal	Normal natural speech

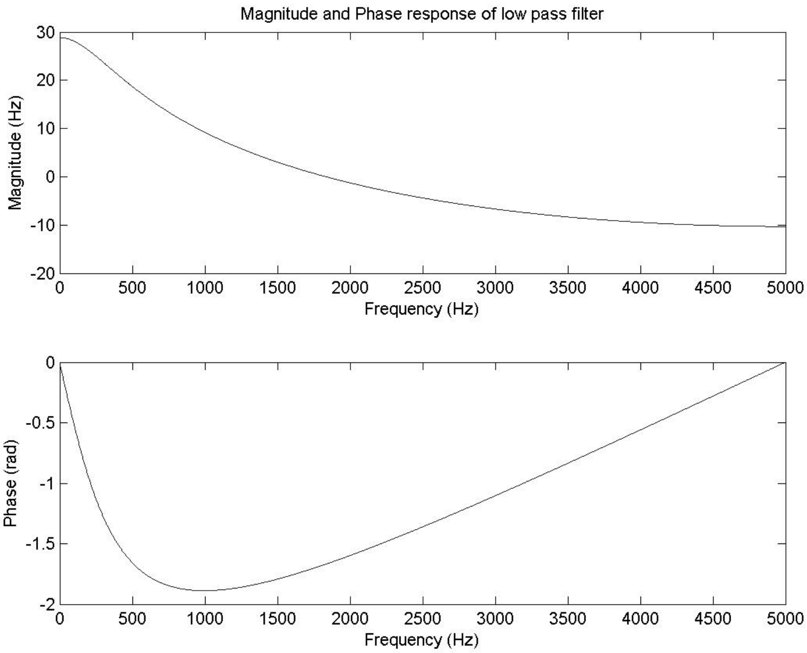


Figure 3.1. The magnitude (top) and phase (bottom) response of the low frequency enhancement filter specified by Qi and Weinberg (1991).

The low frequency enhancement was implemented by processing the sentences through the two-pole low pass filter specified by Qi and Weinberg (1991):

$$H(z) = \frac{1}{\left(1 - az^{-1}\right)^2} \tag{3.1}$$

where $a = 0.81$. The magnitude and phase response of this filter are shown in Figure 3.1.

An example of the effect of employing the low frequency enhancement filter is demonstrated in Figure 3.2.

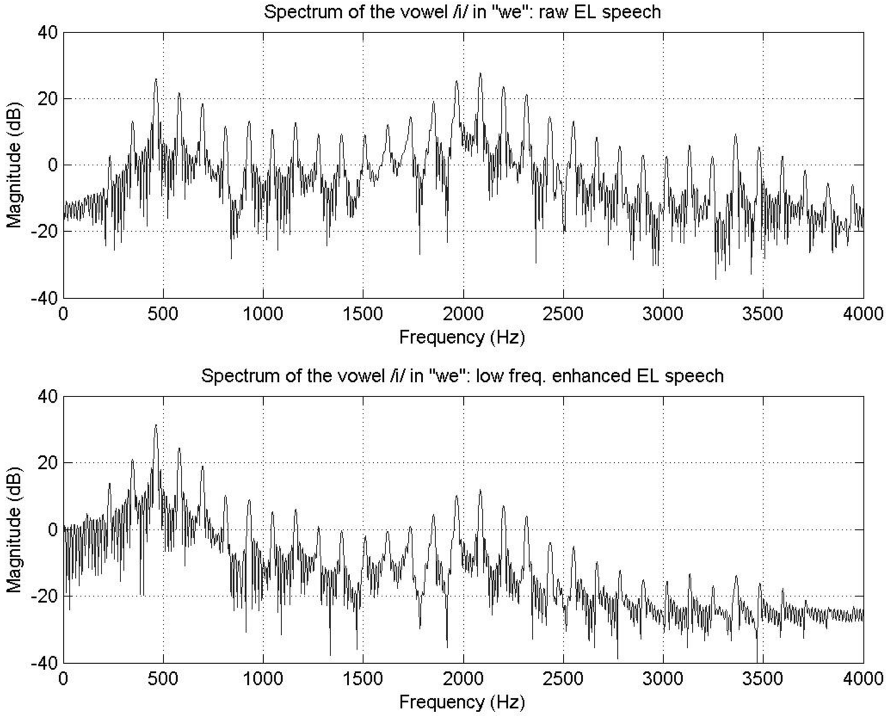


Figure 3.2. The spectrum of the vowel /i/ in "we" spoken with a Servox EL by a male speaker. The spectrum of the raw speech (top) shows the low frequency deficit and a spectral tilt such that the amplitude of the second formant is greater than that of the first. The spectrum of the enhanced speech (bottom) demonstrates that the low pass filter increases the amount of low frequency energy (relative to energy in the overall spectrum) and corrects the spectral tilt.

Because speaking through the port in the door tended to slightly restrict articulatory movements of the jaw and lips, it was decided to make this the default. Therefore, every sentence presented to the listeners was recorded under this condition so as to remove differences in articulation as potential perceptual cues. To construct stimuli representing unprocessed/raw EL speech, a time-aligned estimate of the EL self-noise was added to the EL sentences that were recorded through the port of the sound isolated booth. The self-noise estimates were made from free field recordings in the sound isolated booth while the speakers held the EL to their necks and kept their mouths closed.

The addition of the proper pitch information to the EL speech involved 3 steps. First, the normal and EL sentences were time aligned using the Pitch-Synchronous Overlap-Add (PSOLA) algorithm (Moulines and Charpentier 1990) found in the Praat (www.praat.org) software package, such that the phonemes of both sentences had the same onset times and duration. Both sentences were then analyzed using a modified version of a Mixed Excitation Linear Predictive (MELP) vocoder (McCree and Barnwell 1995). The MELP vocoder was chosen for this task because it effectively separates speech into source and filter parameters that are easily manipulable, while producing high quality resynthesized speech. (A more detailed discussion of the MELP vocoder and how it was modified can be found in Appendix B.) Finally, the pitch track obtained from the MELP analysis of the normal sentence was used in the MELP synthesis of the EL speech, thus giving the EL sentence the same exact pitch contour as that of the normal sentence. Because the second sentence contained unvoiced phonemes, there were sections in which no pitch estimate could be made during MELP analysis. Therefore, before the measured pitch contour was used in the resynthesis of the EL sentences, the sections of the pitch contour corresponding to the unvoiced sections were set equal to the last pitch measurement made prior to the onset of each unvoiced section. As a result, the pitch was set at a fixed value during what were the unvoiced sections of the normal version of the voiced/voiceless sentence. Moreover, during the resynthesis of the EL versions of this sentence, every frame was set as voiced.

The MELP vocoder was also used to set the pitch of the monotonous EL sentences to the mean pitch of the normal sentences. This step was taken to remove the potentially confounding influence that differences in the pitches of the stimuli might have on perceptual comparisons. Similarly, the monotonous normal speech token was generated by fixing the pitch of the whole sentence at the mean pitch. It should be noted that for the female speaker, implementing this step meant that the pitch would be at a frequency beyond what a Servox EL is able to produce, in effect, making the EL speech sentences "better" than they really should be. However, it was decided that removing differences that could act as perceptual cues was more important than keeping the pitch within the Servox range.

3.3. Experimental Procedure

The experimental procedure consisted of using the Method of Paired Comparisons (Torgerson 1957) with an accompanying visual analog scale. For each speaker-sentence condition, all combinations of pairs of speech tokens (45) were presented via computer D/A (Aureal Vortex soundcard) and headphones to a group of 10 naïve, normal hearing listeners (5 male and 5 female). The listeners were required to indicate on a computer response screen which of the two tokens in each pair "sounded most like normal natural speech". Once this decision was made, the listener was then asked to use a mouse–controlled visual analog scale (VAS) to rate how different the chosen token was from normal natural speech. The scale was 10 cm long and ranged from "Not At All Different" to "Very Different", with the distance (in cm.) from "Not At all Different" used as the rating of the stimulus. Each complete set of tokens was presented twice in different random orders to assess listener reliability. Prior to beginning the experiment, all 10 speech tokens were played to the listeners to familiarize them with the range of speech quality that the tokens spanned. Once the experiment began, however, the subjects could only listen to the normal token as a reference. This allowed the normal token to act as an anchor so that all listeners would have a common frame of reference to make their judgments.

3.4. Analysis

3.4.1. Paired Comparison Data: Law of Comparative Judgment

The data collected from the Paired Comparison procedure were analyzed using Thurstone's Law of Comparative Judgment (Thurstone 1927). It is assumed in each subject, a group of stimuli elicits a set of discriminal processes (or perceptions) along a psychological continuum with respect to a certain attribute of the stimuli. However, since human observers tend to be inconsistent, a stimulus will not always elicit the same discriminal process every time it is presented. As such, the most common process is labeled the *modal discriminal process*, while the spread of the discriminal process is called the *discriminal dispersion*. If these discriminal processes are modeled as normal random variables, then the modal discriminal processes and the discriminal dispersions are the mean and standard deviation of the random variables where the mean is taken to be the scale value on the psychological continuum.

If two stimuli, *j* and *k*, are presented to a group of several listeners, and stimulus *j* chosen more often to be "greater" than stimulus *k* (for a certain attribute) then it can be assumed that the scale value, S_j of stimulus *j*, is greater than the scale value, S_k of stimulus *k*. Furthermore, the proportion of times that stimulus *j* is chosen over stimulus *k* is related to the difference between the scale values, i.e. the discriminal difference. This discriminal difference is also a normal random variable with a mean of S_j-S_k and standard deviation of

$$\sigma_{j-k} = \sqrt{\sigma_j^2 + \sigma_k^2 - 2r_{jk}\sigma_k\sigma_j} \qquad (3.2)$$

where σ_j and σ_k are the discriminal dispersions of stimuli *j* and *k* respectively, and r_{jk} is the correlation between the two stimuli. It then follows that the discriminal dispersion between two stimuli can be calculated from

$$S_j - S_k = z_{jk}\sqrt{\sigma_j^2 + \sigma_k^2 - 2r_{jk}\sigma_k\sigma_j} \qquad (3.3)$$

where z_{jk} is the normal deviate corresponding to the theoretical proportion stimulus *j* is judged "greater" than stimulus *k*. Since the theoretical values aren't available, they are estimated from the empirical values obtained from the paired comparisons experiment. Equation 3.3 represents the complete version of Thurstone's Law of Comparative Judgment. It is, unfortunately, impossible to solve Equation 3.3 because there will always be a larger number of unknowns than observable equations (Torgerson 1957) and thus some simplifying assumptions must be made. Thurstone (1927) discusses several different cases of simplifications, however, this discussion will restrict itself to Thurstone's Case V, where it is assumed that that the discriminal dispersions are equal and that correlations between stimuli are also equal. This reduces equation (3.2) to

$$S_j - S_k = z_{jk}\sqrt{2\sigma^2(1-r)}. \tag{3.4}$$

The term $\sqrt{2\sigma^2(1-r)}$ is a scaling constant and can be set equal to 1 without any loss of generality (Edwards 1957) so that

$$S_j - S_k = z_{jk} \tag{3.5}$$

Hence the scale value of each stimulus can be found, thus providing not only a ranking of the stimuli but the psychological distance between them on the psychological continuum.

The following procedure is used to generate the z_{jk}. The proportion of times stimulus j is judged greater than stimulus k, p_{jk} is entered into the *jth* column and *kth* row of a matrix, *P*, such as the one shown in Table 3.2. Because no stimulus is ever presented against itself, the diagonals of the *P* matrix remain empty. The *Z* matrix, whose cells contain the z_{jk}, is found by computing the normal deviates of the entries in the *P* matrix. The diagonal entries of the Z matrix are set to zero. If the *Z* matrix is full (i.e. there are no infinite values in any of the entries) then the S_j are easily computed by averaging each column of the *Z* matrix. However, in many circumstances, one stimulus is always judged to be "better" (or "worse") than another thereby producing a proportion, p_{jk}, of 1 (or 0) and a corresponding infinite z_{jk}. In such cases, simply averaging the columns of the *Z* matrix is not possible and another method of estimating the scale values must be used. Kaiser and Serlin (1978) suggested a least squares method to estimate the scale values that was valid as long as the data collected from every stimulus is at least indirectly connected to each other, i.e. as long as no stimulus is always judged to be better (or worse) than all the others. When the Z matrix is full, the Kaiser-Serlin method reduces to averaging the columns of the matrix.

Unfortunately, because of the nature of the stimuli used in this experiment, in some instances, this necessary condition was violated. Specifically, for some speaker-sentence conditions, the normal sentence was always judged to sound more like normal natural speech than all of the other speech tokens. In such cases, the data collected for the normal sentences can be thrown out and the Kaiser-Serlin method can be applied to the remaining sub-matrix but no information can be obtained on the scale value of the normal sentence (it is effectively infinity).

Therefore, this study made use of the solution to this problem provided by Krus and Krus (1979), who suggest the following transformation from the proportions, p_{jk} to the z-scores, z_{jk}:

Table 3.2: The P Matrix

Stimulus	1	2	…	n
1	-	p_{21}	…	p_{n1}
2	p_{12}	-	…	p_{n2}
…	…	…	-	…
n	p_{1n}	p_{2n}	…	-

$$z_{jk} = \frac{p_{jk} - p_{kj}}{\sqrt{\frac{p_{jk} + p_{kj}}{N}}} \tag{3.6}$$

where N is the total number of times the stimulus pair, (j,k) was presented. This transformation provides a rational z-score even when p_{jk} equals one or zero which is proportional to the square root of the number of observations. The diagonal entries of the Z-matrix are set to zero, the z-score of a proportion of 0.5, i.e. what would be expected if pairs of the same stimuli were presented. The Kramer-Serlin method was applied to these scores to produce the scale values. The scale values were then shifted by the amount necessary to set scale value of the lowest ranked token to zero.

3.4.2. Visual Analog Scale Data

The distance in centimeters from the end of the VAS labeled "Not at all different" was used as an estimate of how different a listener judged a speech token to be from normal natural speech. The lower the rating, the less different from normal speech a sentence is judged to be. These distances were used to compute a mean distance for each speech type. A 3-way analysis of variance (ANOVA) was conducted on the entire data set to look for significant main effects and interactions between the speech ratings, the gender of the speaker and the type of sentence. The rating data were then divided in two ways, based on the speaker gender and sentence type. To determine whether or not the ratings were significantly different, within each subset of data, three one-way ANOVAs followed by Bonferroni corrected (Harris 2001) post-hoc *t* tests were computed: 1) on all 10 sentences; 2) on the lowest rated (i.e. least different from normal) EL speech sentence, the normal monotonous speech sentence and the normal speech sentence; and 3) on the 8 EL speech sentences.

3.5. Results

3.5.1. Scale Values

3.5.1.1 Combined data

To obtain an overview of the paired comparison data, the judgments made on all four speaker-sentence conditions (male-voiced, male-voiced/voiceless, female-voiced, female-voiced/voiceless) were combined and the resulting scale values are shown in Table 3.3.

As expected, raw EL speech received the lowest scale value while normal speech received the highest. In general, combining EL enhancements produced speech that was judged to be more normal and natural than EL speech with only one type of enhancement. The sole exception occurred for the pitch-enhanced speech (EL-P), which was ranked slightly higher than low frequency enhanced, self-noise reduced EL speech (EL-LN). This indicates that adding the proper pitch contour to EL speech would be more effective than combining the other two enhancements. This assertion is further bolstered by the presence of the pitch enhancement in the four highest ranked speech tokens. Nevertheless, the monotonous normal speech, which does not have the proper pitch contour, was judged to be more like normal natural speech than any version of EL speech.

Table 3.3: Overall Scale Values

Speech type	Rank	Scale Value
EL-raw	10	0.00
EL-L	9	0.87
EL-N	8	3.62
EL-LN	7	4.56
EL-P	6	4.85
EL-LP	5	6.42
EL-LNP	4	9.10
EL-NP	3	9.28
norm-mono	2	11.45
normal	1	14.47

Conversely, increasing the low frequency content of EL speech seems to be the least effective enhancement. On its own, it only produces a small increase in scale value (from 0 to 0.87) and when combined with the other two enhancements, it actually reduces the quality of the speech. The self-noise reduction enhancement, while not as effective as the pitch enhancement, produced a noticeable increase in EL speech quality. By itself, it produced an increase in scale value from 0 to 3.62 and when added to the pitch enhanced speech, increased the scale value from 4.85 to 9.28.

Average listener reliability was found to be 88.3% ± 8.9%

3.5.1.2 Speaker gender

The judgments were separated based on the speaker gender to examine the effect gender has on the scale values. Table 3.4 contains the resulting scale values for each speaker type. In general, the ranking of the speech types for both genders agreed with the ranking found for the pooled data, with the results for the female speaker exactly paralleling those for the combined rankings and those for the male speaker differing in two small ways. The scale values are smaller in absolute terms for both genders but this is to be expected since according to equation (3.6), the z-scores are proportional to the square root of the number of observations.

Although the absolute scale values differ somewhat, there is very little distinction between the data from the two speakers, the main discrepancy being in the scale values for EL-P and EL-LN speech tokens. For the male speaker, EL-LN speech was found to be slightly better than EL-P speech while the opposite held true for the female speaker. However, the difference in scale values is small enough to consider the two sentences similar in quality. There is also some difference between the scale values of EL-NP and EL-LNP speech for the two speakers. Whereas for the female speaker, EL-NP received a slightly larger scale value than EL-LNP (6.67 vs. 6.42), for the male speaker the associated scale values were equal.

Table 3.4: Scale Values Based on Gender of the Speaker

Male Speaker			Female Speaker		
Speech type	**Rank**	**Scale Value**	**Speech type**	**Rank**	**Scale Value**
EL-raw	10	0.00	EL-raw	10	0.00
EL-L	9	0.41	EL-L	9	0.82
EL-N	8	2.66	EL-N	8	2.47
EL-P	7	3.07	EL-LN	7	3.29
EL-LN	6	3.16	EL-P	6	3.79
EL-LP	5	4.11	EL-LP	5	4.96
EL-LNP	4	6.45	EL-LNP	4	6.42
EL-NP	3	6.45	EL-NP	3	6.67
norm-mono	2	8.19	norm-mono	2	8.00
normal	1	10.09	normal	1	10.37

3.5.1.3 All Voiced vs. Voiced-Voiceless Phonemic Context

The listener data were also sorted according to whether the judgments were based on the sentence comprised of all voiced phonemes or the one comprised of both voiced and unvoiced phonemes (see Table 3.5). Separating the observations in this fashion reveals a clear difference in the rank ordering and scale values assigned to EL enhancements for the two types of sentences. In general, the scale values for most of the EL enhancements are higher for the all-voiced sentence as compared to the voiced-voiceless sentence. Of particular note is the much lower ranking and scale value for EL-P (pitch) enhancement for the voiced-voiceless sentence, even though pitch is ultimately included in the combined enhancements that were ranked and scaled as the best three for both sentence types.

Table 3.5: Scale Values Based on Phonemic Content.

All-Voiced Sentence			Voiced/Voiceless Sentence		
Speech type	Rank	Scale Value	Speech type	Rank	Scale Value
EL-raw	10	0.00	EL-raw	10	0.00
EL-L	9	0.57	EL-L	9	0.66
EL-N	8	2.21	EL-P	8	2.59
EL-LN	7	2.94	EL-N	7	2.91
EL-P	6	4.27	EL-LN	6	3.51
EL-LP	5	5.50	EL-LP	5	3.57
EL-LNP	4	6.64	EL-NP	4	5.69
EL-NP	3	7.43	EL-LNP	3	6.23
norm-mono	2	7.46	norm-mono	2	8.73
normal	1	10.40	normal	1	10.06

3.5.1.4 Individual Speaker/Sentences cases

Separating the data further into the individual speaker/sentence cases is useful for examining both how the rankings of the versions of a single sentence differ between the two speakers and for examining how the rankings of the versions of the two sentence types differs within a single speaker. The separated data are shown in Table 3.6. The data are separated horizontally by speaker gender and separated vertically by sentence type.

This further separation of the data reveals that for both speakers, the ranking of the different sentences were again dependent on the sentence presented. The rankings of the EL speech versions of the all voiced sentence were generally higher than those of the versions of voiced/voiceless sentence. Consequently, the normal-monotonous and normal speech version received higher rankings for the all voiced sentence. There was also a difference in the rank order of the speech type for both speakers, and for both speakers the sentences with the pitch enhancement generally did better for the all voiced sentence than for voiced/voiceless sentence. These differences are illustrated by examining the difference in rank and scale value for EL-NP speech produced by the female speaker. For the voiced/voiceless sentence, EL-NP only attained a scale value of 4.29 and was ranked below both normal monotonous and EL-LNP speech whereas for the all voiced sentence, EL-NP was the most highly ranked EL speech version (in fact ranked higher than normal monotonous speech) with a scale value of 5.14.

The gender of the speaker appears to have little effect on the ranking and scale values of the different speech types; the ranks and scale values for the versions of each speech type were very similar for both speakers with only two notable exceptions. For the voiced/voiceless sentence, EL-LP speech is ranked higher for the male speaker than for the female speaker. For the all voiced sentence, EL-NP speech was judged to be more like normal natural speech than monotonous normal speech for the female speaker whereas the converse was true for the male speaker.

Table 3.6: Scale Values Based on Sentence Type and Speaker Gender

Male speaker – Voiced/Voiceless			Male speaker – All Voiced		
Speech type	**Rank**	**Scale Value**	**Speech type**	**Rank**	**Scale Value**
EL-raw	10	0.00	**EL-raw**	10	0.00
EL-L	9	0.13	**EL-L**	9	0.45
EL-P	8	1.21	**EL-N**	8	1.61
EL-LP	7	1.83	**EL-LN**	7	1.97
EL-N	6	2.15	**EL-P**	6	3.13
EL-LN	5	2.50	**EL-LP**	5	3.98
EL-NP	4	3.76	**EL-LNP**	4	4.96
EL-LNP	3	4.16	**EL-NP**	3	5.37
norm-mono	2	6.04	**norm-mono**	2	5.55
normal	1	6.84	**normal**	1	7.42
Female speaker – Voiced/Voiceless			**Female speaker - All Voiced**		
Speech type	**Rank**	**Scale Value**	**Speech type**	**Rank**	**Distance**
EL-raw	10	0.00	**EL-raw**	10	0.00
EL-L	9	0.80	**EL-L**	9	0.36
EL-N	8	1.97	**EL-N**	8	1.52
EL-P	7	2.46	**EL-LN**	7	2.19
EL-LN	6	2.46	**EL-P**	6	2.91
EL-LP	5	3.22	**EL-LP**	5	3.80
EL-NP	4	4.29	**EL-LNP**	4	4.43
EL-LNP	3	4.65	**norm-mono**	3	5.01
norm-mono	2	6.31	**EL-NP**	2	5.14
normal	1	7.38	**normal**	1	7.29

3.5.1.5 Reliability of least squares estimate

The computed scale values are only estimates of the true locations of the different speech stimuli on the psychological continuum of natural normal speech and thus it is useful to measure the reliability of these estimates. Kramer and Serlin (1979) suggest the following measure of reliability:

$$r^2 = \frac{\sum_{i \neq j}\sum (S_i - S_j)}{\sum_{i \neq j}\sum z_{ij}^2} \qquad (3.7)$$

where S_i, S_j are the computed scale values and the z_{ij} are the measured z-scores discussed in section 3.4.1. r^2 is bounded between 0 and 1 with values closer to 1 indicating a better least-squares fit. The r^2 values computed for the different sets of scale values ranged from 0.854 to 0.870 with a mean of 0.863 ± 0.005 indicating that the least-squares model accurately fits the measured data.

The analysis used in this study assumed that Thurstone's Case V (all discriminal dispersions are equal) was applicable to the data. It was believed since the stimuli were all versions of the same sentence that it was not unreasonable to make this assumption. Of the 10 speech types, the one most likely to have a unique discriminal dispersion was normal speech as it likely that its presentation doesn't elicit a great deal of variation on the scale of sounding like normal natural speech. If the discriminal dispersion of normal speech were indeed different, the effect on the overall ranking would be minimal and confined to the scale value of the normal speech (Mosteller 1951). Since the scale values of the normal speech were always found to be considerably greater than those of the other speech types, it is likely that equal discriminal dispersion assumption had little effect on the scaling results. Furthermore, the high r^2 values suggest that the least squares method, which assumes equal discriminal dispersions, produces an accurate fit to the data.

3.5.2. Visual Analog Scale (VAS) Ratings

When considering the visual analog ratings, it is important to keep in mind that opposite to the paired comparison data, a lower rating indicates that a speech token is more like normal natural speech.

3.5.2.1 Combined data.

The overall mean VAS ratings for the ten sentences are shown in Table 3.7 along with the corresponding standard errors of the mean, σ_m.

The ranking of the speech types is very similar to the ranking obtained from the paired comparison data except for the reversal of the order of the EL-N and EL-LN sentences. The normal speech token is found to be the least different from normal natural speech, and is closely followed by the normal-monotonous token. Once again, the EL-NP speech token was found to have the best rating, although the difference in rating between this token and that of normal speech is far greater (in relative terms) than the difference in scale values found from the paired comparisons data.

Table 3.7: Overall VAS Ratings

Speech type	Rank	Rating	σ_m
EL-raw	10	8.40	0.20
EL-L	9	8.08	0.19
EL-LN	8	7.93	0.11
EL-N	7	7.89	0.13
EL-P	6	7.11	0.11
EL-LP	5	6.98	0.11
EL-LNP	4	6.50	0.10
EL-NP	3	6.20	0.10
norm-mono	2	1.76	0.08
Normal	1	0.09	0.02

A 3-way ANOVA was performed on the entire data set using the three factors: speaker gender, phonemic content, and sentence type. Significant differences ($p < 0.0001$) were found for all three main effects with smaller (closer to normal speech) average scale values found for male spoken and all-voiced sentences. The overall mean ratings and standard errors for the male and female speakers were 5.98 ± 0.05 and 6.282 ± 0.05 respectively. The mean rating and standard error of the voiced sentence was found to be 5.657 ± 0.052 while that of the voiced/voiceless sentence was computed as 6.604 ± 0.048. The expected overall ordering of average values across the 10 levels of the speech type factor was observed with normal speech having the lowest scale value and EL-raw having the highest (farthest from normal speech) scale value. Significance was also found for the interactions speech type*gender (F=4.1, $p < 0.01$) and speech type*phonemic content (F=59.2, $p < 0.01$). The mean ratings for each sentence type with the corresponding standard errors, separated by both speaker gender and phonetic content are plotted in Figure 3.3.

The curves in the left hand plot in Figure 3.3 look very similar; however the ratings for the male speaker were always slightly lower than the corresponding ratings for the female speaker (except for the normal speech sentence). The differences in ratings based on phonemic content are more pronounced as can be seen in the right hand plot. The voiced sentences were always rated better than their mixed sentence counterparts except in the cases of the normal-monotonous and normal sentences. Although both interactions were found to be statistically significant, the greater F value for the speech type*phonemic content interaction further illustrates it is the larger of the two interactions.

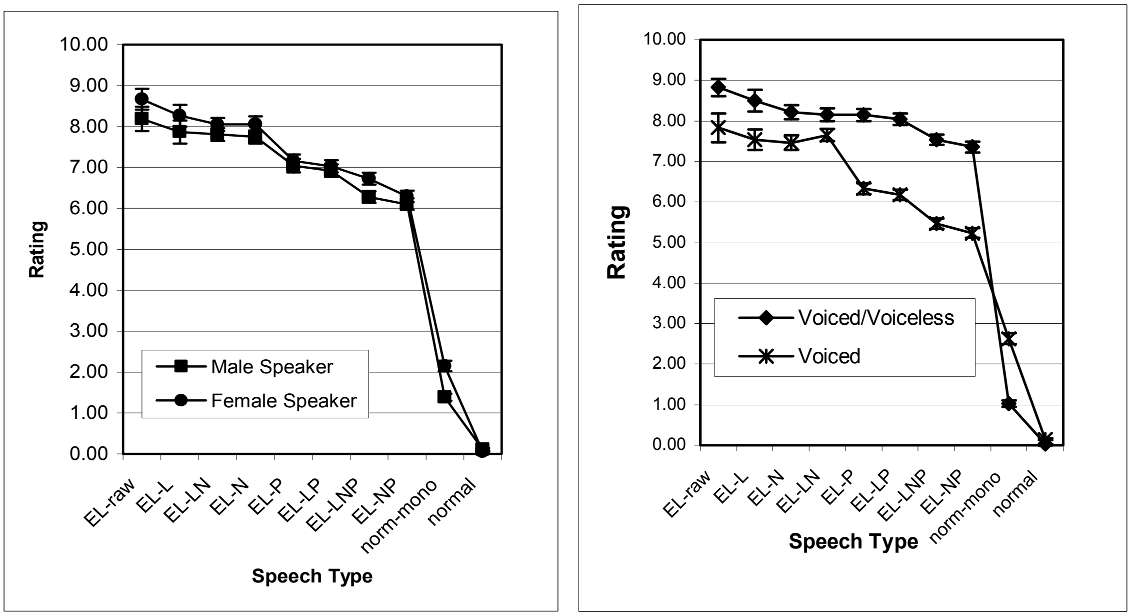


Figure 3.3. The mean and standard error of the ratings for the different sentence types separated by gender (left) and phonemic content (right). Separating the ratings by gender shows that the ratings for the male speaker are consistently lower than those for the female speaker. Separating the data by phonemic content shows that the ratings for the voiced sentence were significantly lower except for the normal monotonous sentence.

A further investigation of the data was conducted by performing 3-way ANOVAs on two subsets of the data: one comprised of all the EL speech versions and one comprised of the normal, normal-monotonous, and lowest rated EL speech versions (EL-NP). The ANOVA performed on the first subset revealed significant differences for all the speech type and phonemic content effects at the $p < 0.000$ level and the speaker gender effect at the $p < 0.01$ level. However, while the speech type*phonemic content interaction was found to be significant ($p < 0.000$), the speech type*speaker gender interaction was not ($p = 0.58$). The ANOVA performed on the second data subset demonstrated significant differences for the sentence type and speaker gender ($p < 0.000$) but not for phonemic content ($p = 0.051$). Both types of interactions were found to be significant ($p < 0.001$).

These findings help clarify some of the results found by the statistical analysis performed on the entire data set. The difference in significance of the speech type*speaker gender interaction between the two data subsets indicates that the significance of this interaction for the entire data set is primarily due to speaker gender based differences between the ratings of the normal, normal-monotonous and EL-NP speech tokens. Figure 3.4, which plots the ratings of the three different sentences for both speakers shows that the greatest discrepancy in ratings is for the normal-monotonous speech. The rating for the female speaker (2.15 ± 0.13) was almost double that for the male speaker (1.38 ± 0.09).

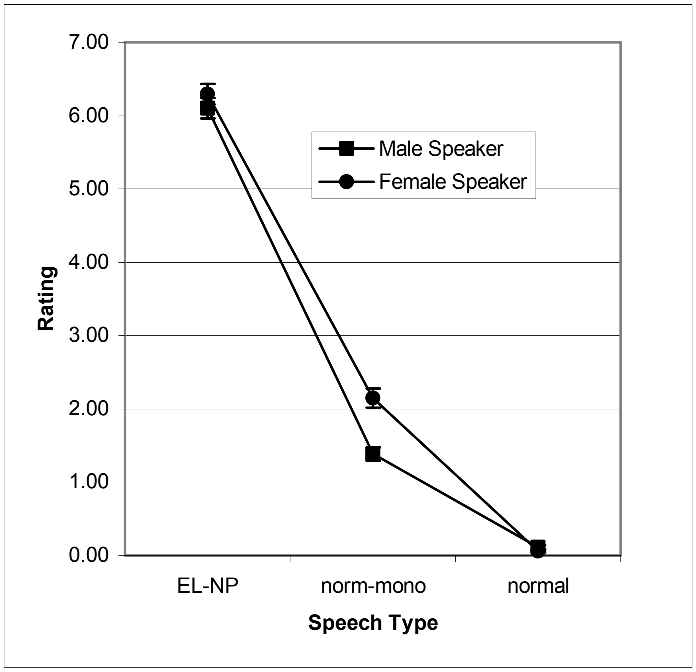


Figure 3.4. The mean and standard error of the ratings for EL-NP, normal-monotonous, and normal speech sentences separated by speaker gender. The ratings of the normal-monotonous speech sentence display the greatest dependence on speaker gender.

Additional one-way ANOVA and post-hoc *t* tests (Bonferroni corrected, $p < 0.01$) demonstrated that the ratings for the normal, normal monotonous and EL-NP speech types are all significantly different (p<0.01) from each other. Post-hoc tests also indicated that: 1) the ratings for the 4 highest rated speech types (EL-raw, EL-L, EL-LN, EL-N) did not differ significantly from each other, 2) the ratings for the EL-P and EL-LP speech types differed from the 4 highest rated speech types but not from each other, and 3) while the EL-LNP rating was significantly different from the 5 highest rated speech types, it was not different from those of the EL-LP and EL-NP speech types.

3.5.2.2 Speaker gender

The mean ratings and standard errors of the ratings separated by speaker gender are displayed in Table 3.8. These data are the same as those plotted in the left half of Figure 3.3. The rank orders for both speakers are very similar except for the reversal in order of the EL-N and EL-LN speech types. However, difference between the EL-N and EL-N ratings for the female speakers is small enough (0.01) that they may be considered equivalent. The similarity of the rank orders is further supported by the lack of speech type*speaker gender interaction for the 8 EL speech types described above. Additionally, as was true for the combined data, the ratings for the normal-monotonous and normal speech types were much lower than the lowest rated EL speech sentence.

Table 3.8. VAS ratings separated by speaker.

Male Speaker					**Female Speaker**				
Speech type	**Rank**	**Rating**	**σ_m**	**N**	**Speech type**	**Rank**	**Rating**	**σ_m**	**N**
EL-raw	10	8.19	0.30	39	**EL-raw**	10	8.67	0.25	32
EL-L	9	7.87	0.28	52	**EL-L**	9	8.26	0.26	58
EL-LN	8	7.81	0.15	139	**EL-N**	8	8.06	0.19	110
EL-N	7	7.75	0.16	123	**EL-LN**	7	8.05	0.16	136
EL-P	6	7.05	0.17	136	**EL-P**	6	7.16	0.16	152
EL-LP	5	6.92	0.15	169	**EL-LP**	5	7.03	0.15	189
EL-LNP	4	6.28	0.14	243	**EL-LNP**	4	6.73	0.14	235
EL-NP	3	6.10	0.14	243	**EL-NP**	3	6.29	0.14	243
norm-mono	2	1.38	0.09	298	**norm-mono**	2	2.15	0.13	285
normal	1	0.11	0.03	358	**normal**	1	0.06	0.01	360

One-way ANOVAs and Bonferroni corrected post-hoc *t* tests ($p < 0.01$) conducted on both sets of data indicated that for both speakers, 1) the normal, normal-monotonous and EL-NP speech types are significantly different from each other, 2) the ratings for the 4 highest rated speech types were not significantly from each other, 3) the EL-NP rating was not significantly different from those of the EL-LP and EL-LNP types, and 4) while the EL-LNP rating was significantly different from the 4 highest rated speech sentences, it was not different from those of EL-LP and EL-P. The two disparities in these statistical results were centered on the EL-P and EL-LP speech types. While for the male speaker, the rating of EL-P was did not significantly differ from the ratings of the 4 higher rated speech types, for the female speaker, it did differ from the ratings of EL-raw and EL-LN. Moreover, for the female speaker the EL-LP rating was also different from those of the 4 highest rated sentence types, while for the male speaker, it only differed from the EL-LN rating. The seemingly unintuitive notion of a lower rated speech type to not differ from one speech type and yet differ from another speech type with a more similar rating can be explained by the disparity in number of times the sentence types were rated. The experiment was designed such that VAS ratings were made only on speech tokens that were judged to sound more like normal natural speech in the paired comparison task. Therefore, by design, the least natural sounding speech tokens were rated less often. The smaller number of observations produces larger standard errors, thereby preventing the higher rated speech tokens from being significantly different from lower ones. Thus, for the female speaker, even though the distance between EL-P rating was much larger than the EL-L rating than the EL-LN rating, the number of fewer number of EL-L ratings (58) precluded it from being significantly different from EL-P.

3.5.2.3 Phonemic content

Table 3.9 displays the rating data separated by the phonemic content of the sentence. These are the same data found in the right hand plot of Figure 3.3. The rank orders of the speech types are again similar except for the EL-LN condition, which attained a lower rank for the voiced/voiceless sentence. There is however, a noticeable difference between the speakers for the numerical ratings assigned to each sentence type. Every speech type received a higher rating for the voiced/voiceless sentence than for the all voiced sentence. This difference, which is supported by the significant sentence type*phonemic content interaction discussed earlier, is punctuated by the fact that the lowest rated EL speech type for the voiced/voiceless sentence (EL-LN) received a rating that was only lower than the 4 highest rated sentences for the all voiced sentence.

Table 3.9. VAS ratings separated by phonemic content.

All Voiced sentence					**Voiced/Voiceless sentence**				
Speech Type	**Rank**	**Rating**	**σ_m**	**N**	**Speech Type**	**Rank**	**Rating**	**σ_m**	**N**
EL-raw	10	7.83	0.36	30	**EL-raw**	10	8.83	0.21	41
EL-LN	9	7.65	0.15	123	**EL-L**	9	8.50	0.27	62
EL-L	8	7.53	0.25	48	**EL-N**	8	8.22	0.17	133
EL-N	7	7.46	0.18	100	**EL-LN**	7	8.15	0.16	152
EL-P	6	6.33	0.14	165	**EL-P**	6	8.15	0.15	123
EL-LP	5	6.17	0.13	204	**EL-LP**	5	8.04	0.14	154
EL-LNP	4	5.47	0.13	240	**EL-LNP**	4	7.54	0.12	238
EL-NP	3	5.23	0.12	265	**EL-NP**	3	7.36	0.13	221
norm-mono	2	2.63	0.13	266	**norm-mono**	2	1.03	0.07	317
normal	1	0.15	0.03	359	**normal**	1	0.02	0.01	359

One-way ANOVAs and Bonferroni corrected post-hoc *t* ($p < 0.01$) tests conducted on both sets of data indicated that for both sentences the normal, normal-monotonous and EL-NP speech types are significantly different from each other. However, in general, these tests showed that whereas several speech types received significantly different ratings in the voiced sentence case, this was only true for two sentence types (EL-NP and EL-LNP) in the voiced/voiceless sentence case. Thus, while EL-P, EL-LP, EL-LNP, and EL-NP were found to be significantly different from the 4 highest rated sentence types for the voiced sentence, only EL-NP was significantly different in this manner for the mixed sentence.

3.5.2.4 Individual Speaker/Sentences cases

As was done for the paired comparison data, the visual analog scale ratings were separated into the four speaker/sentence cases. These data are shown in Table 3.10. An inspection of the ranks of the different speech types reveals that although each speaker/sentence condition produced a different rank order, the four lowest ranked speech types were always normal, normal-monotonous, EL-NP and EL-LNP speech tokens. For all four cases, the normal and normal-monotonous sentences were rated much lower than all of the EL-speech sentences. A one-way ANOVA and Bonferroni corrected post-hoc *t* tests ($p < 0.01$) showed that the normal, normal-monotonous and EL-NP sentence types were all significantly different from each other.

Looking across speaker gender it can be seen that for the voiced/voiceless sentence, the orderings of the speech types are very different, except for the four lowest rated conditions. Surprisingly in the case of the male speaker, the EL-LP and EL-P were rated higher than three speech types that did not have any pitch enhancement. However, a one-way ANOVA followed by the Bonferroni corrected post hoc *t* tests indicated that none of the EL speech ratings were statistically different from each other at least at $p < 0.01$. (At $p < 0.05$ EL-NP was significantly different from the EL-raw and EL-LP speech types.) These statistical tests also indicated that for the female speaker, none of the 7 highest rated speech types were significantly different from each other and that EL-NP was only different from the 4 highest rated speech types. Thus, even though the rankings differ between the two speakers, the lack of statistical significance indicates that these differences aren't that meaningful.

A similar situation is found for the all voiced sentence, except that in this case, only the rankings of the highest 4 speech types differ between the two speakers. Again, the one-way ANOVA followed by the Bonferroni corrected *t* tests ($p < 0.01$) show that for both speakers, the 4 most highly rated speech types are not significantly different from each other ($p = 1$). It was also revealed that for both speakers, EL-NP was different from all EL speech types except for EL-LNP.

For the other speech types the extent of the significance of the differences between them diverged for the two speakers. The following was found for the other speech types: 1) for the female speaker, EL-P was significantly different from EL-raw and EL-LN while for the male speaker, EL-P only differed from EL-LN; 2) EL-P differed only from EL-LN in the male speaker case but differed from both EL-LN and EL-raw in the female speaker case; and 3) EL-LNP differed from the 4 highest rated speech types in the male speaker and differed from the 5 highest in the female speaker.

Table 3.10. VAS ratings separated by sentence type and speaker gender.

Male speaker – Voiced/Voiceless Sentence					Male speaker – All Voiced sentence				
Speech type	**Rank**	**Rating**	**σ_m**	**N**	**Speech type**	**Rank**	**Rating**	**σ_m**	**N**
EL-raw	10	8.58	0.29	26	**EL-LN**	10	7.81	0.21	57
EL-LP	9	8.23	0.18	67	**EL-L**	9	7.70	0.36	23
EL-P	8	8.11	0.25	53	**EL-N**	8	7.51	0.25	49
EL-L	7	8.00	0.42	29	**EL-raw**	7	7.40	0.65	13
EL-N	6	7.90	0.21	74	**EL-P**	6	6.36	0.19	83
EL-LN	5	7.80	0.22	82	**EL-LP**	5	6.06	0.17	102
EL-LNP	4	7.34	0.18	119	**EL-LNP**	4	5.26	0.18	124
EL-NP	3	7.26	0.19	110	**EL-NP**	3	5.14	0.17	133
norm-mono	2	0.57	0.06	161	**Norm-mono**	2	2.34	0.15	137
normal	1	0.04	0.03	179	**normal**	1	0.18	0.05	179
Female speaker Voiced/Voiceless Sentence					**Female speaker – All Voiced Setence**				
Speech type	**Rank**	**Rating**	**σ_m**	**N**	**Speech type**	**Rank**	**Rating**	**σ_m**	**N**
EL-raw	10	9.25	0.26	15	**EL-raw**	10	8.15	0.38	17
EL-L	9	8.95	0.33	33	**EL-LN**	9	7.50	0.22	66
EL-N	8	8.61	0.26	59	**EL-N**	8	7.42	0.26	51
EL-LN	7	8.57	0.21	70	**EL-L**	7	7.36	0.36	25
EL-P	6	8.17	0.18	70	**EL-P**	6	6.29	0.2	82
EL-LP	5	7.90	0.2	87	**EL-LP**	5	6.29	0.19	102
EL-LNP	4	7.74	0.17	119	**EL-LNP**	4	5.69	0.19	116
EL-NP	3	7.45	0.18	111	**EL-NP**	3	5.32	0.17	132
norm-mono	2	1.50	0.13	156	**Norm-mono**	2	2.93	0.22	129
Normal	1	0.01	0	180	**normal**	1	0.11	0.02	180

There were also differences in the ratings between the two sentence types for both speakers. Again, the ratings for the versions of all voiced sentence were always lower than those for the versions of the voiced/voiceless sentence, except in the case of the normal-monotonous condition. Moreover, while the 4 lowest rated speech types for the male (5 for the female) were the same for both types of sentences, the rankings of the rest of the speech types were not. However, as the statistical tests performed on this data demonstrated, the difference in ratings of these higher ranked sentences were not statistically significant, thus minimizing the importance of this discrepancy in rank order.

3.6. Discussion

This comprehensive perceptual study was conducted to better quantify the sources and perceptual impact of abnormal acoustic properties typically found in EL speech. This was done by determining the relative contributions that a set of proposed acoustic enhancements makes towards improving the quality of EL speech. The ultimate goal is to use these results to efficiently direct an effort to improve the quality of EL speech. The results of this study indicate that of the three properties selected, the lack of pitch information contributes the most to EL speech's poor quality. With only a few exceptions, the pitch enhancement, both on its own or in combination with another enhancement was consistently found in the four top rated EL speech tokens for both the paired comparison and the VAS procedures. Next in importance is the competing self-noise followed by the lack of low frequency energy. Thus, in designing an improved EL device, one would gain the most benefit by somehow providing the users with a means of pitch control while removing, or at least reducing the amount of self-noise it generates. Based on the results obtained here, enhancing the low frequency content, at least in the manner described by Qi and Weinberg (1991) can actually reduce the quality of the resulting speech in some circumstances.

Although the results of the paired comparison experiment indicate that using the pitch and noise reduction enhancements can make a substantial improvement in EL speech quality, the analog scaling results demonstrate that even the best enhanced version of EL speech still has significantly degraded quality when compared to normal natural speech. In fact, not one version of EL speech received a mean rating lower than 5 (i.e. the half way point) while both the normal and monotonous normal tokens consistently received ratings below 2.5. Initially, the paired comparison and VAS results may appear somewhat contradictory especially in terms of how similar the EL-NP and normal-monotonous sentences are to each other. However, a detailed explanation of how the scale values are computed helps explain how this discrepancy is inherent in the algorithm used to make these computations.

Table 3.11 is the *P*-matrix of the entire data set and contains the proportion of times a stimulus in column *k* is judged to sound more like normal natural speech than the stimulus in column *j*. As described in Section 3.4.1, the diagonal entries are left blank. Using Eq. 3.6, this matrix is converted to the corresponding *Z*-matrix, which, along with the raw and shifted mean scale values, is shown in Table 3.12. The diagonal entries of the *Z*-matrix are set to zero.

Table 3.11. The P-Matrix for the entire data set.

	EL-raw	EL-L	EL-N	EL-LN	EL-P	EL-LP	EL-LNP	EL-NP	norm-mono	normal
EL-raw		0.625	0.925	0.9375	0.8375	0.9375	0.9125	0.95	0.9875	1
EL-L	0.375		0.7625	0.925	0.8375	0.85	0.9375	0.9375	1	1
EL-N	0.075	0.2375		0.65	0.6875	0.75	0.8625	0.8625	0.9625	1
EL-LN	0.0625	0.075	0.35		0.6625	0.7125	0.8875	0.85	0.9625	1
EL-P	0.1625	0.1625	0.3125	0.3375		0.7	0.9125	0.9625	0.85	1
EL-LP	0.0625	0.15	0.25	0.2875	0.3		0.7875	0.8375	0.85	1
EL-LNP	0.0875	0.0625	0.1375	0.1125	0.0875	0.2125		0.5125	0.8125	1
EL-NP	0.05	0.0625	0.1375	0.15	0.0375	0.1625	0.4875		0.8375	1
Norm-mono	0.0125	0	0.0375	0.0375	0.15	0.15	0.1875	0.1625		0.975
normal	0	0	0	0	0	0	0	0	0.025	

Table 3.12. The Z-Matrix for the entire data set with corresponding scale values.

	EL-raw	EL-L	EL-N	EL-LN	EL-P	EL-LP	EL-LNP	EL-NP	norm-mono	normal
EL-raw	0.00	2.24	7.60	7.83	6.04	7.83	7.38	8.05	8.72	8.94
EL-L	-2.24	0.00	4.70	7.60	6.04	6.26	7.83	7.83	8.94	8.94
EL-N	-7.60	-4.70	0.00	2.68	3.35	4.47	6.48	6.48	8.27	8.94
EL-LN	-7.83	-7.60	-2.68	0.00	2.91	3.80	6.93	6.26	8.27	8.94
EL-P	-6.04	-6.04	-3.35	-2.91	0.00	3.58	7.38	8.27	6.26	8.94
EL-LP	-7.83	-6.26	-4.47	-3.80	-3.58	0.00	5.14	6.04	6.26	8.94
EL-LNP	-7.38	-7.83	-6.48	-6.93	-7.38	-5.14	0.00	0.22	5.59	8.94
EL-NP	-8.05	-7.83	-6.48	-6.26	-8.27	-6.04	-0.22	0.00	6.04	8.94
norm-mono	-8.72	-8.94	-8.27	-8.27	-6.26	-6.26	-5.59	-6.04	0.00	8.50
normal	-8.94	-8.94	-8.94	-8.94	-8.94	-8.94	-8.94	-8.94	-8.50	0.00
Raw scale values	-6.46	-5.59	-2.84	-1.90	-1.61	-0.04	2.64	2.82	4.99	8.01
Shifted Scale values	0.00	0.87	3.62	4.56	4.85	6.42	9.10	9.28	11.45	14.47

Because the *Z*-matrix is full (i.e. there are no cell entries of infinity or negative infinity), the Kramer-Serlin method reduces to simply averaging the columns of the matrix, the averages the result of which is shown in the second to last row of Table 3.11. The averages are then shifted so that the lowest scale value is set to zero. The averaging implies that the scale values are considerably dependent on how one stimulus rates against *all* of the other stimuli. Consider the cases of the EL-NP and normal-monotonous speech tokens, which are judged to be much better than all of the other speech types and much worse than normal speech. In most cases, the entries in the cells of the respective columns of the *P* and *Z* matrices are very similar. Conversely, the cells corresponding to the direct comparison between the two stimuli are quite different. When paired with EL-NP speech, normal-monotonous speech was judged to be more like normal natural speech 83.75% of the time. If the *Z*-matrix were reduced to only the EL-NP and normal-monotonous cells, it would resemble the matrix in Table 3.13, and produce more disparate scale values.

Table 3.13. A Reduced Z-matrix

	EL-NP	norm-mono
EL-NP	0.00	6.04
norm-mono	-6.04	0.00
Raw scale values	-3.02	3.02
Shifted Scale values	0.00	6.04

Moreover, according to Eq. 3.6 the *z*-scores are proportional to $\sqrt{N}$, where N is the number of observations made. Therefore a larger number of observations produces larger *z*-scores and hence larger scale values. So, for example, if stimulus *j* were always rated better than stimulus *k* in 10 observations, the *z*-scores for *j* and *k* would be –3.16 and 3.16 respectively. If the same pattern occurred for 100 observations, the *z*-scores would be -10 and 10. Therefore the limits of the scale values are dependent on the number of observations made and the scale values of all the stimuli must be proportionally scaled within these limits.

Consequently, because the extent of the scale is limited by the total number of observations that were made and because a large and because several stimuli were used in the experiment, the scale values of EL-NP speech and normal-monotonous speech are not as different as they would be had a they been the only two stimuli used and if a larger number of observations were made.

On the other hand, the average rating for a speech token on the visual analog scale is mathematically independent of those made for other speech tokens and of the number of ratings made on each stimulus. Thus the visual analog scale gave the listeners enough flexibility to in effect create an unbalanced scale whereby the ordinal ranking of the speech types were the same as it was for the paired comparison data but with a more compressed set of ratings for the EL speech types

Both the paired comparison and VAS data suggest that these enhancements are not as effective for speech that contains unvoiced phonemes, further limiting the improvement in quality. The pitch contour extracted from the normal speech versions of the voiced/voiceless sentence contained gaps corresponding to the unvoiced parts of the sentence. Prior to being added to the pitch enhanced EL versions, the pitch values within these gaps were set equal to the last measured pitch value thus creating a flat pitch contour for a short period of time. As such, the pitch contour estimate used in the voiced/voiceless sentence was not as accurate as the one used in the all voiced sentence perhaps limiting the effectiveness of the pitch enhancement. However, this reasoning cannot satisfactorily explain the difference in ratings between the non-pitch enhanced EL speech sentences. All of the EL versions of the voiced/voiceless lacked the proper perceptual cues for unvoiced consonants, a problem inherent in electrolarynx speech (Weiss *et al.* 1979, Weiss and Basili 1985). It is likely that this missing information contributed to the reduced ratings of the EL versions of the voiced/voiceless sentence.

While adding pitch information may be the most effective means of improving EL speech quality it is perhaps the most difficult enhancement to implement because it requires finding a way of estimating what pitch the speaker intends to use. In one attempt to provide EL users with pitch control, only 2 of the 16 subjects studied were able to master the control of the device and thereby produce pitch contours that resembled those in normal speech (Uemi *et al.* 1994). And although, EL-LN speech received a scale value similar to that of EL-P speech, the sliding scale results indicated that EL-LN wasn't significantly different from raw EL speech.

However, the fact that normal monotonous speech more closely approximates the quality of normal natural speech than any type of EL speech enhancement (including ones with the proper pitch information) provides some hope that EL speech can be significantly improved without having to add prosodic information. It also suggests that there are other, unexplored properties of EL speech that contribute to its unnatural quality. For example, the limited effectiveness of the three enhancements on speech with unvoiced phonemes suggests that lack of voicing information is another important EL speech property that reduces its quality. Perhaps a reasonable intermediate goal would be to identify and correct other aberrant properties of EL speech that enable a closer approximation to normal monotonous speech. Therefore, the following chapters are devoted to identifying and exploring such additional aberrant properties.

4. Acoustic analysis of pre- and post-laryngectomy speech.

4.1. Introduction

The results of the paired comparison experiments described in the previous chapter show that removing the direct noise from and adding the proper pitch information to EL speech can lead to a substantial improvement in EL speech quality. Yet, according to the subsequent visual analog scale results, there is still a significant gap between the quality of normal speech and that of the best enhanced EL speech. More importantly, however, monotonous (i.e. fixed pitch) normal speech was still judged to sound more like normal natural speech than any version of enhanced EL speech. This suggests that there are other factors (besides the three studied in the preceding chapter) that contribute to the unnatural quality of EL speech.

In order to investigate other properties of EL speech that contribute to its unnatural quality, it is helpful to analyze and compare normal and EL speech. Primarily concerned with intelligibility issues, Weiss *et al.* (1979) conducted such a study that explored both the acoustical and perceptual characteristics of EL speech. Their perceptual experiments demonstrated that the greatest deficit in intelligibility was due to initial stop consonant confusion. In addition, their acoustical analysis revealed the presence of the low frequency spectral deficit as well as a significant amount of direct noise corruption. It is well known that these two acoustic properties contribute to the poor quality of EL speech, yet the perceptual studies described in the previous chapter show that correcting these problems only results in a limited improvement in quality. Thus, once again, a more thorough analysis is needed to look for other attributes of EL speech that contribute to its unnatural quality.

The Weiss *et al.* study used both normal and EL sentences spoken by a group of normal (i.e. non-laryngectomized) speakers. Although using of such a group is valuable for isolating surgery induced changes from the speech, comparing pre- and post-laryngectomy speech could potentially be more informative since laryngectomy patients are by far the primary users of the EL. Qi and Weinberg (1991) took such an approach, but because it is rare to find pre-laryngectomy speech recordings of current EL users, they were forced to compare average normal speaker data with average EL speaker data

The most effective method for looking at the differences between normal and EL speech would be to compare an EL user's EL speech with his/her normal speech from before being laryngectomized. Hillman *et al.* (1998) recorded laryngeal cancer patients both before and after treatment as part of a large multi-institutional study carried out by the Cooperative Studies Program at the Veterans Administration (VA-CSP#268). In both cases, speech recordings consisting of sustained vowels, reading of a standard passage, a verbal description of a picture, and reading a randomized list of 50 phrases (carrier phrase with different target words) were made pre and post treatment. Post-treatment recordings were made at regular follow up visits after treatment. This data set provides a unique opportunity to study all of the changes that occur in an individual's speech when he/she is forced to migrate from normal speech to EL speech and eliminates the inter-speaker variability incurred when comparing one normal speaker with another EL speaker. Moreover, not only can general trends be found in differences between normal and EL speech, but individual variability can be separated from these common differences as well. This chapter discusses the results of analyzing this corpus of pre- and post-laryngectomy speech data.

4.2. Methods

4.2.1. Description of VA-CSP268 Recordings

A total of 332 subjects who were all diagnosed with stage III or IV squamous cell carcinoma of the larynx were used in the study. The subjects were randomly assigned to one of two treatment groups: 166 underwent laryngectomy surgery followed by radiation therapy and 166 received chemotherapy followed by radiation. Those patients who underwent laryngectomy surgery were offered instruction in esophageal or electrolaryngeal speech or, in some cases, the surgical implantation of a tracheoesophageal shunt. All subjects were functionally assessed before treatment and at 1,6, 12, 18 and 24 months afterwards. Some subjects continued to be assessed beyond the 24 month interval, sometimes lasting up to 60 months.

At each assessment session, recordings were made of the following tasks: 1) the sustained production of three vowels: /a/ (as in "bot"), /i/ (as in "beet"), and /u/ (as in "boot"); 2) the reading of The Zoo Passage (see Appendix C); 3) a description of a picture, The Cookie Jar Picture (Goodglass and Kaplan 1983); and 4) reading a randomized list of 50 words, each contained in a carrier phrase. All recordings were done in a quiet environment and made on an audio cassette using a Marantz model 220 recorder and a Radio Shack model 33-1071 microphone, situated 6 to 12 inches from the subjects. This uniform protocol was used at the15 participating Veteran's Administration Hospital throughout the United States. The recordings were analyzed both in terms of intelligibility and acoustic properties. Of particular interest to this study, estimates of the amount of spectral noise present in the pre-laryngectomy speech were made. A rating between 1 and 5 was given to the recorded speech based on this estimate with a higher rating indicating more noisy speech.

4.2.2. Subject selection and data preparation

Of the 166 laryngectomy patients in this study, 13 were found to have had good normal (i.e. pre-treatment) speech and to have spoken with a neck type EL after laryngectomy surgery. All 13 of these subjects received a spectral noise rating of 3 or below and the quality of their speech was subjectively verified to approximate normal speech by the author. Of these 13 subjects, 9 of them were used in this analysis, while the remaining 4 were set aside to be used in testing of future developed enhancement algorithms. The 9 subjects (all male) were recorded at 7 different VA hospitals and used 3 different EL devices. Table 4.1 contains the location of the hospital and the type of EL device used:

Table 4.1 Location of VA Subjects and Type of EL Used

Subject:	Location	EL Used
1	Boston, MA	Servox
2	Buffalo, NY	Servox
3	Buffalo, NY	Servox
4	Dallas, TX	Romet
5	Dallas, TX	Romet
6	Allan Park, MI	Servox
7	East Orange, NJ	Aurex
8	Tampa, FL	Servox
9	Tucson, AZ	Servox

Of the several post treatment recordings that were made for each subject, only the final EL speech recordings were used in this study. It often takes speakers a long time to learn to the master speaking with an EL and thus it was believed that the final EL speech recordings represented each subject's best possible EL speech.

The analog recordings were digitzed as follows: a Marantz PMD 501 cassette player was connected to the input of a PC sound card (Creative Labs Sound Blaster Live! Platinum). An audio signal acquisition and editing software package (Syntrillium Software's Cool Edit 2000) was used to digitize the speech at 32 kHz.

4.2.3. Analysis

Because the focus of the measurements was on the frequency range that included the first three formants, prior to analysis, the speech was appropriately low pass filtere and then downsampled to 10 kHz. This sampling frequency was chosen because the frequency range of interest was below 5 kHz. From the running speech contained in the Zoo Passage recordings, 9 vowels were isolated for analysis: /i/ in "eat", /I/ in "sister", /ɛ/ in "get", /æ/ in "basket", /a/ in "Bob", /U/ in "took", /u/ in "zoo", /^/ in "brother", and /ɝ/ in "service." (Because of the limitations of the software used to generate the figures in this document, substitutions were used for the proper phonemic symbols. These are described in Appendix A). For each vowel, the spectrum was computed by performing a 4096-point Discrete Fourier Transform on a 50 ms Hamming windowed section of the vowel. Linear predictive (LP) coefficients were computed for this vowel segment as well. The LP order was chosen to be the smallest number that accurately captured (based on visual observation) the first three formants. Typically, the LP orders used were 14 for normal speech and 18 for EL speech. The LP coefficients were then converted into conjugate pairs of poles, p_i, which were converted into frequencies, f_i and bandwidths, bw_i by:

$$f_i = \frac{(\angle p)F_s}{\pi} \tag{4.1a}$$

$$bw_i = \frac{-\ln|p|F_s}{\pi}, \tag{4.1b}$$

where F_s is the sampling frequency. Finally, a long-term average spectrum was computed over the vowel by averaging the spectrum computed on a sliding 5 ms window with a 2.5 ms overlap. This spectrum was used in the computation of the peak-to-valley ratios which are discussed below.

The first three formant frequencies (*F1, F2, F3*) and amplitudes (*A1, A2, A3*) were measured by manually marking the approximate frequency locations of the formants and then finding the harmonic with the largest amplitude within 100 Hz of the marked frequency location. The formant bandwidths (*BW1, BW2, BW3*) were defined as the bandwidths of the conjugate pole pair whose frequency was closest to the formant frequency. The first harmonic frequency (*F0*) and its amplitude (*H1*) were computed by taking the inverse of an estimate of the pitch period obtained by using an autocorrelation of the vowel segment. Based on these measurements, the following acoustic parameters were calculated: the relative formant amplitudes (*A1-A2, A2-A3, A1-A3*), the spectral tilt (*H1-A3*), and the amplitude of the first harmonic relative to that of the first formant *(H1-A1*).

In addition, as a measure of the amount of low frequency energy contained in the vowel, the quantity, E_r, normalized low frequency energy was calculated. This was done by low pass filtering the vowel at 500 Hz using a 64-point FIR filter, and then dividing the energy in the resulting signal by the energy in full band (i.e. up to 5000 Hz) vowel signal. Moreover, because the self noise tends to fill in the spectral valleys between the formants, the peak-to-valley ratios between the first two formants (*ptvrF1F2*) and the second and third formants *(ptvrF2F3*) were measured to assess the effect of EL self-noise on EL speech. To compute *ptvrF1F2*, the minimum of the long-term average spectrum between the first two formants was marked and subtracted from the amplitude of the first formant, *A1*. Similarly, *ptvrF2F3* was defined as the difference between minimum of the long-term average spectrum between *F2* and *F3* and the amplitude of the second formant, *A2*. Finally, the frequency locations of any visible spectral zeros (anti-resonances) were marked.

To determine the significance of any differences found between the various measured quantities, a repeated measures ANOVA was conducted for two trial factors: speaking condition (pre- and post-laryngectomy) and vowel.

4.3. Results

4.3.1. General Description

Typically, post-laryngectomy speech spectra demonstrated higher formant frequencies, narrower formant bandwidths (especially for F1), reduced low frequency energy and differing relative formant amplitudes compared to the corresponding pre-laryngectomy spectra. Moreover, spectral zeros were visible in certain post-laryngectomy spectra (but not in the pre-laryngectomy spectra). An example of the measured spectra from the two different speech conditions is shown in Figure 4.1. In this particular example, the first three formant frequencies increased by 40 Hz, 267 Hz, and 1168 Hz respectively. The extreme shift in third formant frequency may be due to the presence of the spectral zero at 3050 Hz that could have attenuated the true third formant to the point that it is indistinguishable from the contribution of the self-noise in that part of the spectrum. Thus, the fourth formant is effectively acting as the third formant.

In this example, compared to the pre-laryngectomy speech, the first and third bandwidths were narrower for EL speech (by 82.0 Hz and 58.9 Hz respectively), but the second formant bandwidth widened by 58.9. The difference between the amplitudes of the first and second formants, the first and third formants, and the second and third formants increased by 10.9 dB, 3.2 dB and 14.1 dB, respectively. The changes in the peak-to-valley ratios were small with *F1F2ptvr* decreasing by of 4.2 dB and *F2F3ptvr* by 5.4 dB. The reduction in low frequency energy in EL speech is demonstrated by the 95% decrease in the normalized low frequency energy, $E_{r,}$ and the 33.5dB and 47.5 dB decrease in *H1-A3* and *H1-A1* respectively.

Although this particular example was representative of the data, there was a significant amount of variation between both subjects and vowels. The mean values and variability of the measured quantities are discussed in the following section.

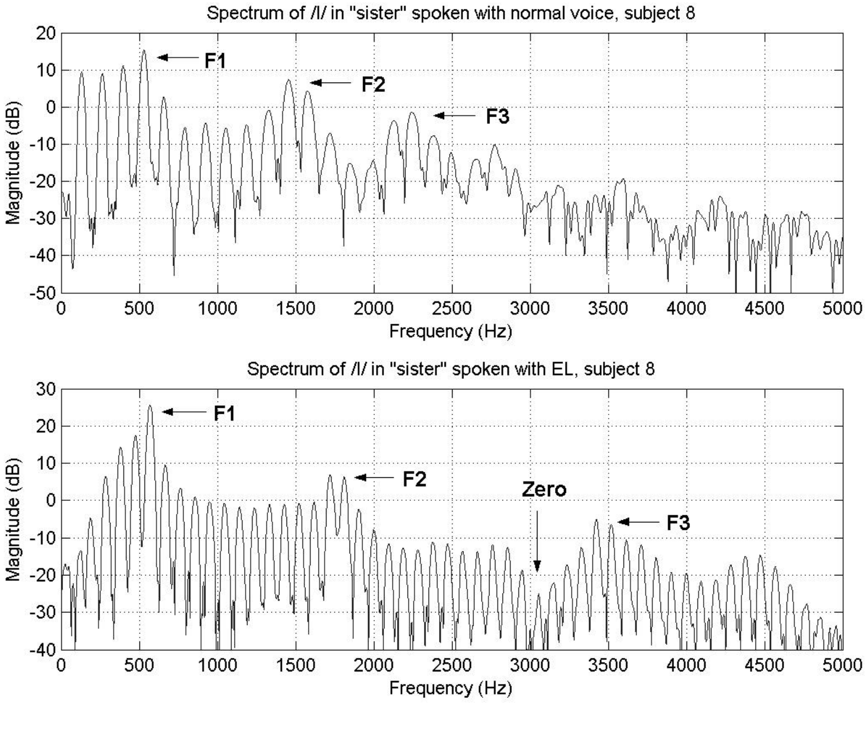


Figure 4.1. The spectra of the vowel /I/ in “sister” in the pre- *(top)* and post-laryngectomy *(bottom)* speech of a single subject. The formant and zero locations are marked accordingly.

4.3.2. Mean values

4.3.2.1 Formant Frequencies

The mean and range of the first three formant frequencies across all subjects are shown in Figure 4.2 for both pre- and post-laryngectomy (i.e. normal and EL) speech. The data demonstrate that the formant frequencies of EL speech are clearly and consistently higher than they are for normal, pre-laryngectomy speech. On average, F1, F2, and F3 showed an increase of 124 ± 36 Hz, 212 ± 91 Hz, and 388 ± 154 Hz respectively. The detailed mean and range data can be found in Table 13.1 in Appendix E. Figure 4.3, presents an alternative view of this increase in formant frequency by plotting the measured normal and EL speech F1 and F2 values of all the vowels for each subject. Although there is some degree of overlap, the EL speech formants (black) tend to cluster towards the upper right of the graph while the normal speech formants (white) cluster towards the lower left.

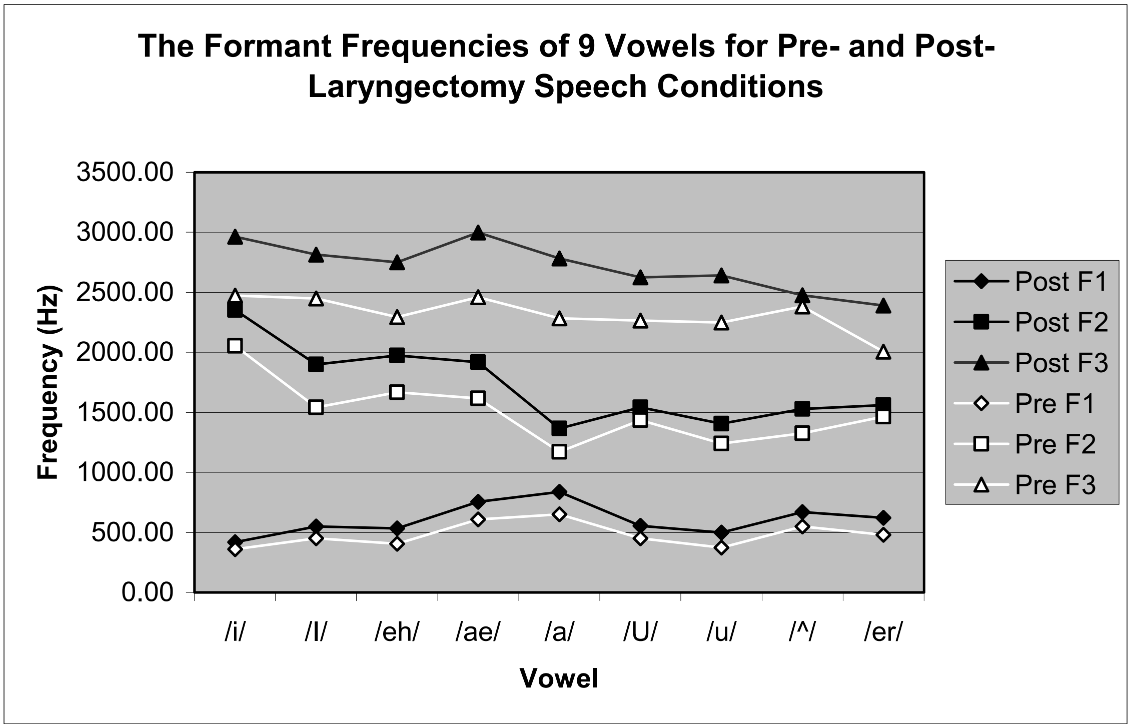


Figure 4.2. The mean values of the first three formants of the nine vowels for both EL (black) and normal (white) speech. In every case, except for F3 of the vowel /^/, the formant frequencies of EL speech are clearly higher than those of normal speech.

The ANOVA revealed significant differences for both main effects (vowel and speaking condition) for all three formant frequencies ($p < 0.01$), but only showed a marginal significance for the vowel*speaking condition interaction ($p = 0.016$). The vowel dependence of the F1 difference is visible in Figure 4.2, where the difference is smaller for the high vowels, /i/ and /I/ than for other vowels.

4.3.2.2 Formant Bandwidths

The first three bandwidths, BW1, BW2, and BW3, were reduced on average by 36 ± 15 Hz, 20 ± 35 Hz and 38 ± 59 Hz respectively. That the range of bandwidth change for BW2 and BW3 actually produces a negative reduction (i.e. a bandwidth increase) indicates that the EL speech bandwidths were not always narrower than their normal speech counterparts. This situation is clearly illustrated in Figure 4.4 where the mean bandwidths are plotted for each vowel. While the first formant bandwidth is always smaller in the post-laryngectomy case, for certain vowels, such as /æ/, the second and third formant bandwidths were actually larger. The detailed mean and range data for the formant bandwidths can be found in Table 13.2 in Appendix E.

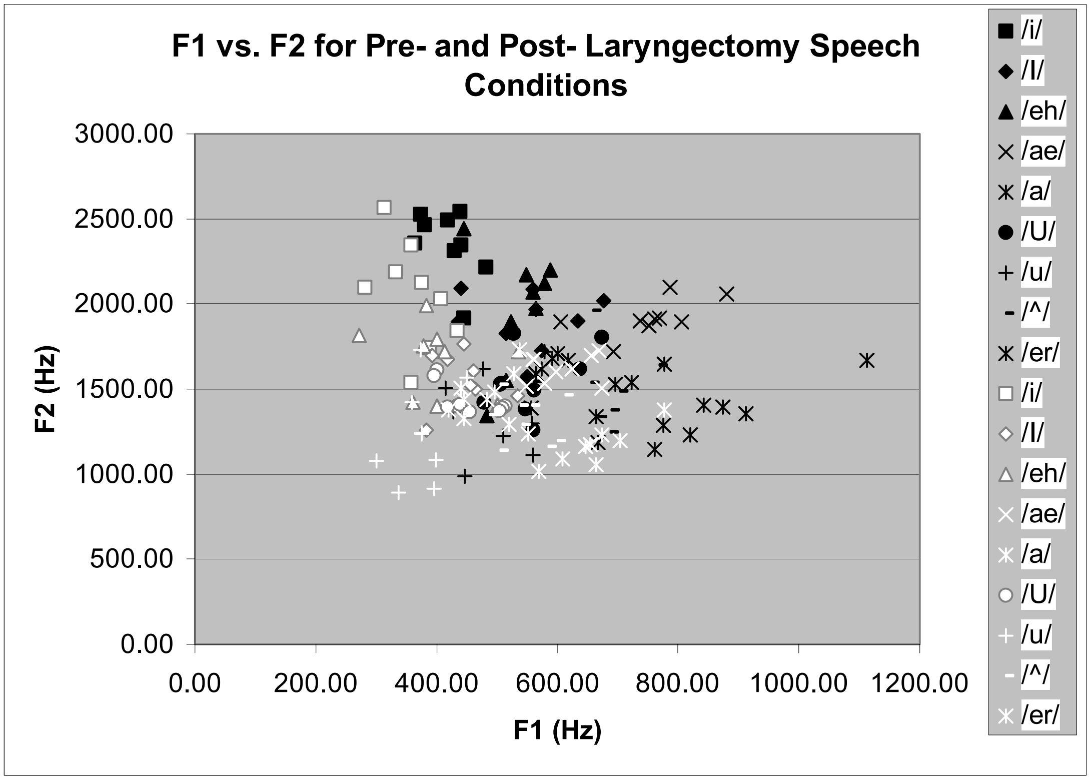


Figure 4.3. F1 vs. F2 of all nine vowels for all nine subjects for both EL *(black)* and laryngeal *(white)* speech. The formant values for EL speech tend to cluster towards the upper right (i.e. higher frequency values) while those of the normal speech tend towards the lower left.

4.3.2.3 Formant Amplitudes

Because the original recordings of the VA subjects were not calibrated for absolute intensity, the absolute formant amplitudes are products of the voicing source, the vocal tract acoustics and the recording conditions and thus would not be useful measures for tracking the effects of changes to the voicing source and vocal tract acoustics. However, the relative formant amplitudes (i.e. the differences between the formant amplitudes) are better suited to this task as they are independent of the recording conditions (since the same microphone was used in every recording session).

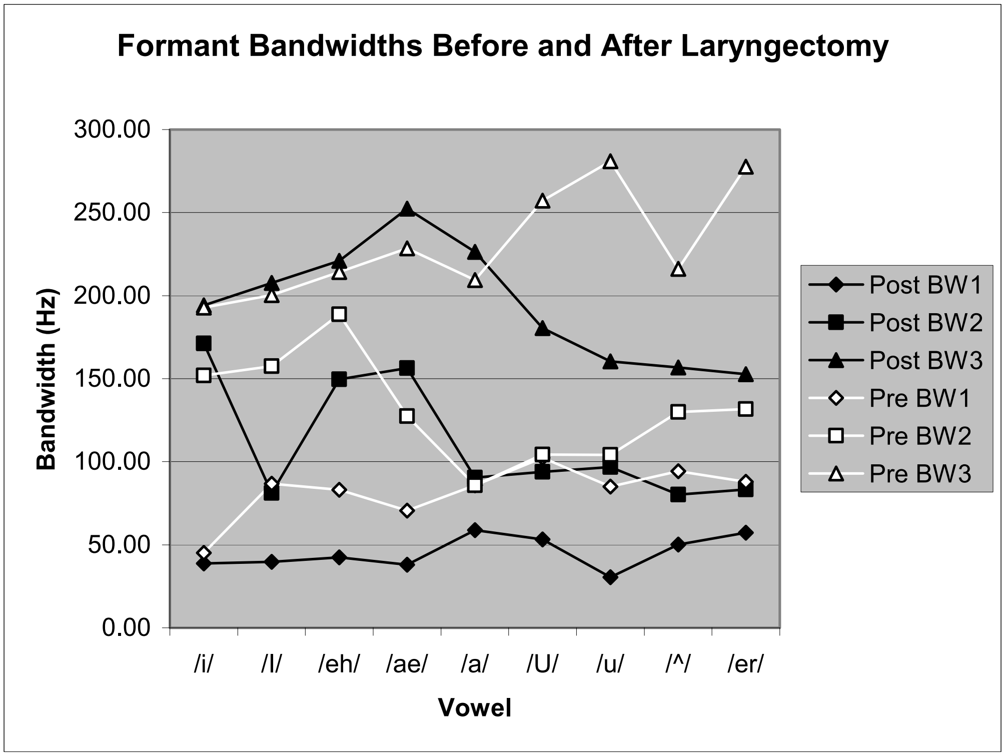
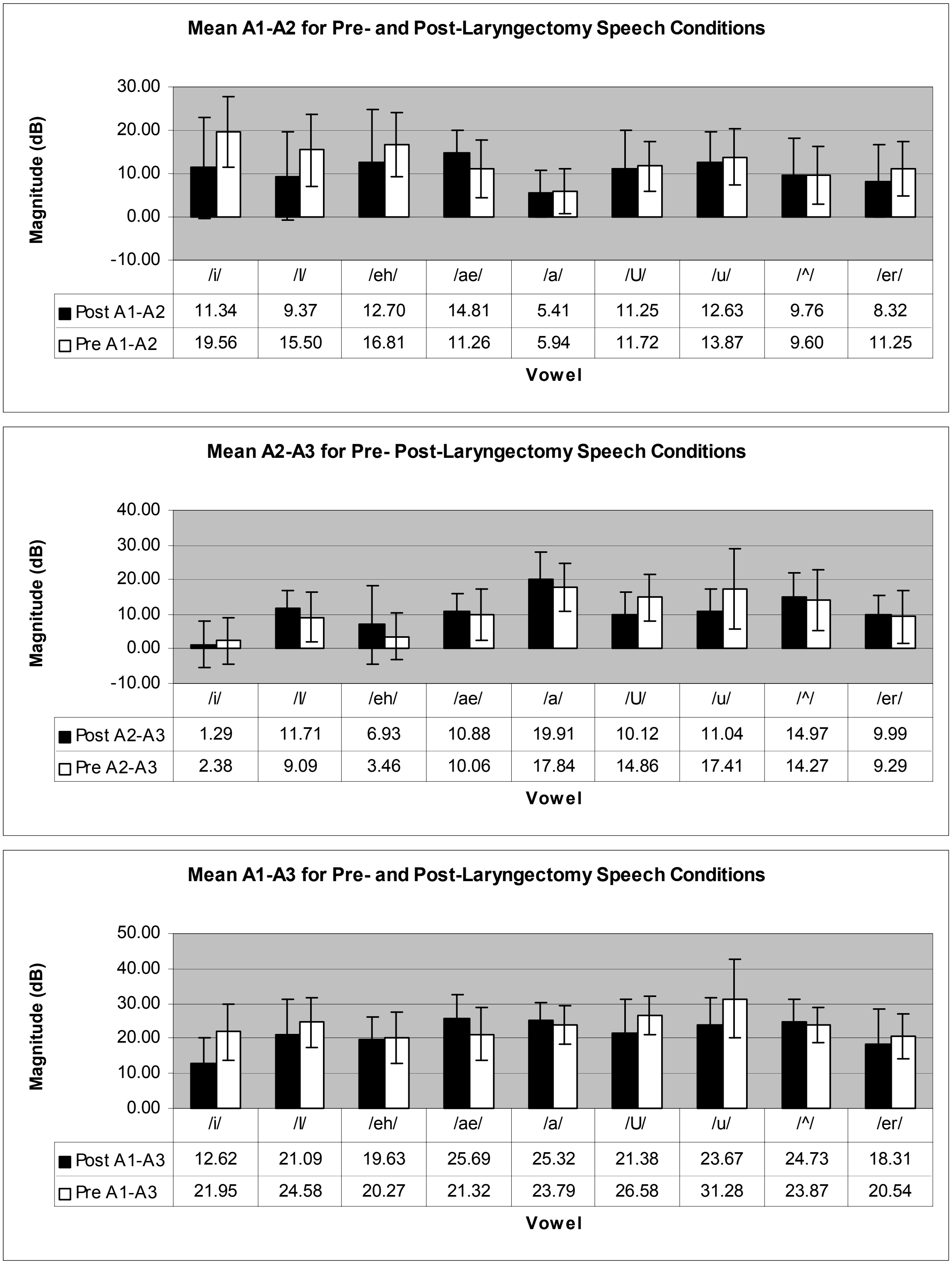


Figure 4.4. The mean bandwidths of the first three formants for pre- *(white*) and post-laryngectomy (*black*) speech. The first formant bandwidth was always smaller in post-laryngectomy speech, but this was not always true for the higher formant bandwidths.

Figure 4.5 shows the mean relative formant amplitudes, A1-A2 (top), A2-A3 (middle) and A1-A3 (bottom) of each vowel for both EL and normal speech. The differences between the relative formant amplitudes appear to be largely vowel dependent; even within a single vowel, there is a great deal of variability as shown by the large error bars. For example, while the mean values of A1-A2 greatly differs between the two speech types for the vowels, /i/ and /I/, they are very similar for the vowel /a/. Yet the range of obtained values indicates a considerable amount of overlap between the A1-A2 values for /i/ and /I/ despite the large gap in the means. These observations are supported by the statistical analysis. Significance was found for the vowel main effect (p < 0.01) for all



Mean A1-A2 for Pre- and Post-Laryngectomy Speech Conditions

	/i/	/I/	/eh/	/ae/	/a/	/U/	/u/	/^/	/er/
■ Post A1-A2	11.34	9.37	12.70	14.81	5.41	11.25	12.63	9.76	8.32
□ Pre A1-A2	19.56	15.50	16.81	11.26	5.94	11.72	13.87	9.60	11.25

Mean A2-A3 for Pre- Post-Laryngectomy Speech Conditions

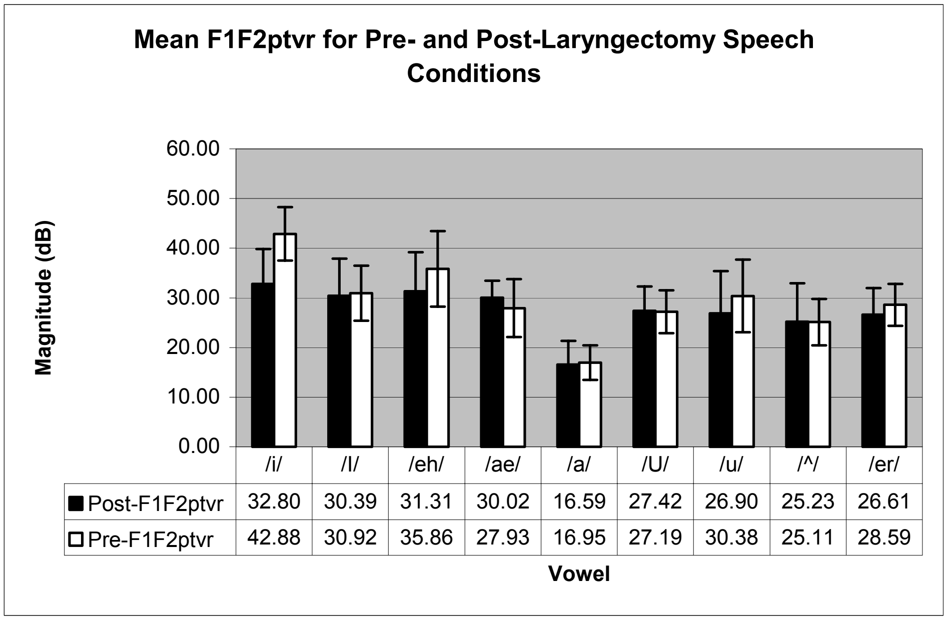
	/i/	/I/	/eh/	/ae/	/a/	/U/	/u/	/^/	/er/
■ Post A2-A3	1.29	11.71	6.93	10.88	19.91	10.12	11.04	14.97	9.99
□ Pre A2-A3	2.38	9.09	3.46	10.06	17.84	14.86	17.41	14.27	9.29

Mean A1-A3 for Pre- and Post-Laryngectomy Speech Conditions

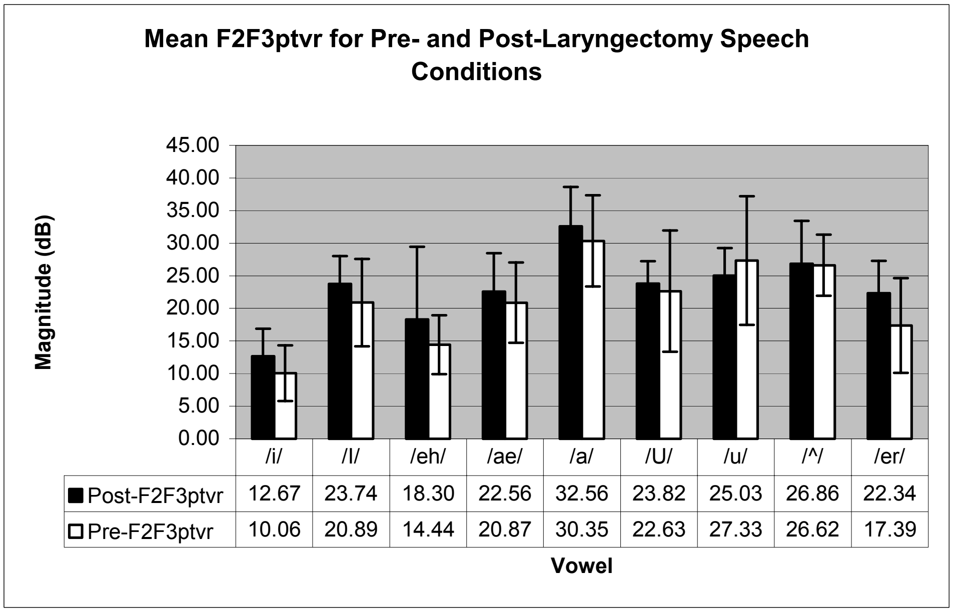
	/i/	/I/	/eh/	/ae/	/a/	/U/	/u/	/^/	/er/
■ Post A1-A3	12.62	21.09	19.63	25.69	25.32	21.38	23.67	24.73	18.31
□ Pre A1-A3	21.95	24.58	20.27	21.32	23.79	26.58	31.28	23.87	20.54

Figure 4.5. Plots of A1-A2 (*top*), A2-A3 (*middle*) and A1-A3 (*bottom*) for all 9 vowels for pre-(*white*) and post-(*black*) laryngectomy speech.

three measures but not for speaking condition (*p* = 0.373, 0.271, 0.943 for A1-A2, A2-A3 and A1-A3 respectively). Moreover, the vowel*speaking condition interaction was significant for A1-A3 (*p* = 0.003), almost significant for A1-A2 (*p*= 0.021) but not for A2-A3 (*p* = 0.172).



	/i/	/I/	/eh/	/ae/	/a/	/U/	/u/	/^/	/er/
■ Post-F1F2ptvr	32.80	30.39	31.31	30.02	16.59	27.42	26.90	25.23	26.61
□ Pre-F1F2ptvr	42.88	30.92	35.86	27.93	16.95	27.19	30.38	25.11	28.59



	/i/	/I/	/eh/	/ae/	/a/	/U/	/u/	/^/	/er/
■ Post-F2F3ptvr	12.67	23.74	18.30	22.56	32.56	23.82	25.03	26.86	22.34
□ Pre-F2F3ptvr	10.06	20.89	14.44	20.87	30.35	22.63	27.33	26.62	17.39

Figure 4.6. Mean F1F2ptvr *(top)* and F2F3ptvr *(bottom)* for pre- and post-laryngectomy speech. Except for F1F2ptvr for the vowel /i/, the mean peak-to-valley ratios appear to be independent of speech type.

4.3.2.4 Peak-to-valley ratios

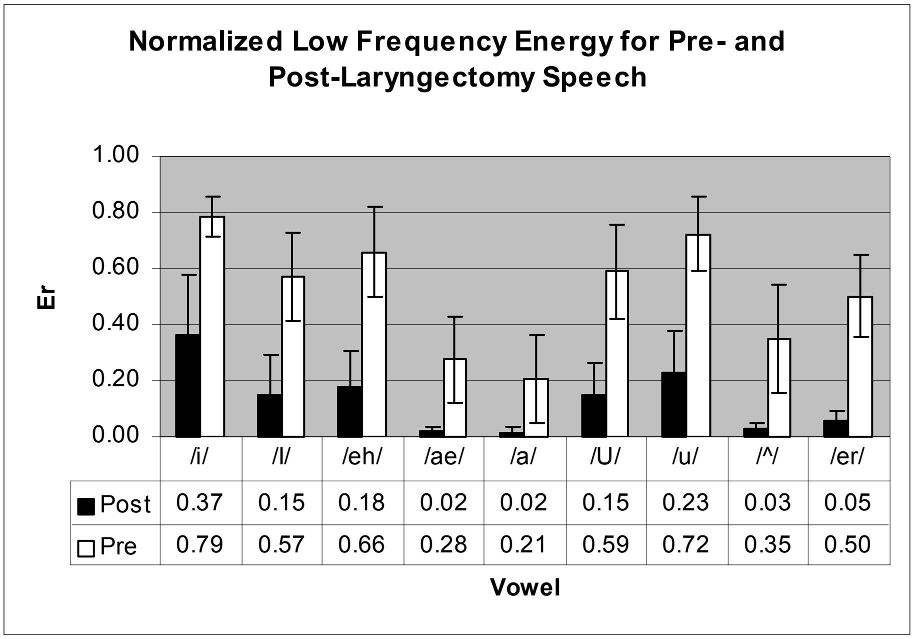
As demonstrated by Figure 4.6, in general there appears to be little difference between the peak-to-valley ratios of normal and EL speech. On average there was a change of –2.1 ± 3.6 dB and 1.8 ± 2.3 dB for *F1F2pvtr* and *F2F3ptvr* respectively. The notable exception to this trend was for the vowel /i/ which presented a 10 dB decrease in *F1F2ptvr*. The reduced low frequency energy in EL speech has most likely attenuated the low frequency first formant of /i/ thus producing a decreased *F1F2ptvr*. Significant differences were found for the vowel main effect ($p < 0.01$) but not for speaking condition ($p = 0.373$) for both peak-to-valley ratios. Moreover, a significant interaction effect ($p <0.01$) was found for *F1F2ptvr* but not for *F2F3ptvr*.

4.3.2.5 Low frequency energy measures

The quantities E_r, *H1-A1*, and *H1-A3*, were measured to quantify the differences between the two speech types that occur at low frequencies (i.e. below 500 Hz). All three quantities showed significant differences for the condition effect. However, only E_r and *H1-A3*, presented a significant vowel effect ($p < 0.01$) and significant vowel*speaking condition interaction effect ($p < 0.01$). The mean E_r of each vowel for each speech type is shown in Figure 4.7. Consistent with the statistical results, the difference in E_r between the two speaking conditions is considerable, with means of 0.5 ± 0.2 and 0.1 ± 0.1 for pre- and post-laryngectomy speech, respectively. Figure 4.5 clearly displays the vowel dependency of both the absolute values of E_r as well as the changes in E_r between the two speech conditions. In general it appears that there was a greater percentage change for the low vowels (which have high F1s) such /a/ and /æ/ while there was a greater absolute difference for the high vowels such as /i/ and /I/.

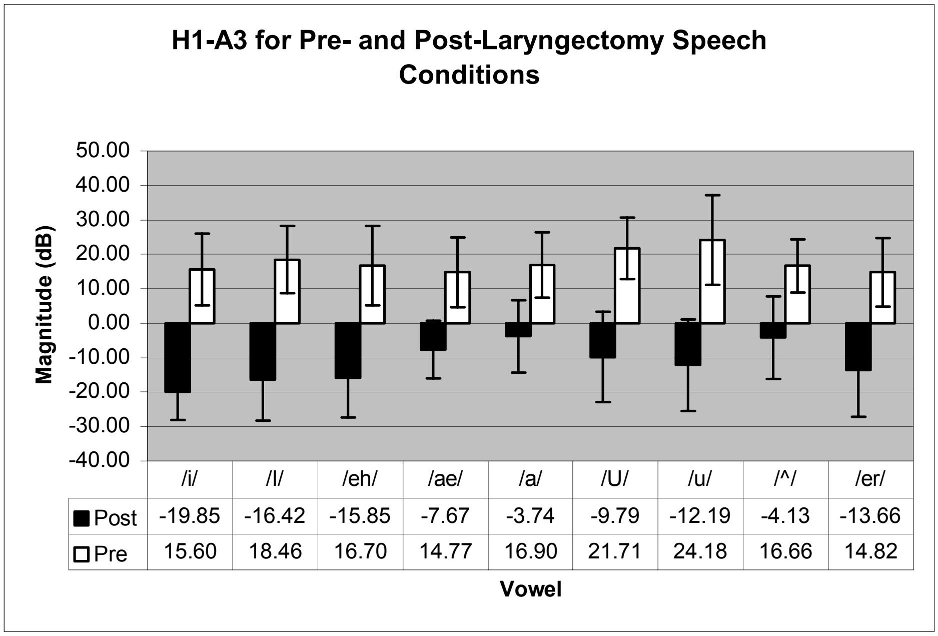
Similarly, *H1-A3*, plotted for each vowel in Figure 4.8, also demonstrates a significant vowel and speaker dependency. Where as *H1-A3* was always a positive quantity in pre-laryngectomy speech, it was always negative in post-laryngectomy speech. Moreover, *H1-A3* was significantly dependent on the vowel produced as would be expected since *A3* is affected by all three formant frequency locations.

The disparity in low frequency characteristics between the two speech types is further confirmed by the difference between the *H1-A1* values. As demonstrated by Figure 4.9, the difference in amplitudes between the first harmonic and the first formant was reduced on average by 26.7 ± 3.8 dB in EL speech. This difference was relatively independent of vowel, as demonstrated by the small standard deviation and the non-significance of the vowel*speech condition interaction ($p = 0.06$).



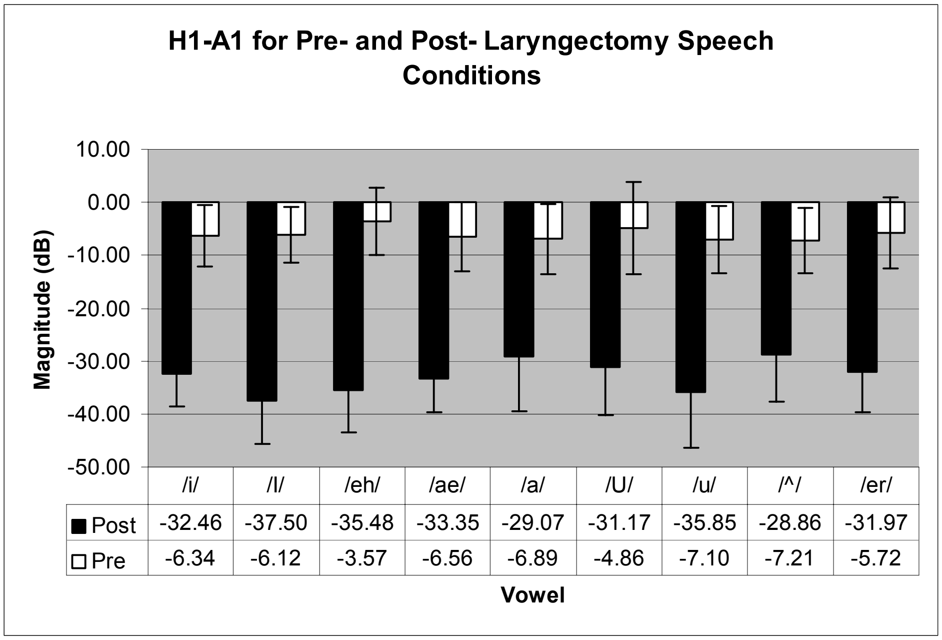
	/i/	/I/	/eh/	/ae/	/a/	/U/	/u/	/^/	/er/
■ Post	0.37	0.15	0.18	0.02	0.02	0.15	0.23	0.03	0.05
□ Pre	0.79	0.57	0.66	0.28	0.21	0.59	0.72	0.35	0.50

Figure 4.7. Mean normalized low frequency energy, E_r, of each vowel for each speaking condition. E_r was found to be dependent on both speaking condition and vowel.



	/i/	/I/	/eh/	/ae/	/a/	/U/	/u/	/^/	/er/
■ Post	-19.85	-16.42	-15.85	-7.67	-3.74	-9.79	-12.19	-4.13	-13.66
□ Pre	15.60	18.46	16.70	14.77	16.90	21.71	24.18	16.66	14.82

Figure 4.8. Mean *H1-A3* of each vowel for each speaking condition. *H1-A3* was found to be dependent on both speaking condition and vowel.



	/i/	/I/	/eh/	/ae/	/a/	/U/	/u/	/^/	/er/
■ Post	-32.46	-37.50	-35.48	-33.35	-29.07	-31.17	-35.85	-28.86	-31.97
□ Pre	-6.34	-6.12	-3.57	-6.56	-6.89	-4.86	-7.10	-7.21	-5.72

Figure 4.9. Mean *H1-A1* of each vowel for each speaking condition. *H1-A1* was found to only be dependent on speech condition.

4.3.2.6 Spectral Zeros

The presence of spectral zeros (i.e. anti-resonances) was unique to the EL speech spectra, an example of which is shown in Figure 4.1. Figure 4.1 demonstrates the potential impact of these zeros: the attenuation of one or more formants. The zeros were believed to be an important distinguishing characteristic of EL speech and as such, the frequencies of any observable zeros in the computed spectra were measured. The number of zeros marked for each spectrum ranged from 0 to 2.

The frequency locations of the zeros were quite variable across both vowel and speaker. Figure 4.10, which divides the measured zero frequencies by speaker, illustrates the extent of this variability. Some speakers, such as speakers 3 and 4, presented zeros that cluster in one or two discrete parts of the spectrum. Other speakers, however, such as speakers 7 and 8 presented zeros that are to be distributed across almost the entire frequency band. The disparity in the number of zeros measured for each speaker is indicative of the fact that the number of observable zeros varied for each spectrum. Figure 4.11 plots the zero frequencies separated by vowel. Although Figure 4.11 demonstrates that there isn't a clear relationship between vowel and zero-frequency (at least across speakers), it does support the notion of a speaker dependency. This is illustrated by the data collected for speaker 4, for whom zero-frequencies were found at around 1500 Hz for several vowels. Similarly, zeros were observed near the 3000 Hz neighborhood in several vowel spectra of Speaker 3.

4.4. Discussion

4.4.1. Implications of spectral characteristics

The goal of this study was to investigate differences in the properties of pre- and post-laryngectomy speech in order to find properties of EL speech that could contribute to its unnatural quality. This analysis has revealed several inherent properties that differentiate EL speech from laryngeal speech. The increase in frequency of the first three formants that was common to all the subjects analyzed is consistent with the findings of Sisty and Weinberg (1972) who found a systematic increase in the formant frequencies of esophageal speakers. These changes in formant frequencies are certainly due to the shortening of the pharyngeal section of the vocal tract that occurs as a result of the laryngectomy surgery. The concentration of the change to the pharyngeal region helps explain why the increase in first formant frequency was less pronounced for the vowels /i/ and /I/ than for other vowels. Consider the simple models of the vowels /i/ and /a/ shown in Figure 4.12. For the vowel /a/, the formant frequencies are approximately equal to the resonant frequencies of the two small tubes that comprise the entire vocal tract (assuming that the area of one tube is much greater than that of the other).

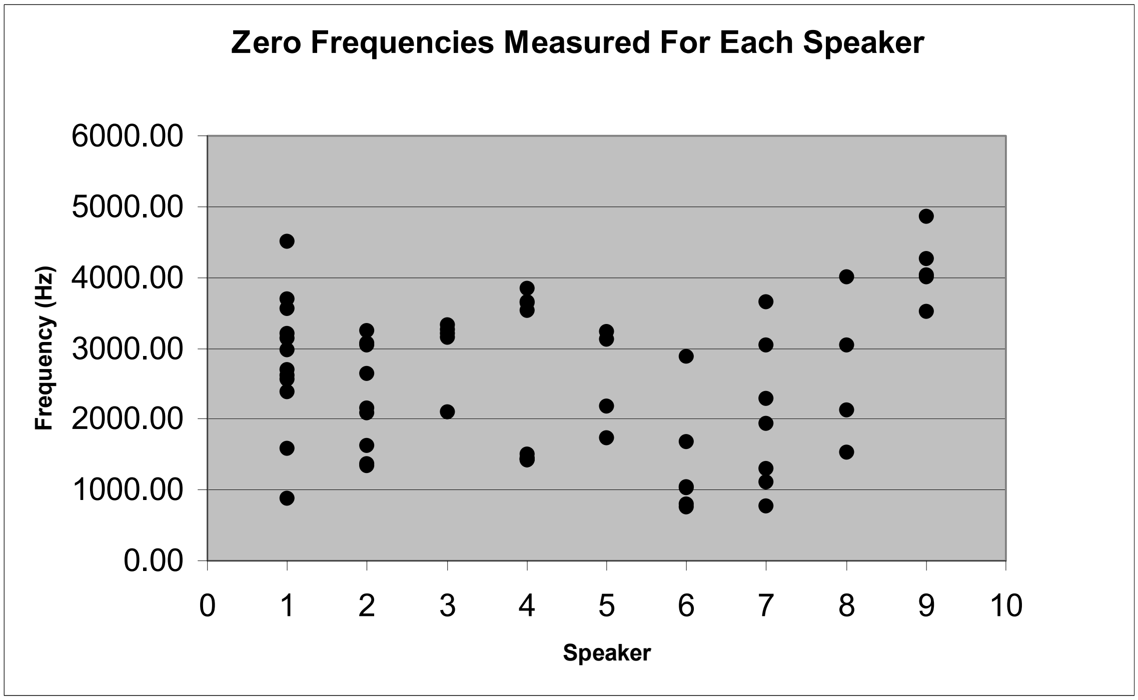


Figure 4.10. Zero frequencies separated by speaker across all vowels. For some speakers, the measured zero frequencies tended to cluster around certain frequencies. For others, however, the zero frequencies were spread out over much of the entire frequency band.

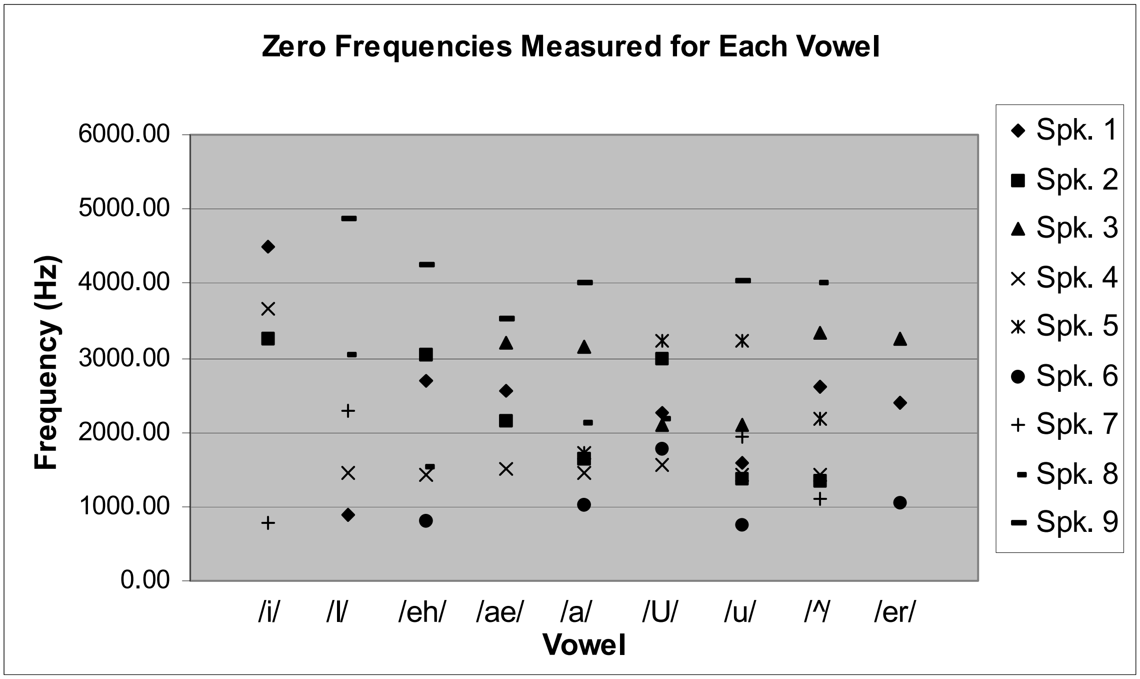


Figure 4.11. Zero frequencies separated by vowel. Although there doesn't appear to be a clear relationship between the zero frequencies and vowels, the zeros measured for speakers 3 and 4 are clustered around 1500 Hz and 3000 Hz respectively.

For both tubes, these frequencies occur at the quarter-wave frequencies, i.e.

$$F_n = \frac{(2n-1)c}{4l}, \ n = 1,2, \ldots \tag{4.2}$$

where c is the velocity of sound and l is the length of the tube in question. In this case, the formant frequencies are inversely proportional to the lengths of the tubes. Similarly, the formant frequencies for the vowel /i/ are also roughly equal to the resonant frequencies of the component tubes, except that in this case, these frequencies are the half-wave frequencies, i.e.

$$F_n = \frac{nc}{2l}, \ n = 1,2, \ldots \tag{4.3}$$

At low frequencies, however, this configuration of the vowel /i/, is equivalent to a Helmholtz resonator whose natural frequency is

$$F_h = \frac{c}{2\pi}\sqrt{\frac{A_1}{A_2 l_1 l_2}} \tag{4.4}$$

where A_1, l_1, A_2, l_2 are the cross sectional areas and lengths of the front and back tubes respectively. This frequency is significantly lower than the half-wave frequencies of either tube and is thus the first formant frequency for /i/. Unlike the quarter wave frequencies, the Helmholz frequency is inversely proportional to the square root of the lengths of the tubes. Thus a reduction in the length of the back tube will have a greater effect on a vowel such as /a/ than it will for vowel such as /i/.

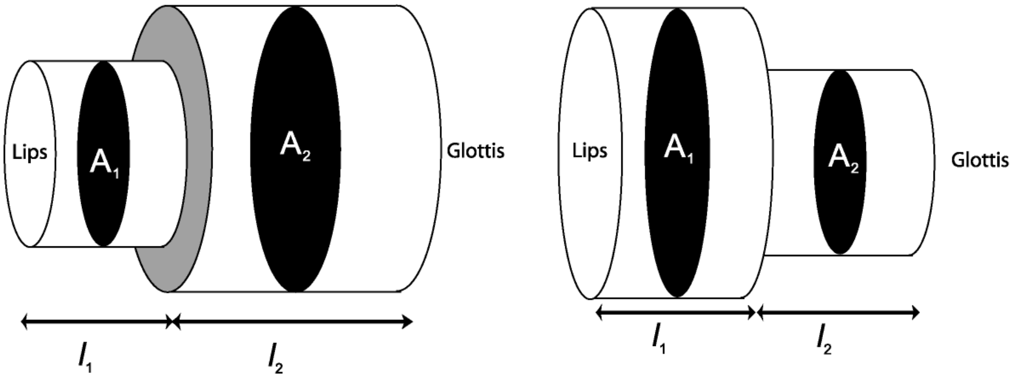


Figure 4.12. Simple tube models for the vowel /i/ (left) and /a/ (right).

It should also be noted that the nature of EL speech often made it difficult to properly measure the formant frequencies. The reduced low frequency energy sometimes significantly reduced the amplitudes of the first formants of high vowels (i.e. vowels with low first formants) making it difficult to isolate the first formant peak. Likewise, the presence of the self-noise combined with the presence of attenuated second and third formant frequencies occasionally made it difficult to pick out the proper higher formants. An example of this phenomenon may have occurred in the vowel /I/ for speaker 8, which was discussed earlier. Based on the spectrum shown in Figure 4.1, the only prominent higher formant is found at 3427 Hz, which is about 1180 Hz higher than the third formant for the same vowel in this subject’s pre-laryngectomy speech. It is improbable that a change in the vocal tract length would cause such an extreme increase in formant frequency. Therefore it’s more likely that the true third formant amplitude has been reduced to such a degree that it has been masked by the EL self-noise. Although this situation occurred only a few times during the analysis, it does help explain the upper limits of the range of measured F3 values.

The perceptual effect of this formant shift is unclear. The distance between the first two formants determines the type of vowel perceived. Weiss *et al.* (1979) found a vowel intelligibility rate of 80% for EL speakers, but it is unlikely that the increase in formant frequencies is responsible for this phenomenon since the formant distances of EL speech are still within normal limits. The absolute formant frequencies, however, are similar to those found in the speech of normal adult females (Peterson & Barney, 1952). As such, while the formant shift may cause the resulting speech to sound somewhat more “feminine”, it probably does not contribute to the unnatural quality of EL speech.

The narrowing of the formant bandwidths, especially that of the first formant, agrees with the findings of House and Stevens (1958) who found significantly narrower bandwidths when the glottis was closed. The closed glottis position is analogous to the anatomical situation of laryngectomy patients. Their vocal tracts have been completely separated from the subglottal system and thus are effectively rigidly terminated. The rigid termination removes any losses associated with subglottal coupling, thus reducing the formant bandwidths. The effect is most pronounced on the first formant bandwidth because the losses due to the glottal impedance are inversely proportional to the frequency squared (Liljencrants 1985). Consistent with this model, the statistical analysis of the results in this study revealed that only the narrowing of the first formant bandwidth was found to be significant.

The narrower the bandwidth of a resonance, the more sinusoidal it becomes. This can be seen by inspecting the speech of speaker 7, whose mean EL first formant bandwidth was 20.0 ± 15.42 Hz. An example of this subject's speech (the vowel /æ/ in "basket") is shown in Figure 4.13. The spectrum is dominated by the narrow first formant at 605 Hz that has a bandwidth of 37 Hz. The corresponding waveform appears to be very sinusoidal, oscillating with a period of about 0.0017s. This period is exactly the inverse of the first formant frequency, thus illustrating the effect of such a narrow first formant bandwidth.

Perceptually, narrower formant bandwidth adds a shriller, more tonal quality to EL speech. This seems to be the case with speaker 7, whose speech was often uncomfortable to listen to. In terms of intelligibility, the reduced formant bandwidths may actually help the situation by counteracting the masking properties of the EL self-noise. This is supported by the lack of difference in the peak-to-valley measures between the two different speech types. It was believed that one of the main effects of the self-noise was that it filled in the spectral valleys between the formant peaks thereby producing smaller peak-to-valley ratios. However, the average difference between the two speech types was only a few decibels, and was found not to be statistically significant. It appears that the reduced formant bandwidths have increased the corresponding formant amplitudes, thus offsetting the masking properties of the self-noise.

The results of the three measurements of the low frequency spectral characteristics support the previous observations of a low frequency deficit in EL speech. (Weiss *et al.* 1979, Qi and Weinberg 1991). The normalized low frequency energy, and the amplitude of the first harmonic relative to both the first and third formants were significantly reduced in EL speech. The cause of this low frequency deficit is the EL sound source itself and not the acoustic transmission properties of the neck (Meltzner *et al.* 2003). Qi and Weinberg (1991) postulated that compensating for the lack of low frequency energy would improve the quality of the speech, but the results of the perceptual experiments described in Section 3 demonstrate that doing so (at least in their suggested manner) only produced a limited improvement.

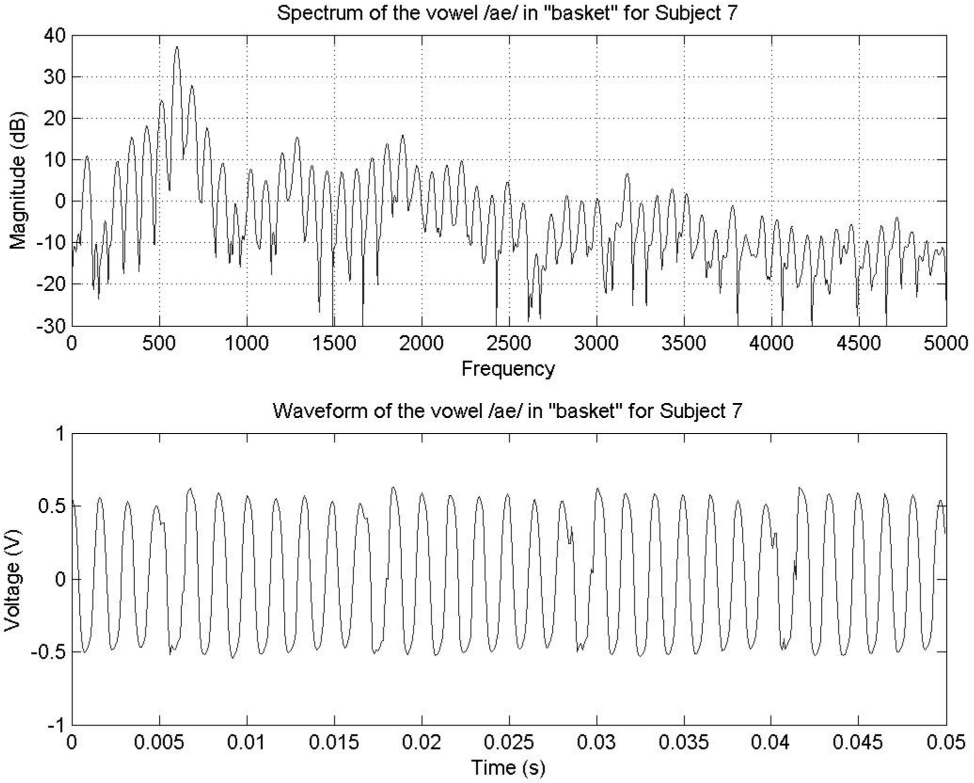


Figure 4.13. The spectrum and waveform of the vowel /æ/ in "basket" for speaker 7. The sinusoidal nature of the waveform is indicative of the narrow first formant that dominates the spectrum.

The lack of significance of the main speech condition effect (i.e. the difference in voicing sources) on the differences between the relative formant amplitudes of the two speech types was somewhat surprising. It was initially believed that the radically different voicing source characteristics of the EL devices would cause a systematic change in relative formant amplitudes. A typical natural voicing source spectrum decays at a rate of $1/f^{2}$ whereas the EL source spectrum resembles Figure 4.14. However, the significance of the voice*speaking condition interaction for *A1-A3* (and its near significance for *A1-A2*) indicates a difference between the speaking conditions but that the difference is dependent on the vowel being produced.

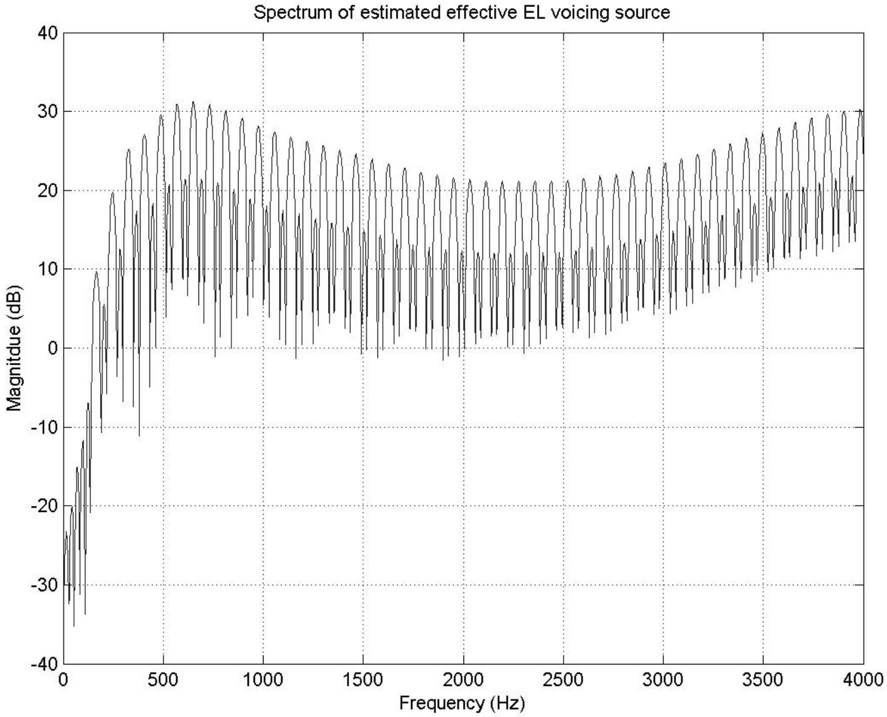


Figure 4.14. The estimated spectrum of the EL voicing source. An acceleration signal obtained when an accelerometer was placed on the vibrating head of a Servox EL was filtered by the neck frequency response function (Meltzner *et al.* 2003) and then differentiated (to account for the lip radiation characteristic). The spectral shape greatly differs from that of a natural glottal source.

There were large variations in the relative formant amplitudes within a single vowel in both the pre- and post-laryngectomy conditions. The variability was expected in the pre-laryngectomy case since the formant amplitudes are dependent on both the vocal tract acoustics and the spectral characteristics of the glottal voicing source. The somewhat disordered voices of these subjects only added to that variability. On the other hand, it was expected that given similar formant frequencies, the relative formant amplitudes of two different EL speakers would be quite similar. Some of the inconsistency in these measures may be due to individual differences in the transmission properties of the neck wall (Meltzner *et al.* 2003) or differences in the outputs of the types of EL used. However, it appears that in many cases one or more formant amplitudes are being attenuated. Consider the two spectra displayed in Figure 4.15 which were obtained from the vowel /I/ in “sister” in speakers 1 and 2. Both subjects used a Servox EL and yet there is an obvious difference between the relative amplitudes of the first two formants. For speaker 1, *A1-A2* was –5.9 dB whereas it was 9.5 dB for speaker 2. Only a small amount of this discrepancy can be attributed to the difference in formant frequencies. This value can be calculated as follows. If the formants above F3 are ignored, then the vocal tract transfer function can be written as:

$$T(s)=\prod_{n=1}^{3}\frac{s_n s_n^*}{(s-s_n)(s-s_n^*)} \tag{4.5}$$

where the s_n , s_n^* are the conjugate pole pairs that correspond to the formant frequencies. The conjugate poles are related to the formant frequencies and bandwidths by

$$s_n=\frac{bw_n}{\pi}+j2\pi F_n \tag{4.6}$$

where F_n and bw_n are the formant frequencies and bandwidths of the nth formant, respectively. The magnitude of the transfer function at any frequency can be found by letting $s = j2\pi f$, where f is frequency and then computing the magnitude. A change in formant frequency can be introduced to the transfer function by multiplying the transfer function by a ratio of the old conjugate pole pair to the new conjugate pole pair, i.e.

$$R(s)=\frac{(s-s_n)(s-s_n^*)s_m s_m^*}{(s-s_m)(s-s_m^*)s_n s_n^*} \tag{4.7}$$

where s_n , s_n^* are the old conjugate pole pair and s_m , s_m^* are the new conjugate pole pair. The change in magnitude at any frequency caused by this substitution is found by computing the magnitude of $R(s)$ at the desired frequency. It can be shown using Eq. 4.7, that by altering the transfer function corresponding to speaker 2 by setting the first and third formant frequencies and bandwidths equal to those found for speaker 1 only decreases the amplitude of the second formant by 1.6 dB. Moreover because of the proximity of F3 to F2 after this shift is made, Eq 4.7 somewhat overstates this decrease. This indicates that there must be another reason for this difference in A1-A2. It appears likely that there is a spectral zero located at about 880 Hz in the spectrum of speaker 1 and that this zero is attenuating the first formant such that its amplitude is actually less than that of the second formant.

Thus, it is probable that the presence of these spectral zeros affects the spectral characteristics of EL speech. Although not readily visible in every spectrum that was analyzed, spectral zeros were common occurrences in the EL speech spectra, as shown by Figures 4.10 and 4.11. The frequency locations of these zeros varied widely across vowels but there was some evidence of commonality within a single speaker. It is possible that these zeros are the result of destructive interference between the sound filtered by the vocal tract and the directly radiated self-noise. However, this would require that the two sounds be 180 degrees out of phase at the frequency where the zero appears, i.e. that the difference in the paths traveled by the sounds be equal to half a wavelength. The large majority of the zeros were found at frequencies below 3500 Hz, meaning that the difference in path length would have to be at least 5 cm, which is unlikely, given the dimensions of the human head. Instead, these zeros are the effects of placing the EL sound source at a location other than the terminal end of the vocal tract. This introduces a back cavity in the vocal tract. At frequencies at which the impedance of this back cavity acts as a short circuit, a zero is introduced into the vocal tract transfer function. Because the impedance of the back cavity is dependent on both the cavity's shape and length, it would not be unexpected to find that the resulting zero frequencies vary both across vowels and across speakers. The complex relationship between zero frequency, source location and vowel type warrants a more thorough investigation. Such a study is discussed in Chapter 6.

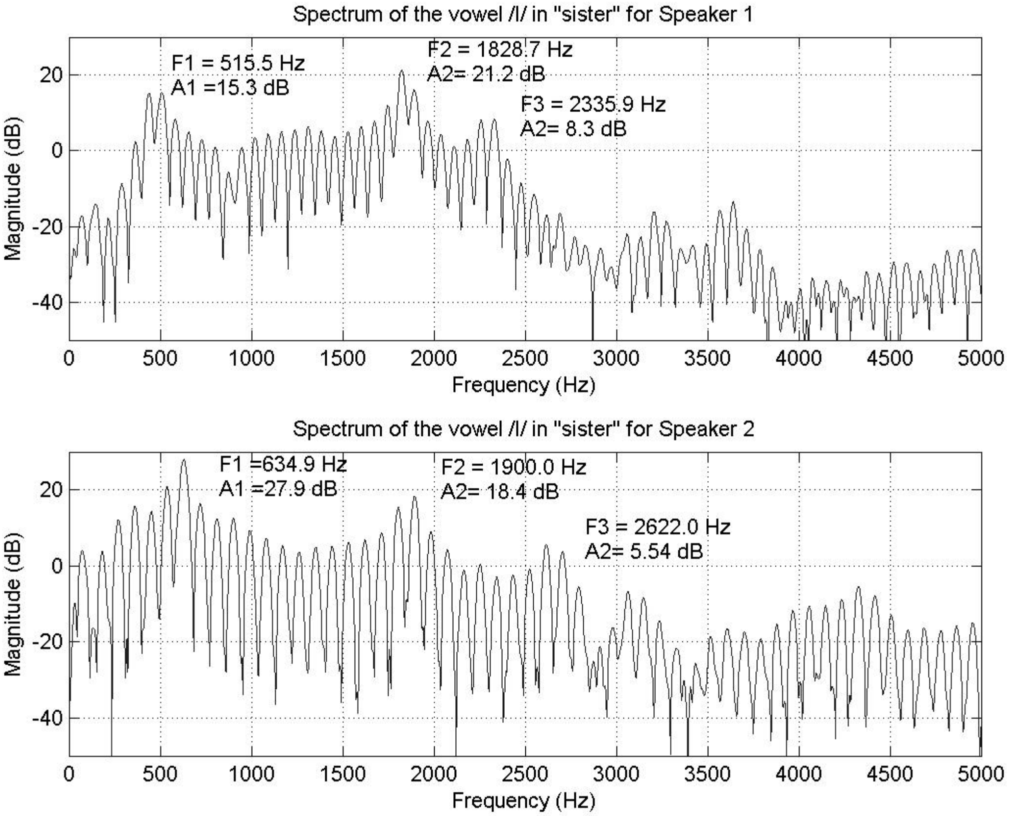


Figure 4.15. Spectra of the vowel /I/ in "sister" for subject 1 *(top)* and speaker 2 *(bottom)*. It appears that a spectral zero at about 880 Hz is significantly attenuating the first formant of speaker 1.

The perceptual effect of these zeros will most likely depend on their proximity to the formant frequencies. The closer a zero is to a formant frequency, the larger the effect it will have. If the frequencies of a zero and a formant are almost equal, they will cancel each other out, thus visibly affecting the resulting speech. This effect may explain why certain formants were attenuated in a manner such that there were barely observable above the EL self-noise. Thus, the effect of the spectral zeros on EL speech could potentially be quite significant.

4.4.2. Limitations

While the VA database provided a unique opportunity to analyze the pre- and post-laryngectomy speech of the same speakers, it must be noted that the pre-laryngectomy voices of many the speakers were at least somewhat disordered. This is not surprising given that the disordered quality of their voices was probably one of the symptoms fo their laryngeal cancer. As such, the measurements conducted on these speakers' pre-laryngectomy speech may differ from those that would have been made on the speech of more normal speakers. Many of the voices had pressed voice quality which most likely affected some of the measurements. Despite this limitation, it was felt that the benefit of being able to compare the pre- and post-laryngectomy speech within a single speaker outweighed the drawbacks of using somewhat disordered speech.

4.5. Summary

An analysis of the pre- and post-laryngectomy speech of EL users has revealed several differences between the speech conditions all of which could have a noticeable impact on the quality of EL speech. Some conditions, such as the low frequency deficit of EL speech, have been well studied, while others, such as the spectral zeros found in EL speech spectra have not. As the perceptual results of Chapter 3 demonstrated, there must be factors other than the low frequency deficit, the self-noise, and lack of pitch information that contribute to the unnatural quality of EL speech. The findings discussed in this chapter suggest that the spectral zeros found in EL speech could have an adverse effect on EL speech quality and warrants a thorough investigation. Such an investigation is carried out in the next chapter.

5. Acoustic effects of source location on EL speech.

5.1. Introduction

The analysis of both the pre- and post- laryngectomy speech of several laryngectomy patients in the preceding chapter revealed that the relative formant amplitudes differed between normal and EL speech. This discrepancy can largely be attributed to difference between spectral content of the normal and EL voicing sources. Yet, even among the EL speech spectra there were differences in the relative formant amplitudes that could not be adequately explained by individual disparities in vocal tract acoustics. In several instances, one or more formants appeared to be attenuated by nearby spectral zeros, suggesting that these zeros can have an adverse effect on the quality (and possibly the intelligibility) of EL speech. Myrick and Yantorno (1993) demonstrated that positioning the voicing source away from the end of the vocal tract introduces these zeros into the vocal tract transfer function. However, the Myrick and Yantorno study only investigated a single vowel and reported results for only one location along the vocal tract. Speech contains a number of different phonemes and different EL users place the EL at different positions along the neck. To more comprehensively examine this phenomenon, the study described in this section modeled the effect of source location for several vowels and at several positions along the length of vocal tract and compared the model results with recorded speech data.

5.2. Theoretical Background

In normal speech production, the voicing source is at the glottis, located at the terminal end of the vocal tract. In EL speech production, the voicing source is positioned at some point in the middle of the vocal tract thus altering the acoustics of the system. To illustrate the effect of source location, consider a uniform tube model of the vocal tract. Although a uniform tube greatly simplifies the shape of the vocal tract, it is still useful for understanding the impact of moving the excitation source away from the terminal end.

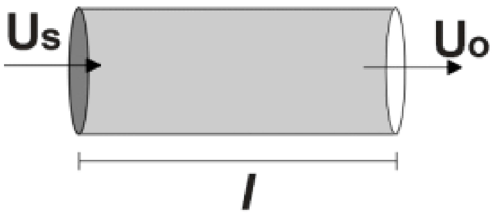


Figure 5.1 Schematic of a uniform tube with the driving source, U_s at one end of the tube.

Suppose we have a uniform tube of length, *l*, and area, *A*, with an input volume velocity source, U_s, and an output volume velocity, U_o, as shown in Figure 5.1. This configuration is a good representation of the schwa vowel /∂/ spoken with a normal voice. The Vocal Tract Transfer Function is defined as:

$$VTTF(f) = \frac{U_o(f)}{U_s(f)} = \frac{Zeros(f)}{Poles(f)}. \quad (5.1)$$

In this case, VTTF(*f*) consists only of an infinite number of poles (*i.e.* an all pole transfer function) whose frequencies are independent of the source, and are determined by the length of the tube. Specifically, the pole frequencies, fp_n are found at

$$fp_n = \frac{(2n-1)c}{4l} \qquad n = 1,2,... \quad (5.2)$$

where *c* is the velocity of sound in air.

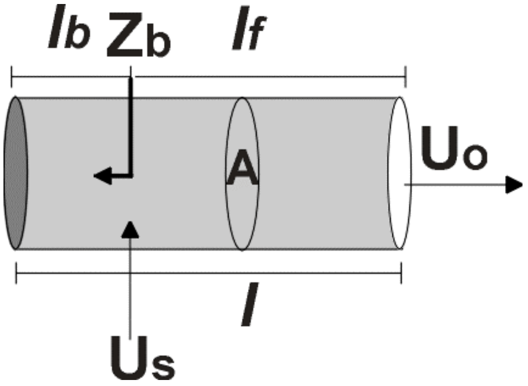


Figure 5.2. Schematic of a uniform tube with the driving source, U_s at a distance l_b from the end of the tube.

In the case of EL speech, however, the situation is changed, as displayed in Figure 5.2. Here, although it is assumed that the source is a volume velocity source, it is no longer at the end of the tube, but located at a distance, l_b, from the end. The VTTF is still defined as in Eq. 5.1; however, while the poles remain the same as in Eq. 3.2, the VTTF now contains an infinite number of zeros as well. The locations of the zeros are determined by which frequencies make the impedance looking into the back cavity, $Z_b(f)$ equal to zero, causing the back cavity to act as a short circuit. The back cavity is effectively a uniform tube with a rigid termination at one end, and thus the impedance, $Z_b(f)$, can be written as,

$$Z_b = -j\frac{\rho c}{A}\cot\left(\frac{2\pi f l_b}{c}\right). \tag{5.3}$$

The zero frequencies, fz_m, occur at

$$fz_m = \frac{(2m-1)c}{4l_b} \qquad \mathrm{m} = 1,2,\ldots \tag{5.4}$$

With Eqs. 5.1, 5.2, and 5.4, the VTTF can now be computed for different source locations, i.e. different l_b as shown in Figure 5.3. The transfer functions in this figure were computed using a tube length, l, of 17.7 cm., a speed of sound, c, of 35,400 cm/s and three different source locations, l_b, 0 cm, 5.9 cm, and 8.0 cm. These particular locations were chosen to emphasize the effects of the zeros on the VTTF. In everyday use, an EL can be held at almost any location along the length of the vocal tract, depending on the user's anatomy, comfort, and location of the spot where EL sound transmission through the neck is most efficient (i.e. the sweet spot). Therefore, the effect of the zeros will vary from user to user.

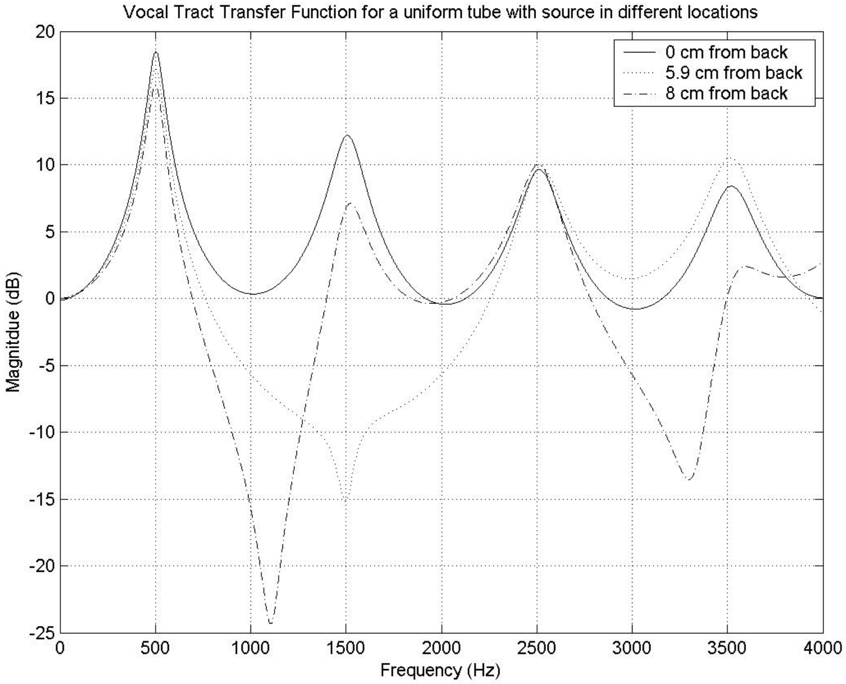


Figure 5.3. Computed vocal tract transfer functions for different source locations.

Setting l_b equal to 0 cm is equivalent to reproducing the situation of natural speech. The source is at the end of the tube, thus producing an all pole transfer function. The first four poles (formants) are found at 500 Hz, 1500 Hz, 2500 Hz, and 3500 Hz as shown. It should be noted that the radiation impedance at the lips was neglected when computing these formant frequencies. The radiation impedance effectively extends the length of the vocal tract, thus lowering the formant frequencies. However, in this case, accounting for this effect is not necessary since even if the formants were decreased, the source location could be altered to compensate for the change in formant frequencies.

When l_b is set at 5.9 cm, the first two zeros occur at 1500 Hz and 4500 Hz. The first zero coincides with the second formant, essentially canceling each other out, thus making the third pole appear to be the second formant. In the third source position, l_b = 8 cm, the lowest frequency zero falls at 1100 Hz, in between the first two formants. Both formants are attenuated, although the attenuation of the second formant is larger due to its closer proximity to the zero. This source position also produces another zero that has a low enough frequency to be visible in this plot. While this zero reduces the amplitude of the fourth formant, it probably would have little effect on the perception of the vowel.

It is clear that changing the source location affects the spectrum of the vocal tract transfer function, and in certain cases can actually effectively eliminate a formant. It is likely that the modified vocal tract acoustics affect the perception of EL speech. However, this simple tube model is overly simplistic since vowels are produced with a more complex vocal tract configuration than a uniform tube. These more complex configurations will change the impedance of the back cavity (depending on the source location), thus altering the frequencies of the zeros. Therefore a model that is more representative of vocal tract shapes is needed to more accurately predict the effects of changing the source location. The details of such a model as well as a comparison of its output with experimental data are discussed in the following sections.

5.3. Methods

This study consists of two parts: an experimental component and a modeling component.

5.3.1. Experimental Component

Two normal (i.e. non-laryngectomized) speakers, one male and one female, produced two sentences using both their natural voices and a neck-placed Servox electrolarynx (Siemens Corp.). Both speakers were proficient at speaking with an electrolarynx. The speakers were instructed to hold their breaths and maintain a closed glottal position while talking with the Servox, in order to approximate the anatomical condition of laryngectomy patients. The subjects were asked to sustain 10 vowels: /i/ as in “beet”, /I/ as in “bit”, /ɛ/ as in “bet”, /æ/ as in “bat”, /a/ as in “bot”, /ɔ/ as in “bought”, /u/ as in “boot”, /U/, as in “put”, /^/ as in “but”, and /e^r/ as in “bert” both with their normal voice and with the Servox. To minimize the differences in the formant frequencies between the same vowel spoken in the two different conditions, the subjects began each task by producing the vowel with their normal voices and then closed off the glottis in mid-vowel, while simultaneously activating the EL device. This procedure was repeated three times, each with the EL placed at one of three places on the neck: at approximately the level of the glottis, and at approximately 1/3 and 2/3 of the way up the length of the neck. Recordings were made with the subjects' faces sealed in a specially constructed port in the door of a sound isolated booth (see Appendix B). This was done to essentially eliminate the self-noise of the neck placed EL from the audio recording of the speech. All recordings were made with a Sennheiser (Model K3-U) microphone placed 15 cm. from the lips. The speech signals were low pass filtered at 20 kHz by a 4 pole Bessel Filter (Axon Instruments Cyberamp) prior to being digitized at 100 kHz (Axon Instruments Digidata acquisition board and accompanying Axoscope software). The signals were then appropriately low pass filtered and downsampled to 12 kHz in MATLAB. This sampling rate was chosen to allow for the observation of any higher frequency anti-resonances that may have been present in the EL speech spectra.

For each vowel, the spectrum was computed by performing a 4096-point Discrete Fourier Transform on a 50 ms hamming windowed section of the vowel. Linear predictive (LP) coefficients were computed for this vowel segment as well. The first three formant frequencies (*F1, F2, F3*) and amplitudes (*A1, A2, A3*) were measured by manually marking the approximate frequency locations of the formants and then finding the harmonic with the largest amplitude within 100 Hz of the marked frequency location. . Based on these measurements, the relative formant amplitudes (*A1-A2, A2-A3, A1-A3*), were calculated. In addition, the frequency locations of any observed spectra zeros were recorded.

5.3.2. Vocal Tract Modeling

The vocal tract modeling required four successive steps that are described as follows:

5.3.2.1 Step one: Determining Vocal Tract Area Function

Previous studies have reported the vocal tract shapes for different vowels, (Chiba & Kajiyama 1941, Story *et al.* 1996), but all of the reported vocal tracts would produce formant frequencies that differed from the ones measured in this study since they were obtained from different individuals. As such, a different method was required to determine that proper vocal tract area functions. Story and Titze's (1998) algorithm based on principal components analysis, determines a vocal tract area function based on the frequencies of first two formants of a vowel. The mean values of F1 and F2, obtained from the three natural speech recordings of each of the 10 vowels were used to generate the area functions. This algorithm assumed that the vocal tract lengths of the male and female speakers were 17.5 cm and 14.5 cm respectively and produced vocal tract area functions that were divided into 17 and 14 sections of equal length (of 1.029 cm 1.036 cm). However, since the third formant is not used in the algorithm, the third formants associated with the generated vocal tract area function noticeably differed from those measured from the speech recordings. As such, the vocal tract areas were corrected using a Matlab implementation of Maeda's VTcalcs (1992) program, which generates a vocal tract transfer function for a given vocal tract area. The individual cross sectional areas were adjusted so that the algorithm generated third formants more closely matched with the measured data.

Because the vowel /e^r/ is very much distinguished by its low third formant, Story and Titze's algorithm was unable to produce an accurate area function. Instead, an area function for the phoneme /r/ was modified using VTcalcs such that the resulting formants approximated those from the measured vowels.

5.3.2.2 Step two: Determining Vocal Tract Impulse Response

Unfortunately, the VTcalcs software does not allow the vocal tract excitation source to be placed anywhere other than at the end of vocal tract. Therefore, another software package, LAMINAR, was used to produce vocal tract impulse responses. LAMINAR uses an enhanced version of the Kelly-Lochbaum (1962) model, which like the VTcalcs software, divides the vocal tract into several smaller segments of equal length. Each element is specified by its cross-sectional area, a shunt loss factor, a series loss factor, and either a volume velocity or pressure source as shown in Figure 5.4. The series loss factor, $D = \frac{R}{Z_a + Z_b}$, where Z_a and Z_b are the characteristic impedances of the adjacent elements, and $Z = \frac{\rho c}{A}$. R is the series resistance of the lossy element and is related to losses associated with airflow in the vocal tract (Liljencrants 1985). Similarly, the shunt loss factor, $E = \frac{G}{Y_a + Y_b}$, where Y_a and Y_b are the characteristic conductances of the adjacent elements. G is the shunt conductance of the lossy element and incorporates heat conduction losses and losses associated with the impedance of the vocal tract walls. (Liljencrants 1985).

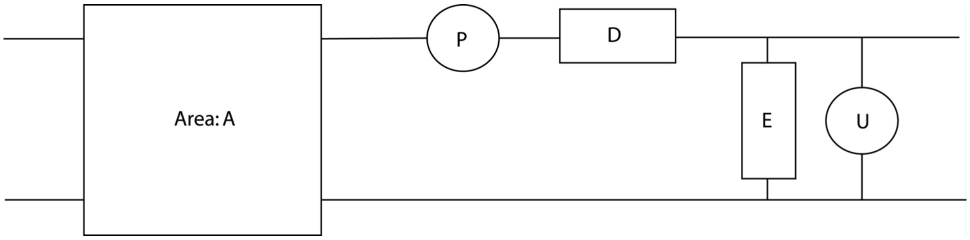


Figure 5.4. Schematic of an individual element in LAMINAR. Each element is specified by an area, *A*, a series loss factor, *D*, a shunt loss factor, *E*, a pressure source, *P* and a volume velocity source, *U*.

Based on the values in Liljencrants (1985). *D* and *E* were set equal to 0.0422 and 0.0538. Although the two loss factors are dependent on the areas of the adjacent elements and thus should vary from element to element, they were set at constant values. The losses were so small in magnitude that applying small area related changes to them would have little impact on the vocal tract impulse response.

The shunt loss of the first element (at the glottis) also includes glottal losses. However, because the model was being used to simulate EL speech where (in laryngeal speakers) the glottis is closed, the glottal shunt loss was set to zero. At the other end of the vocal tract, the series loss of the final segment (at the lips) also includes the loss due to the radiation impedance. Finding the proper value for the radiation loss was problematic because while LAMINAR only permits the loss factor to be constant it is in fact dependent on frequency. Specifically, the radiation resistance can be written as

$$R_{rad} = \frac{\rho \pi f^2 K_s}{c} \tag{5.5}$$

where ρ is the density of air, *c* is the velocity of sound, and K_s is a frequency dependent resistance factor. For a simple source K_s is unity while for a piston in an infinite baffle, $K_s = 2$ (Stevens 1998). It can be shown that for uniform tube of length, *l,* and of area, *A*, the contribution of the radiation resistance to the formant bandwidths is:

$$B_r = \frac{f^2 A K_s}{lc} \tag{5.6}$$

As Eq. 5.6 shows, the contribution to the formant bandwidths is proportional to square of frequency, meaning that higher frequency formants will have much wider bandwidths. Because LAMINAR does not provide for a frequency dependent radiation loss, the resulting higher formant bandwidths will be far too narrow, thus producing an inaccurate vocal tract transfer function. Nevertheless, the LAMINAR model remained useful for determining the formants and the zeros of the vocal tract transfer function. As such, a series of transfer functions were generated for each vowel by moving a volume velocity source along the length of the vocal tact, one element at a time, up to and including the penultimate element. This produced 16 transfer functions for the male model and 13 for the female.

5.3.2.3 Step Three: Producing the Final Vocal Tract Transfer Functions

The vocal tract transfer function can be represented as the ratio of conjugate pairs of poles and zeros, i.e.

$$T(s) = \frac{\prod_m (s - s_m)(s - s_m^*)}{\prod_n (s - s_n)(s - s_n^*)} \tag{5.7}$$

where $s = j2\pi f$, and *m* and *n*, are the number of zeros and poles, respectively. The conjugate poles (and zeros) can be written in terms of a center frequency and bandwidth:

$$s_n = \frac{bw_n}{\pi} + j2\pi F_n \tag{5.8}$$

where the bandwidth is the half-power bandwidth.[4] Thus, a more accurate vocal tract transfer function can be computed by combining the poles and zeros measured from the LAMINAR outputs with the proper bandwidths. The first three bandwidths used in this study were the mean values obtained from measurements of the three recordings of each vowel. The bandwidths of formants that were significantly attenuated were not used in this computation. Because it was difficult to discern the higher formants in the recorded speech spectra, the higher formant bandwidths, F4, F5, F6, and F7 were set at 450 Hz, 635 Hz, 650 Hz and 650 Hz respectively. These values were chosen to approximate the broadness of any visible higher formants. Using this technique, the LAMINAR outputs were converted into new vocal tract transfer functions for each speaker and for all vowels and source positions. Thus, for each of the ten vowels, a total of 16 vocal tract transfer functions (13 for the female case) were generated.

Although using measured bandwidth data in the vocal tract models may appear to be somewhat circular, it should be pointed out that the purpose of the models are to predict the vocal tract behavior when the source is moved. Fixing the bandwidths in this fashion helps ensure that the model outputs are not adversely affected by inadequacies in the model.

Finally, these transfer functions were generated in the discrete domain so that the poles located above the Nyquist frequency served to act as higher frequency poles that occur in the physical world. The sampling frequency used in this case was the same specified by the LAMINAR software and is a function of vocal tract length, number of sections and the velocity of sound.

5.3.2.4 Step Four: generating model based EL speech

An estimate of the Servox excitation signal was made by filtering the measured acceleration produced by a loaded Servox with the estimated neck frequency response functions reported in Meltzner *et al.* (2003). The acceleration was measured using an Endevco model FC-11 accelerometer that was attached to the vibrating head of the Servox. The spectrum of this signal is shown in Figure 4.14. The excitations were then filtered with the model generated vocal tract transfer functions and then differentiated (to account for the effects of the lip radiation characteristic), thus producing a set of synthesized sustained ELl vowels. Measurements of the relative formant amplitudes of the synthesized vowel spectra as well as the frequency locations of any observed spectral zeros were done using the procedure described for the recorded speech vowels described in section 5.3.1.

[4] Although it is not common to refer to the bandwidth of a zero pair, one could define an analogous quantity that refers to the difference between the frequencies at which the frequency response function of the zero pair is 3 dB greater than its minimum value.

5.4. Results

5.4.1. Recorded Speech

5.4.1.1 Relative Formant Amplitudes

The spectra of the vowel /U/ for the male speaker shown in Figure 5.5 is useful in illustrating the typical effects of moving the EL source from the level of the glottis towards the chin. The spectrum of the vowel spoken with a normal voice is provided at the top of the figure for reference. When compared to the spectrum of the normal vowel, the spectrum of the EL vowel at position 1 demonstrates several differences in the spectral attributes that were discussed in the previous chapter: a reduced amount of low frequency energy, narrower bandwidths, and different relative formant amplitudes. When the source is placed at position 2, there is change in the relative formant amplitudes where the amplitudes of the second and third formants have decreased relative to that of the first formant. It is likely that the apparent zero at 2280 Hz is causing this attenuation. As the EL is moved closer towards the lips in position 3, the formant amplitudes change once again, with A1, A2 and A3 decreasing by 18.1 dB, 7.7 dB and 10.4 dB respectively. As a result, the spectrum at position 3 resembles that of position 1 except for the decreased prominence of F3, whose amplitude looks to be attenuated by a zero at 2760 Hz.

The relative formant amplitudes of all ten vowels at all three positions presented in Table 5.1 clearly demonstrate that while EL source location affects the relative formant amplitudes for all vowels, the relationship between vowel and the effect of the source position is complex. In general, as shown by the increases in *A1-A2* and *A1-A3*, as the EL device was moved from position 1 (at the level of the glottis) to position 2 (1/3 of the way up the neck), both F2 and F3 decreased in amplitude relative to F1. This trend was more pronounced for the female speaker as the average increase in *A1-A2* from position 1 to position 2 was 5.4 ± 5.1 dB and 11.5 ± 4.5 dB for the male and female speaker respectively. The notable exception to this rule was the vowel /u/ for the male speaker which actually showed a decrease in *A1-A2*. Similarly, the mean increase in *A1-A3* was 3.8 ± 6.0 dB for the male speaker and 14.3 ± 9.3 dB for the female speaker. The situation is less straightforward for *A2-A3*. When the EL is moved from position 1 to position 2, *A2-A3* decreased in 5 vowels for both speakers, but with /ɔ/ being the only vowel in common between them.

Relocating the EL to position 3 once again produced vowel dependent changes in the relative amplitudes. In many cases, such as the female vowels /u/ and /ɛ/, the relative amplitudes reverted to values similar to those found at position 1. Yet for several other vowels, the relative amplitudes deviated even further from those measured at original position.

5.4.1.2 Spectral Zeros

Zeros were observed in the vowel spectra at all positions, although there were a number of spectra in which the presence of a spectral zero was not obvious. The measured zeros for all vowels grouped by position for both the male and female speakers are shown in Figure 5.6. The data in the figure indicate that the frequencies of the spectral zeros were dependent on the source location. The presence of the zeros in the spectra measured while the EL was in position 1 was unanticipated, as it had been intended for the EL to be located at the level of the glottis. The zeros were found between 3800 Hz and 5000 Hz for both speakers and to produce zeros at these frequencies, the EL would have had to be placed between 1.8 and 2.2 cm from the glottis. Despite the difficulty in estimating the exact location of the glottis, it is still rather surprising to have erred by as much 2 cm.

At position 2, there is a visible shift in the zero frequencies. For the female speaker, the majority of zeros were found between 2200 Hz to 3000 Hz although there was another small group centered around 4200 Hz. Similarly, for the male speaker, the zeros were about evenly divided between a group centered at 2000 Hz and another around 5000 Hz.

The zeros again shift as the EL is moved to position 3. For the male speaker, the zeros can be separated into three groups: one concentrated around 3000 Hz, another around 2000 Hz and one below 1000 Hz. Only two clusters of zeros, one between 3000 Hz and 5000 Hz and another between 1000 Hz and 2000 Hz, were observed for the female speaker.

The shifting of the zeros to lower frequencies as the EL is moved further away from the level of the glottis is consistent with the basic theory described in Section 5.2 and is at least qualitatively similar to effects displayed in Figure 5.3. As the EL is moved away from the glottis, the length of the back cavity increases and produces lower frequency zeros, as predicted by Eq. 5.4. Eq. 5.4 also predicts that zeros will occur at multiples of the quarter-wave frequency, which would explain the presence of multiple groups of zeros at positions 2 and 3. Only one group is observed at position 1 because the frequencies of the other groups are beyond the bandwidth used in this study.

However, the large spread of the zeros is not predicted by the basic theory because it assumes that the back cavity is a uniform tube. The vocal tract area functions of different vowels deviate quite a bit from a uniform tube and from each other. How the different area functions affect the frequencies of the zeros is explored in the following section.

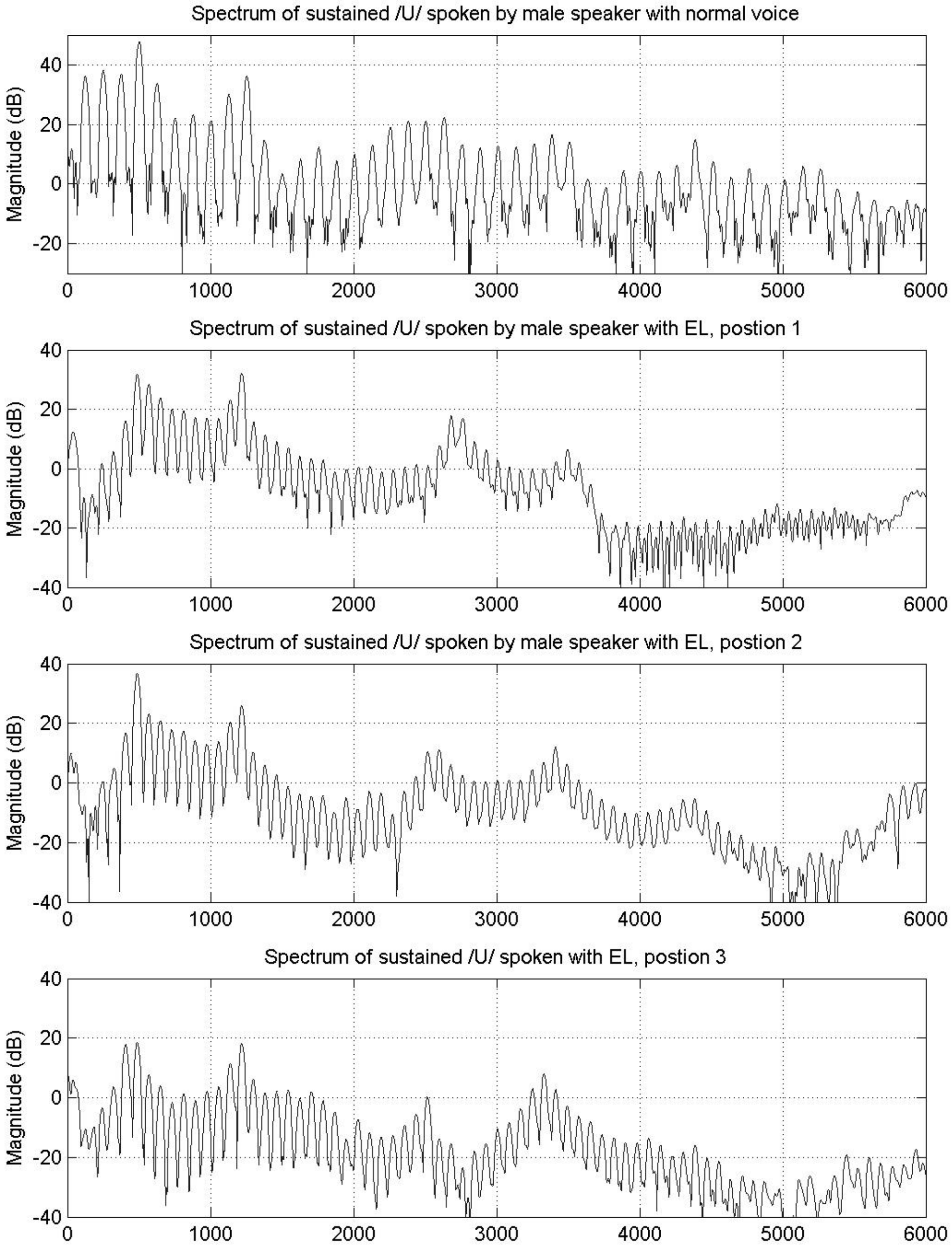
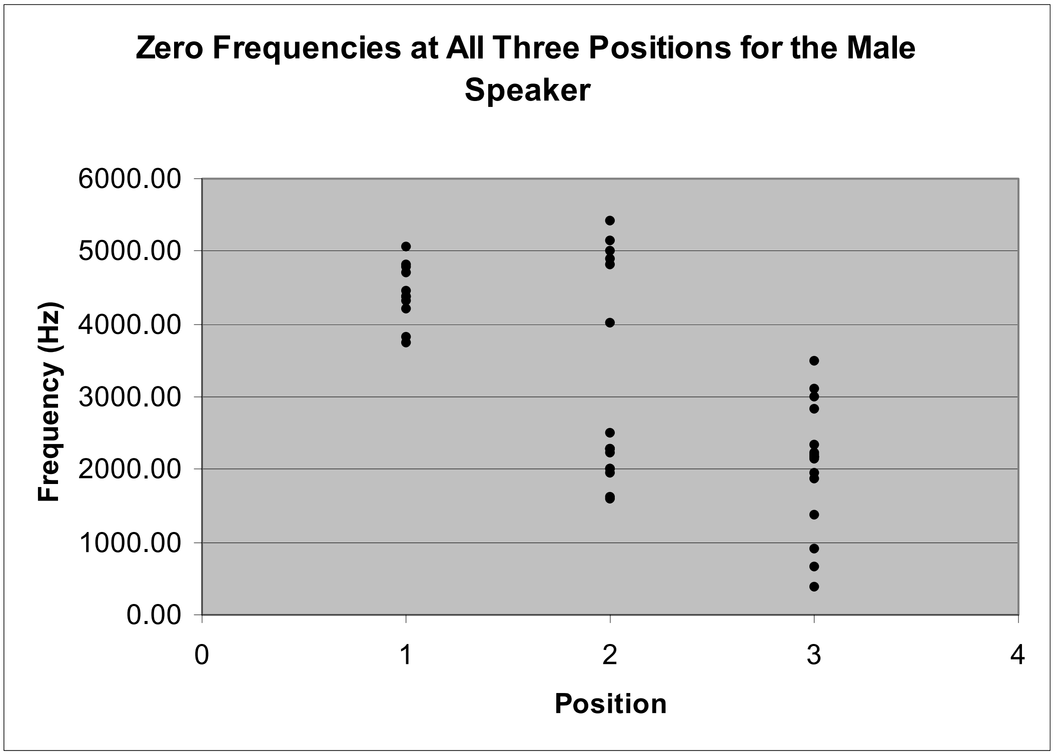


Figure 5.5. The spectrum of the vowel /U/ spoken by a male speaker with normal voice and with a Servox EL at three different positions on the neck. The spectra indicate that the amplitudes of all three formants are dependent on the location of the EL sound source.

Table 5.1. The Relative Formant Amplitudes for 10 Vowels spoken by a Male and Female Speaker.

		Female Speaker			Male Speaker		
Vowel		A1-A2 (dB)	A2-A3 (dB)	A1-A3 (dB)	A1-A2 (dB)	A2-A3 (dB)	A1-A3 (dB)
/i/	Pos 1	-3.1	-2.8	-5.9	-11.8	11.9	0.0
	Pos 2	3.9	22.1	25.9	-6.1	9.9	3.8
	Pos 3	1.9	6.9	8.8	-8.5	7.5	-1.0
/I/	Pos 1	5.7	1.2	6.9	1.1	7.7	8.8
	Pos 2	19.7	8.0	27.7	12.9	2.9	15.8
	Pos 3	5.3	7.0	12.3	9.5	3.1	12.7
/ɛ/	Pos 1	7.2	3.1	10.3	4.0	7.9	11.9
	Pos 2	18.2	17.8	35.9	9.7	5.3	15.0
	Pos 3	10.9	0.2	11.1	16.1	1.2	17.3
/æ/	Pos 1	3.2	12.5	15.7	6.1	10.2	16.3
	Pos 2	13.5	19.4	32.9	12.4	10.9	23.3
	Pos 3	14.2	4.2	18.5	20.9	6.1	27.0
/a/	Pos 1	9.5	14.1	23.6	3.1	16.6	19.7
	Pos 2	12.6	21.0	33.6	8.1	12.1	20.2
	Pos 3	20.0	2.3	22.2	14.7	7.4	22.2
/ɔ/	Pos 1	3.1	21.2	24.3	5.8	21.6	27.4
	Pos 2	14.6	12.2	26.7	8.1	8.9	17.0
	Pos 3	8.7	12.5	21.1	21.4	5.9	27.3
/U/	Pos 1	5.5	8.3	13.8	-0.5	14.5	14.0
	Pos 2	19.8	5.1	24.9	10.8	14.7	25.5
	Pos 3	6.8	2.9	9.7	0.4	17.8	18.2
/u/	Pos 1	-3.9	11.2	7.2	-5.0	17.2	12.2
	Pos 2	9.8	7.4	17.2	-10.2	23.7	13.5
	Pos 3	-5.3	10.3	5.0	-4.3	23.2	18.9
/^/	Pos 1	8.4	10.1	18.5	7.5	12.3	19.8
	Pos 2	18.3	5.6	23.9	4.9	12.9	17.8
	Pos 3	12.6	7.0	19.6	6.1	9.7	15.8
/e^r/	Pos 1	7.0	12.5	19.5	5.8	1.8	7.7
	Pos 2	26.9	1.4	28.2	7.2	9.4	16.6
	Pos 3	2.9	9.2	12.1	2.2	17.7	19.8



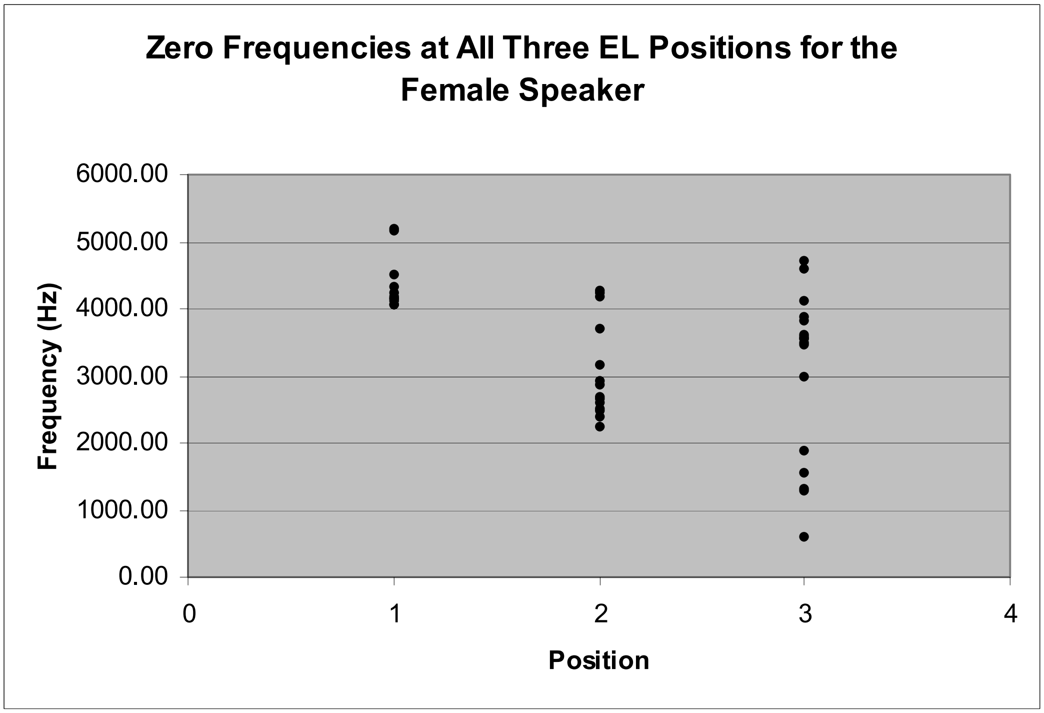


Figure 5.6. The zero frequencies at all three positions for the male (*top*) and female (*bottom*) speaker.

5.4.2. Vocal Tract Modeling

5.4.2.1 General Discription

Figure 5.7 displays an example of the model output, in this case for the vowel /a/ produced by the male vocal tract. (Similar figures for each vowel can be found in Appendix F.) Each panel in the figure represents the spectrum of the vowel produced with the excitation source at a different position in the vocal tract. The spectrum of position 1 is omitted because it was almost exactly the same as that of position 0 within the specified frequency range (the zero was located at 8600 Hz). The effect if moving the voicing source further away from the end of the vocal tract is clearly displayed in this figure. At position 0 (i.e. at the glottis), no spectral zeros are visible, but at position 2, a zero is clearly visible at about 4500 Hz. As the source is moved further up the vocal tract, the frequency of the zero decreases. At position 4, the proximity of the zero to the third formant causes a severe attenuation of the formant. The presence of a second, high frequency zero is also visible at position 4. The spectrum at position 5 is a useful example of how in certain cases the source location does not noticeably affect the relative formant amplitudes. The lowest frequency zero falls directly between the second and third formants, causing only a small change in their relative amplitudes. The effect on of the high frequency zero (at 5000 Hz) on the first three formants is limited as well. As the source is moved even further away from the glottis, the number of zeros in the spectrum increases until, at position 16, the formants and zeros appear to be interleaved.

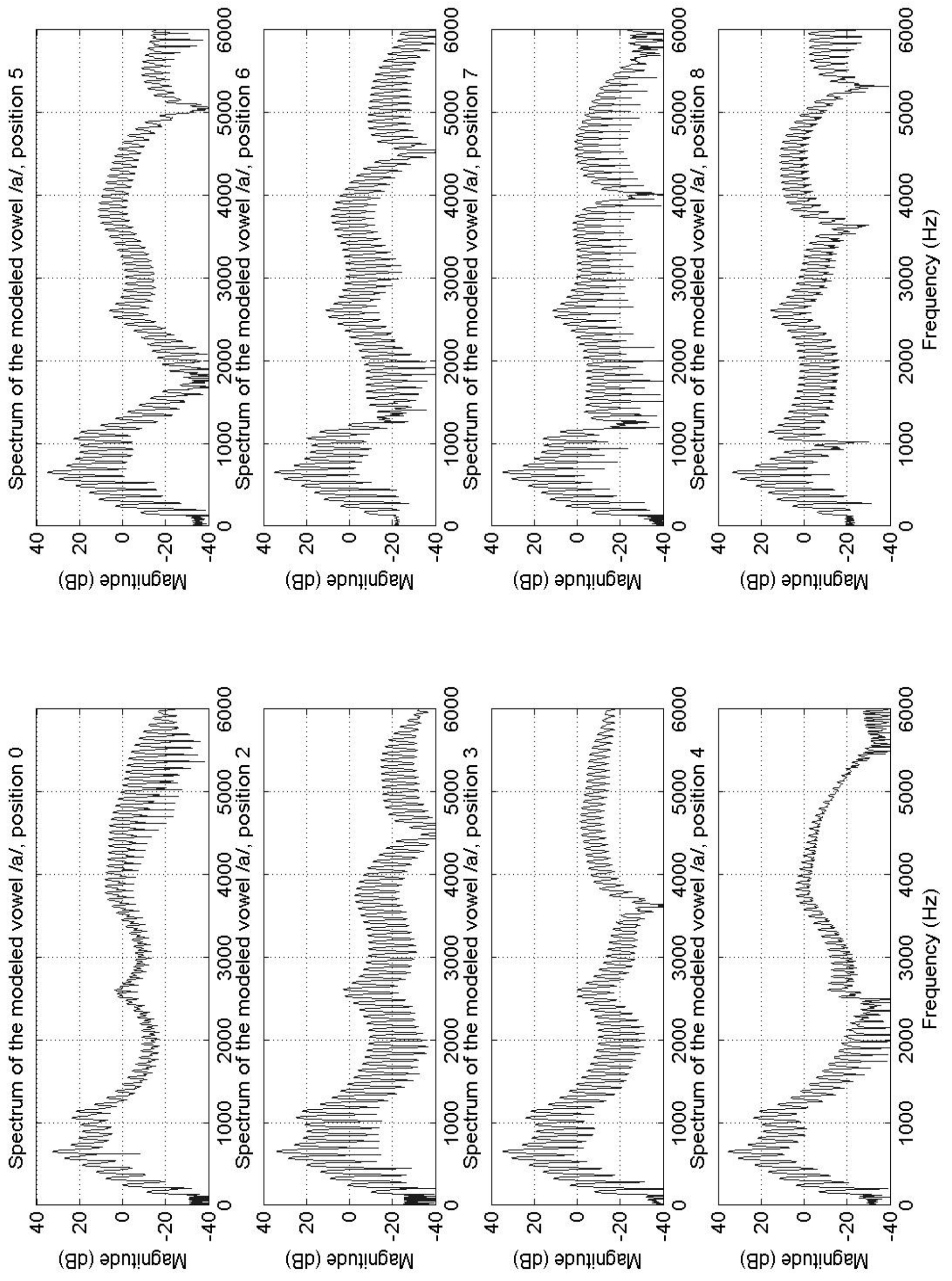


Figure 5.7a. The spectra of the vowel /a/ for positions 0 through 9. The spectrum of Position 1 is omitted because within in the viewable frequency range, it is indistinguishable from that of Position 0.

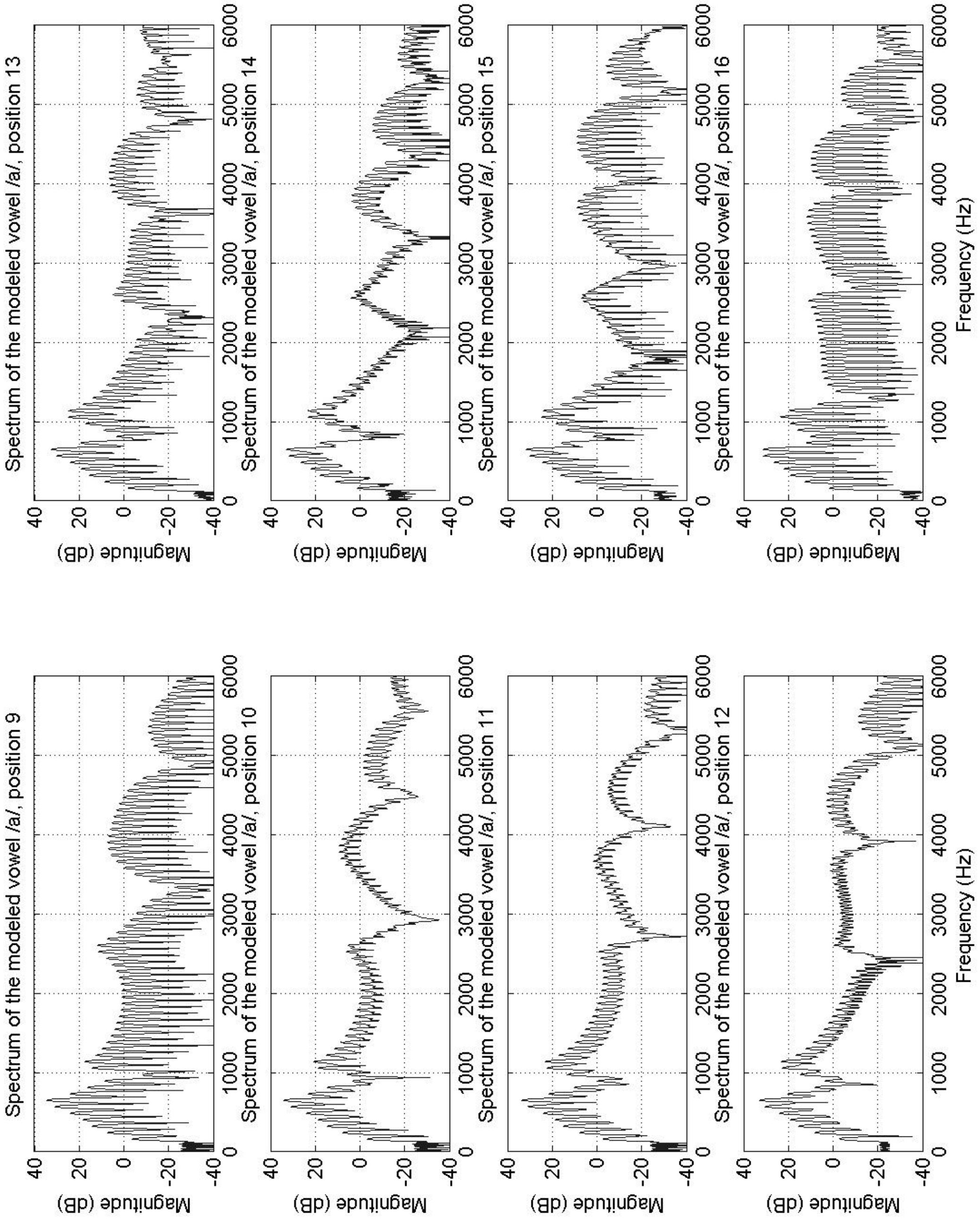


Figure 5.7b. Spectra of the model output for the male vowel /a/ when the source was placed at positions 9 through 16.

This set of vowel spectra also illustrates how the relative formant amplitudes are affected by the location of the EL voicing source. The variability in relative formant amplitudes is made more apparent in Figure 5.8. It is clear that the formant amplitudes are greatly dependent on source location, with the greatest deviation occurring at about the midpoint of the vocal tract (position 8).

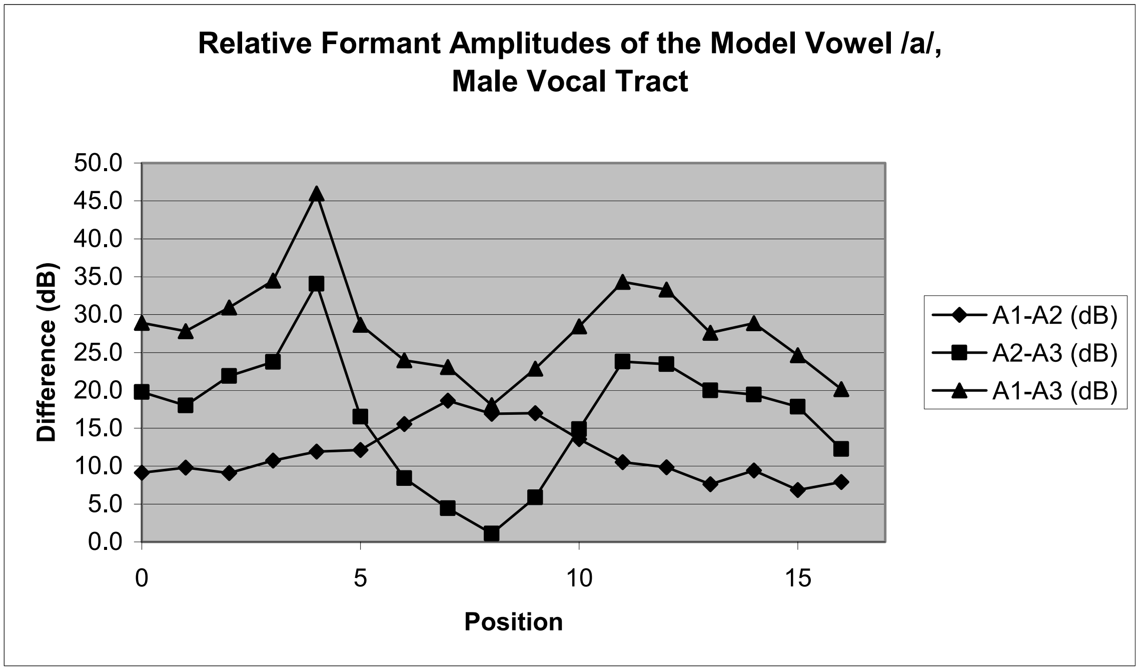


Figure 5.8. The relative formant amplitudes of the model vowel /a/ for the male speaker at each source position. The positions 0, 5, and 10 correspond to to the recorded Positions 1, 2,and 3 respectively.

5.4.2.2 Comparison with Recorded Speech

One of the main goals of this study was to determine how effectively the model of changing the source position captured the observed behavior of recorded EL speech. To this end, three modeled source locations were selected to approximate the locations used in the recording experiment. For the male vocal tract, the source positions were positions 0, 5, and 10 while for the female vocal tract the positions were 0, 4, and 9. While the latter two positions in both models are somewhat shy of the one-third and two-thirds points along the vocal tracts, the modeled spectra at those points most closely matched with the recorded spectra. The reason for the positional discrepancy is most likely due to the uncertainty of the exact location the EL was placed along the vocal tract. From now on, unless otherwise specified, the chosen source locations for the model will be referred to as Positions 1, 2, and 3.

Table 5.2. Relative Formant Amplitudes For the Modeled Vowels.

		Female Vocal Tract			Male Vocal Tract		
Vowel		A1-A2 (dB)	A2-A3 (dB)	A1-A3 (dB)	A1-A2 (dB)	A2-A3 (dB)	A1-A3 (dB)
/i/	Pos 1	6.6	2.8	9.4	-5.0	-0.7	-5.6
	Pos 2	15.9	14.4	30.3	7.0	10.4	17.4
	Pos 3	1.6	6.8	8.5	-1.9	3.6	1.7
/I/	Pos 1	16.2	5.0	21.3	9.0	7.3	16.4
	Pos 2	16.1	10.1	26.3	13.9	22.8	36.6
	Pos 3	-7.3	17.2	9.8	7.5	10.8	18.3
/ε/	Pos 1	13.3	3.2	16.5	8.9	5.2	14.1
	Pos 2	18.6	19.2	37.8	21.7	7.5	29.2
	Pos 3	6.2	9.3	15.6	8.0	11.4	19.4
/æ/	Pos 1	9.9	7.9	17.7	11.3	9.0	20.3
	Pos 2	12.1	15.7	27.8	17.1	18.3	35.4
	Pos 3	11.8	6.5	18.4	14.6	12.7	27.3
/a/	Pos 1	14.7	14.4	29.0	9.1	19.8	28.9
	Pos 2	16.4	33.0	49.4	12.1	16.5	28.7
	Pos 3	26.5	2.2	28.7	13.6	14.9	28.5
/ɔ/	Pos 1	9.4	17.1	26.5	11.3	18.5	29.8
	Pos 2	12.7	9.5	22.2	12.4	19.1	31.5
	Pos 3	9.3	3.8	13.1	17.7	13.4	31.1
/U/	Pos 1	10.3	11.5	21.8	7.7	11.4	19.1
	Pos 2	18.2	11.7	29.9	17.3	15.1	32.5
	Pos 3	17.4	6.1	23.5	7.3	17.9	25.2
/u/	Pos 1	12.8	14.8	27.6	5.6	21.9	27.5
	Pos 2	16.4	21.2	37.6	4.9	28.6	33.5
	Pos 3	-6.9	23.3	16.5	7.2	16.4	23.6
/^/	Pos 1	12.9	12.1	25.0	10.7	14.0	24.8
	Pos 2	15.5	18.7	34.1	13.5	32.0	45.5
	Pos 3	14.5	8.9	23.3	11.8	15.2	27.0
/e^r/	Pos 1	13.2	8.3	21.5	6.2	8.2	14.4
	Pos 2	19.2	20.3	39.4	14.3	21.3	35.7
	Pos 3	25.2	-1.3	24.0	17.5	4.4	21.9

Table 5.2 presents the relative formant amplitudes of all ten vowels at the three positions for both vocal tract models. Consistent with the experimental results, the vocal tact models also demonstrate a complicated relationship between source location and relative formant amplitudes. Once more, as the EL source is moved from position 1 to position 2, both *A1-A2*, and *A1-A3* exhibit notable increases and again, the notable exception to this trend was the male vowel /u/. However, unlike the recorded the data, this effect was slightly greater for the male vocal tract than the female. The mean change in *A1-A2* and *A1-A3* was 6.0 ± 4.6 dB and 13.7 ± 8.4 dB for the male vocal tract and 4.2 ± 2.9 dB and 11.9 ± 8.3 dB for the female vocal tract. Furthermore, the model also somewhat deviates from the experimental data in regards to the change in *A2-A3* between positions 1 and 2. The experimental data indicated that the number of vowels for which *A2-A3* increased was the same as the number for which it decreased, whereas only one modeled vowel for each speaker showed a decrease in *A2-A3*.

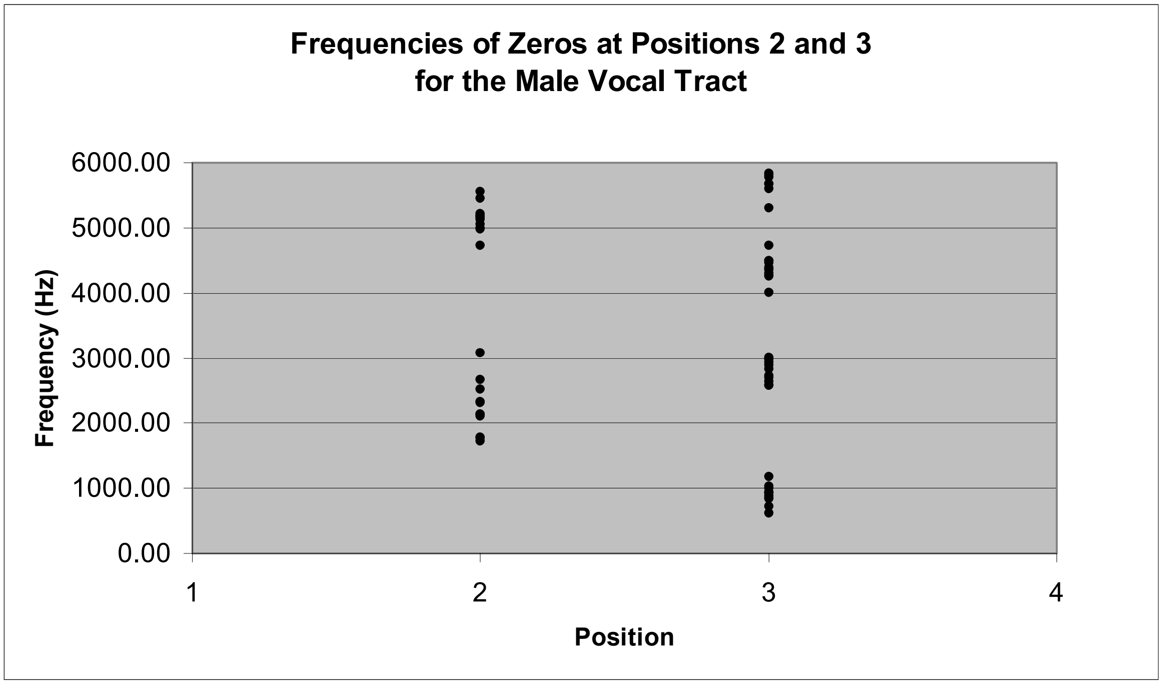
As was the case for the experimental spectra, as the voicing source is moved further away from the glottis to position 3, the change in the relative formant amplitudes was very much dependent on the vowel. For some vowels such as /^/, the relative formant amplitudes returned to values similar to those found at position 1. For most of the modeled vowels, however, the opposite was true – the formant amplitudes deviated even further from the values found at the original source position.

A more quantitative comparison between the model and the recorded speech was made by subtracting the relative formant amplitudes measured from the modeled spectra from those measured from the recorded spectra. The results of this comparison, are given in Table 5.3. The prevalence of negative values in the table indicates that the models typically overestimated the difference in the formant amplitudes at the different source locations. The most conspicuous example of this overestimation occurs for *A1-A3* of the male speaker at position 2, where on average, the model deviates from the measured data by 14.3 ± 4.7 dB. The modeled vowels also had the most difficulty predicting the proper relative formant values for the vowel /e^r/, most likely because the area functions for this vowel were based on vocal tract measurements made on a speaker not used in this study. Despite these shortcomings, the data in Table 5.3 (which provides the mean absolute differences and the corresponding standard deviations as well as the differences for each vowel at each position) demonstrate that on average, the vocal tract models can predict the measured relative formant amplitudes within 5 dB.

The models also agree with the recorded data in that the zeros in the vocal tract transfer function change as the position of the source is moved. Moreover, in agreement with the recorded spectra, the frequency shifts of the zeros in the models are dependent on the vowel, causing the zeros at one source location to noticeably differ in frequency. Nevertheless, despite the variability, the models also reveal that the zeros tend to cluster into multiple discrete frequency ranges. The frequencies of the zeros obtained from the models are shown in Figure 5.9 for positions 2 and 3. Unlike the recorded data, however, there were no zeros observed for the spectra at position 1 because position 1 was defined to be located at the glottal end of the vocal tract.

Table 5.3. Differences Between Predicted and Measured Relative Formant Amplitudes For Both Speakers.

Female				Male			
Position 1				Position 1			
Vowel	A1-A2	A2-A3	A1-A3	Vowel	A1-A2	A2-A3	A1-A3
/a/	-5.4	-0.3	-5.6	/a/	-6.0	-3.1	-9.2
/æ/	-6.3	5.2	-1.2	/æ/	-5.2	1.2	-3.9
/ɔ/	-6.5	4.7	-1.8	/ɔ/	-5.4	3.0	-2.4
/ɛ/	-6.1	-0.1	-6.2	/ ɛ /	-7.0	-6.4	-13.4
/e^r/	-5.0	3.6	-1.3	/e^r/	-0.4	-6.3	-6.7
/i/	-9.8	-5.6	-15.3	/i/	-4.3	4.9	0.6
/I/	-10.6	-3.8	-14.4	/I/	-7.9	0.4	-7.6
/u/	-16.8	-3.6	-20.4	/u/	-10.5	-4.7	-15.3
/^/	-4.5	-2.0	-6.5	/^/	-3.2	-1.8	-5.0
/U/	-4.8	-3.2	-8.0	/U/	-8.2	3.1	-5.2
mean	-7.6	-0.5	-8.1	mean	-5.8	-1.0	-6.8
std dev	3.6	3.6	6.2	std dev	2.7	3.9	4.6
Position 2				Position 2			
Vowel	A1-A2	A2-A3	A1-A3	Vowel	A1-A2	A2-A3	A1-A3
/a/	-3.7	-2.5	-6.2	/a/	-4.0	-4.4	-8.4
/æ/	1.4	3.7	5.1	/æ/	-4.7	-7.4	-12.1
/ɔ/	1.8	2.7	4.5	/ɔ/	-4.3	-10.2	-14.5
/ ɛ /	-0.4	-1.5	-1.9	/ ɛ /	-12.0	-2.4	-14.4
/e^r/	13.0	-9.0	4.0	/e^r/	-7.2	-11.9	-19.1
/i/	-12.0	7.6	-4.4	/i/	-3.8	-5.9	-9.7
/I/	-12.3	-4.9	-17.2	/I/	-1.0	-19.9	-20.9
/u/	-4.3	-17.4	-21.7	/u/	-15.1	-4.9	-20.0
/^/	1.6	-13.0	-11.4	/^/	-14.0	-3.2	-17.3
/U/	1.7	-6.6	-4.9	/U/	-6.5	-0.5	-7.0
mean	-1.3	-4.1	-5.4	mean	-7.3	-7.1	-14.3
std dev	7.0	7.3	8.7	std dev	4.6	5.4	4.7
Position 3				Position 3			
Vowel	A1-A2	A2-A3	A1-A3	Vowel	A1-A2	A2-A3	A1-A3
/a/	-6.6	0.1	-6.5	/a/	1.1	-7.4	-6.3
/æ/	2.4	-2.3	0.1	/æ/	6.3	-6.6	-0.3
/ɔ/	-0.6	8.7	8.0	/ɔ/	3.7	-7.6	-3.8
/ ɛ /	1.7	1.3	3.0	/ ɛ /	3.8	-10.3	-6.5
/e^r/	-22.4	10.4	-11.9	/e^r/	-15.4	13.3	-2.1
/i/	0.3	0.1	0.4	/i/	4.7	-1.0	3.7
/I/	9.2	-10.2	-1.0	/I/	2.1	-7.7	-5.6
/u/	1.6	-13.0	-11.4	/u/	-11.5	6.8	-4.7
/^/	-3.7	-2.5	-6.2	/^/	-5.6	-5.5	-11.1
/U/	-10.5	-3.3	-13.8	/U/	-1.1	-5.4	-6.6
mean	-2.9	-1.1	-3.9	mean	-1.2	-3.1	-4.3
std dev	8.3	6.9	6.8	std dev	6.9	7.1	3.8



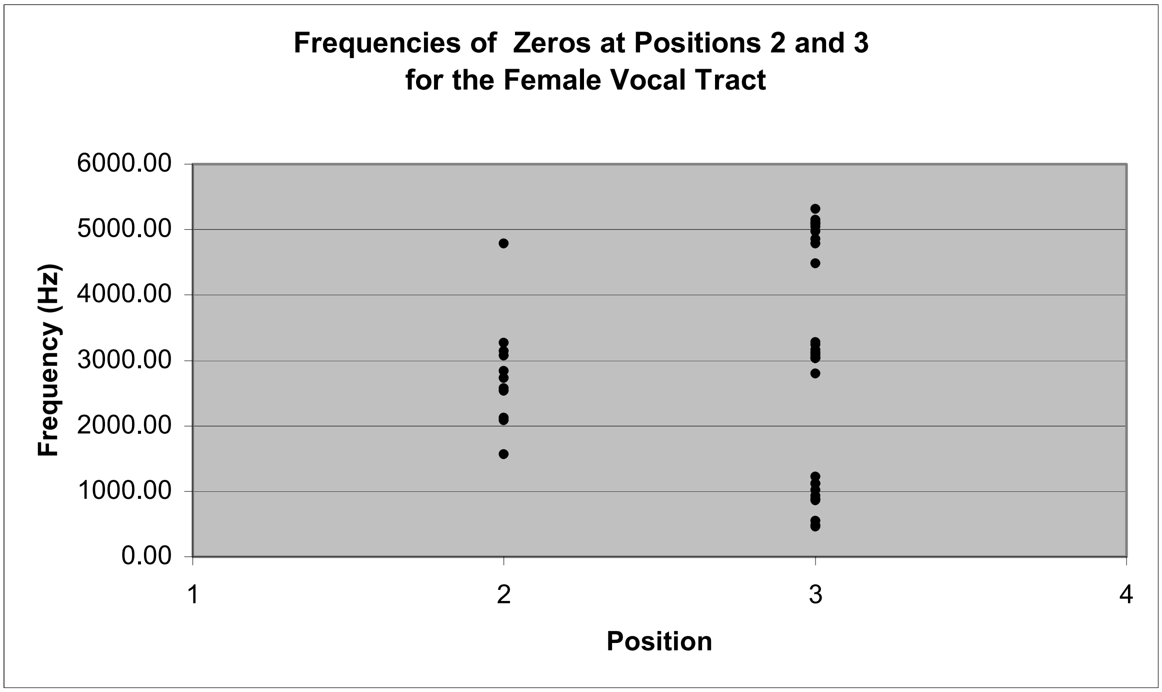


Figure 5.9. The zero frequencies at positions 2 and 3 for the positions for the models of the male (*top*) and female (*bottom*) vocal tracts.

It was not unexpected to find additional clusters of zeros in the model output than found in the recorded data. While the Door was effective at attenuating the EL self-noise (less so for the female speaker) it was not perfect, thus allowing some EL self-noise to leak into the recording environment, and mask any zeros that may have been present. The details of the Door's effectiveness are discussed in Appendix B. Moreover, at higher frequencies, the formant structure in the recorded data was not clear, making it difficult to discern the presence of a zero.

5.5. Discussion

This study was motivated by the observation of zeros in several of EL speech spectra obtained from the VA Database described in Chapter 4. Because it appeared that these zeros were altering the formant amplitudes, and in some cases actually canceling out formants, it was believed that these zeros were contributing to the unnatural quality of EL speech. Based on the basic theory discussed in Section 5.2, it was hypothesized that these zeros were the result of placing the EL sound source at a location other than at the "glottal" end of the vocal tract. The analysis of the recorded speech data and the vocal tract modeling effort were conducted in order to test this hypothesis.

The results obtained from both parts of this study confirm that positioning the EL sound source at a location other than at the terminal end of the vocal tract introduces zeros into the vocal tract transfer function and that these zeros affect the formant amplitudes in a manner than can potentially degrade EL speech quality. The data also demonstrate that while Eq. 5.2 correctly predicts that the frequencies of the zeros will decrease as the EL is moved further away from the terminal end of the vocal tract, they also show that Eq. 5.2 only suffices as a first order approximation. As both Figures 5.5 and 5.8 demonstrate, at any single source location, the spectral zeros greatly depend on the vowel that is being produced. Given that the ultimately the results of this study will be used to guide an enhancement effort to improve the state of EL speech, this vowel dependence is vitally important. Myrick and Yantorno (1993) proposed designing an all pole inverse filter based on the zeros measured in the spectrum of one vowel to compensate for the effects of the zeros. This filter would be effective for the vowel upon which the measurements were made (and depending on the vowel that was used, perhaps a couple of other vowels with similar vocal tract transfer functions), but may in fact be disastrous for other vowels. To explain why, it is helpful to consider Figure 5.10, which plots the first three zeros at Position 3 for the model of the female vocal tract as well as for a uniform tube. The lowest frequency zero is the one most likely to affect the perception of the vowel because of its proximity to both the first and second formants. However, trying to counteract the effects of this zero with a single filter based on measurements made on one vowel will be problematic. Because this zero frequency fluctuates between 460 Hz and 1220 Hz, when one vowel is filtered with a filter designed to compensate for the zero in another vowel, the output will contain an extra non-formant pole in the spectrum.

Different EL users place the EL device at different locations, ranging from the lower neck to just under the chin. As such, it is worth extending this analysis to other potential EL source locations. Figure 5.11 shows the range of the frequencies of the lowest frequency zero at each source location along the length of both the male and female vocal tract models. In general, once the source is moved beyond the first segment, the frequency range decreases as the source is moved further from the end of the vocal tract. However, this decrease isn't monotonic, especially in the male case.

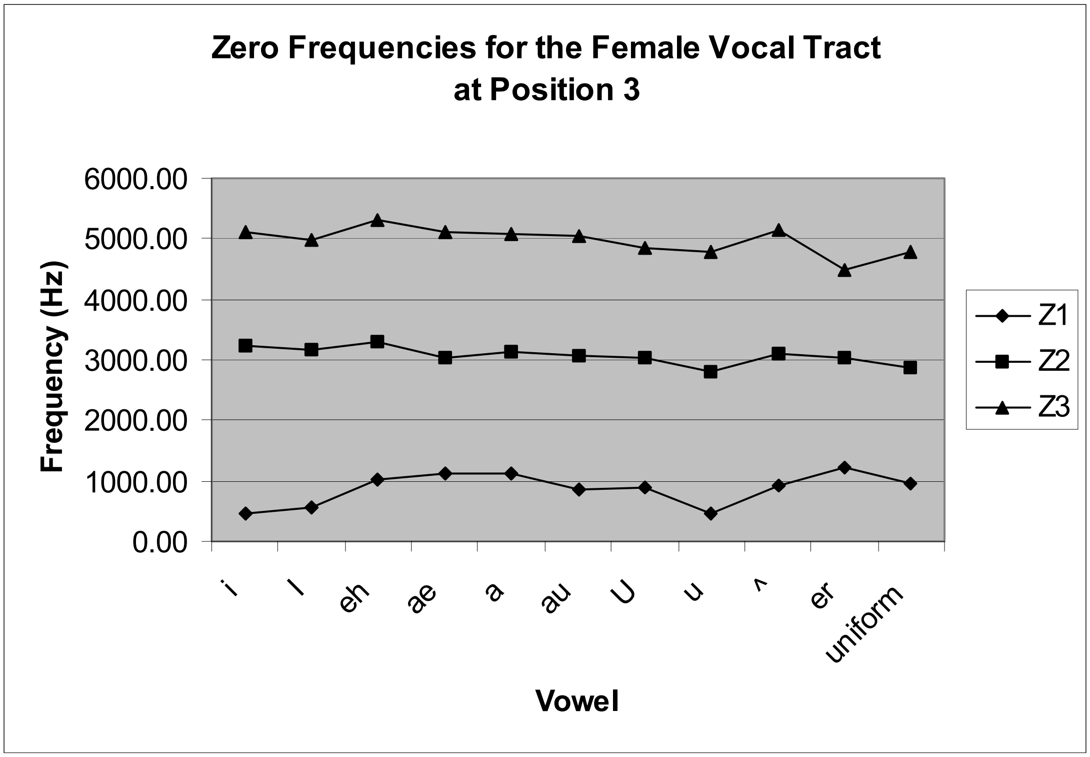
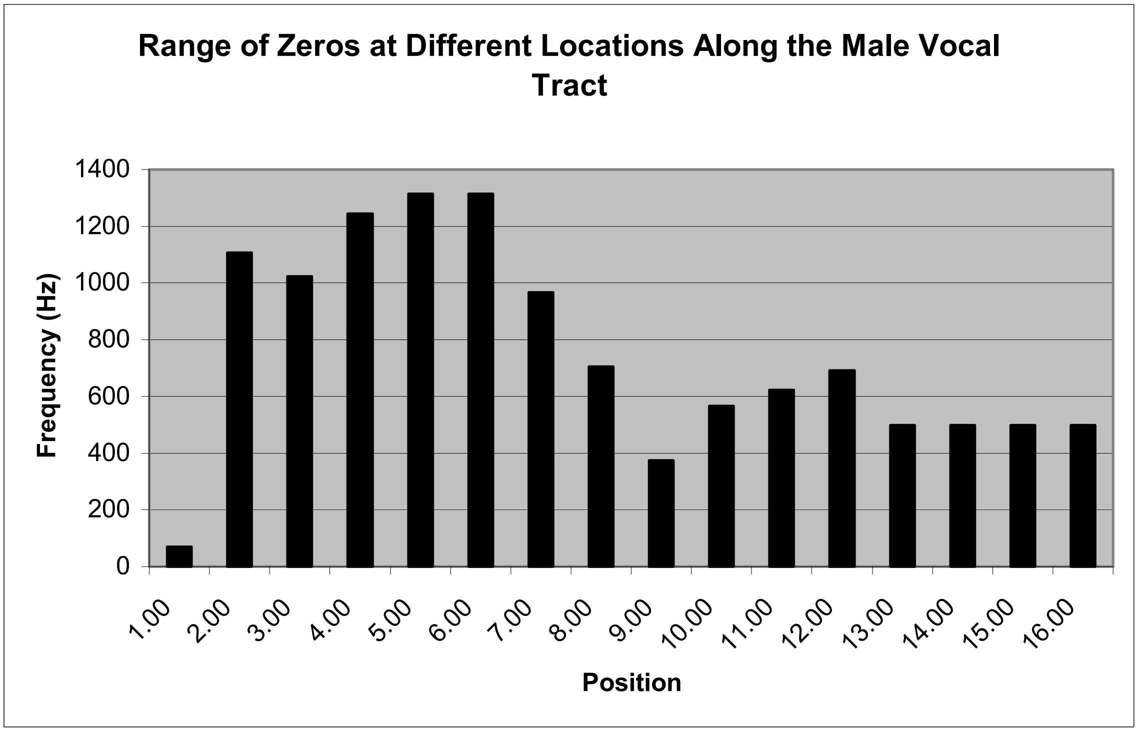


Figure 5.10. The frequencies of the zeros for each vowel produced at Position 3 of the female vocal tract model. The corresponding zero frequencies computed for a uniform tube length are presented for comparison.

The similarity of the zero locations when the source is only located at 1 segment away from the end is not surprising since the back cavity at this point resembles a uniform tube for all of the vocal tract configurations. However, moving the source just one segment vastly increases this spread of zero frequencies, especially in the female model. The cause of this variability can be attributed to differences in cross-sectional areas between back cavities of the different vowels. The zeros occur at what essentially are the natural frequencies of the back cavity, and, as section 5.2 demonstrates, for a uniform tube, these are the quarter-wave frequencies. However, as the cross-sectional areas in Tables 14.1 and 14.2 (see Appendix F) show, the back cavities that are formed when the source is placed closer towards the open end of the vocal tract significantly deviate from a uniform tube. Perturbation theory is useful for explaining the effect of these deviations in cross-sectional area. If a short uniform tube is open at both ends, then the impedance at the one of the openings can be approximated as



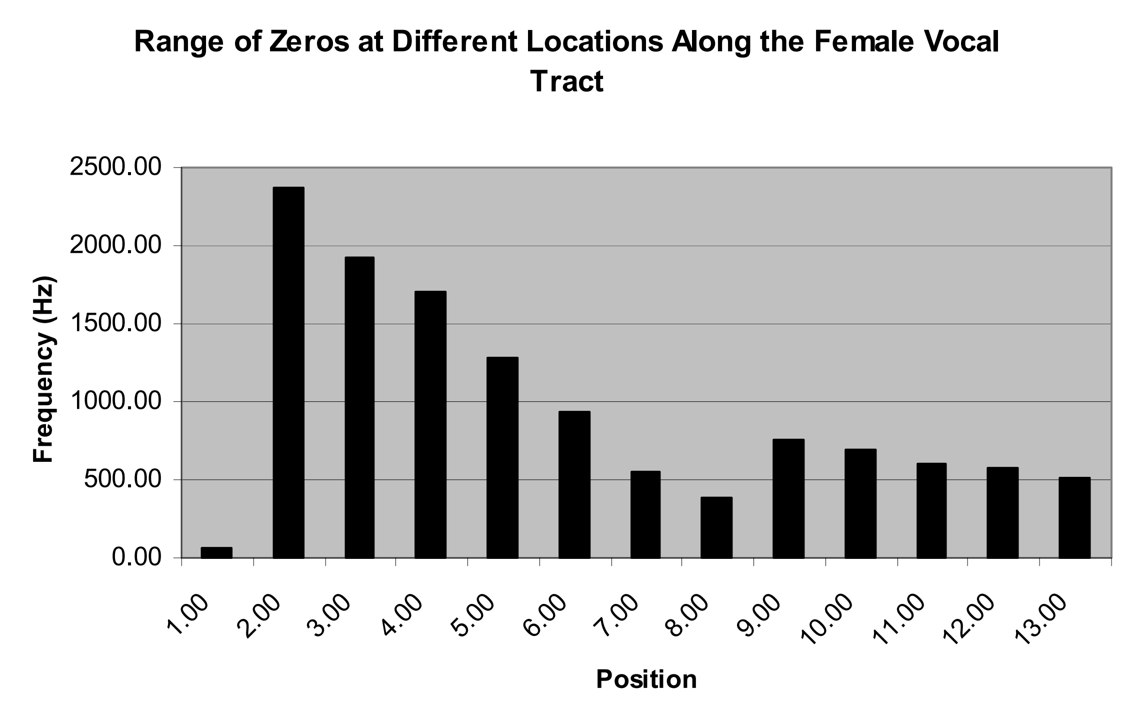


Figure 5.11. The variability of the zero frequencies at each source position along the length of the vocal tract for both speakers. In general as the source moved further away from the terminal end of the vocal tract, the variability decreased.

$$Z_l \cong j2\pi f \frac{\rho l}{A} \tag{5.9}$$

where ρ is the density of air, and l and A are the length and cross-sectional area of the tube respectively. Since Z_l is proportional to $\frac{l}{A}$, any increase in A is equivalent to a decrease in l. If the back cavities of the vocal tracts are considered to be comprised of concatenated small tubes (2 tubes in this case), then increasing the area of the front tube effectively shortens the length of the back cavity, increasing its natural frequencies. Conversely, decreasing the area of the front tube effectively lengthens the cavity. Thus, the source locations with the largest variability in zero frequency will also generally be associated with the largest variability of cross-sectional areas (across vowels) at the open end of the back cavity.

Perturbation theory also clarifies why the zeros of the high vowels, /i/, /u/, and to a lesser extent, /I/, are significantly lower than those of the other vowels in Figure 5.10. Referring again to the cross-sectional areas contained in Table 14.2 in Appendix D, at segment 9, the vocal tract is much narrower for these three vowels than it is for the others. The area of the end segment of the back cavity has been reduced, effectively lengthening the tube and hence decreasing the natural frequencies. The higher frequency zeros are not as affected by the constriction because at these small wavelengths, the relative length of the constriction is much larger than it was for the lower frequency zeros. Constrictions at different points along the vocal tract also account for why the range of zero frequencies does not decrease monotonically as the distance from the glottis increases.

Fortunately, the extent of the zero frequencies tends to decrease as the source is moved closer towards the lips. The zeros that are produced when the source is placed near the terminal (closed) end of the vocal tract are at frequencies that are high enough to be perceptibly insignificant. For these situations, it is probably not worth developing an enhancement scheme that could account for the resulting wide range of zero frequencies. At locations more distant from the glottis, however, the perceptual effects of the zeros can be far more detrimental. Consider the situation in the male vocal tract when the source in this case is located at the segment 7 (i.e. about 7 cm from the glottis), a distance that is somewhat less than halfway point of the vocal tract. Placing the source at this position is not uncommon among laryngectomy patients. Figure 5.12 displays the resulting lowest frequency zeros as well as the corresponding formant frequencies for each vowel. In this configuration, the zeros are located in the neighborhood of the second formants. For some vowels, such as /u/, the distance between the zero and the second formant is large enough not to have too much of an effect on the perception of the vowel. For the rest of the vowels, however, the zeros are much closer to the second formants, sometimes as little as 80 Hz away. The close proximity of the zeros to the second formants results in their attenuation, which will affect the perception of the vowels both in terms of quality and intelligibility. The attenuation of the second formant reduces its prominence in the spectrum to a point where the third formant effectively acts the second formant. Because the difference between the first and *effective* second formant frequencies has now increased to values similar to that of the vowel /i/, vowels such as /a/ and /ε/ sound more like /i/. This vowel confusion could contribute to the reduced vowel intelligibility reported by Weiss *et al.* (1979).

In this situation, developing a means of counteracting the effect of the zero would appear to be helpful in improving the quality (and intelligibility) of EL speech. Although the zero is not stationary, the fact that the zero appears to roughly parallel the trajectory of the second formant may simplify any enhancement algorithm.

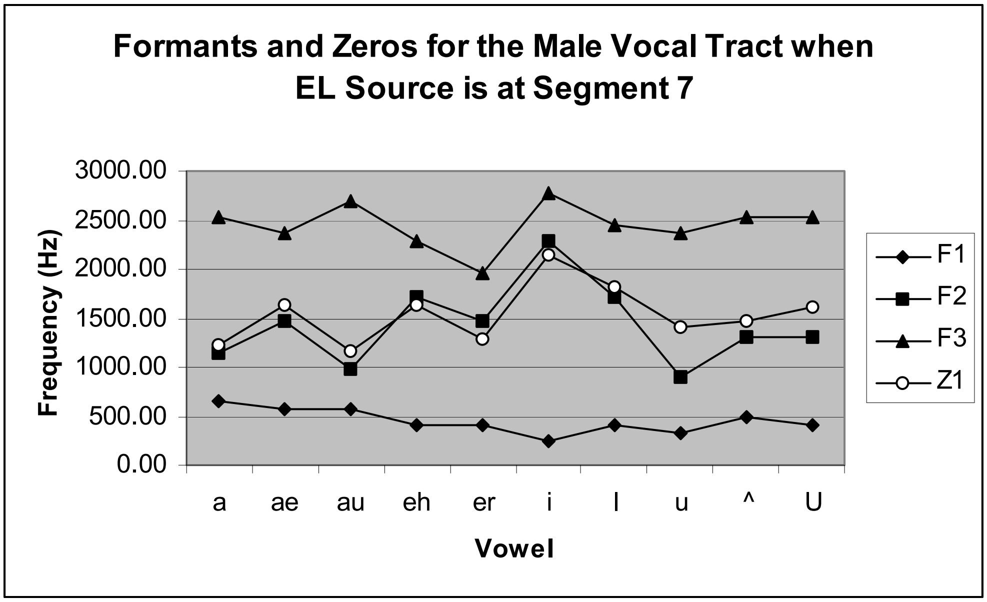


Figure 5.12. The formant and zero frequencies when the source is located at segment 7 in the model of the male vocal tract. The zeros resulting from this configuration appear to track the second formant, potentially having an adverse affect on both quality and intelligibility.

5.5.1. Limitations

Although the model was able to capture the effect of source location on the spectra of EL speech, as demonstrated by Table 5.2, there were several instances where it noticeably differed from the recorded speech data. One possible reason for this inaccuracy could be in the neck frequency response function used to generate the estimated EL excitation source. Digital filters that approximated the mean neck frequency response function were used to filter the measured acceleration. However, as reported in Meltzner *et al.* (2003) there was a notable degree of deviation from the mean in certain subjects. Because the male speaker in this study was a subject in the neck measurement study, it was known that the digital filter closely approximated his neck frequency response. However, no such measurement was made on the female speaker, so it is possible that her neck frequency response deviated from the approximation that was used.

Another difficulty arose in estimating the proper bandwidths to use in the model. The model bandwidths for the first three formants were taken from the bandwidths of the recorded speech data. However, it was often difficult to measure the higher formant bandwidths and so some approximations were made to best capture the higher frequency behavior. It's feasible that the bandwidths used for these higher formants were too large in certain cases thus overstating the spectral tilt.

Dividing the vocal tract models into approximately 1 cm. long segments sometimes made it difficult to choose a proper source location to compare with the recorded data. In some cases, the location of the zeros in the recorded data appeared to correspond with a source location between two adjacent model segments. The model was thus limited in its ability to reproduce the effect of placing the EL at its exact location on the neck. Using a model comprised of a larger number of shorter segments would help alleviate this problem.

Finally, the model assumed that the EL excitation source could be represented by a point volume velocity source. However, the plastic cap of the Servox EL that couples to the neck has a non-zero length (of about 2.5 cm) meaning that the EL excitation source might be better modeled as a distributed source. Being only a function of vocal tract shape, the formants would not be affected by the source type. However, this is not the case with the zeros. Using a source that excited the vocal tract at multiple adjacent locations along the vocal tract will produce multiple zeros in the output that are very close in frequency, effectively producing a single zero with a widened bandwidth. A wider zero would reduce the amount of attenuation it introduces into the nearby formants. To demonstrate this point, let us assume that the EL excitation source, *e*(*t*), can be written as the sum of two sources $e_1(t)$ and $e_2(t)$ that are located 1 cm. apart and are in phase with each other, i.e.

$$e(t) = \frac{1}{2}(e_1(t) + e_2(t)) \tag{5.10}$$

if $s_1(t)$ and $s_2(t)$ are the respective speech outputs for $e_1(t)$ and $e_2(t)$, then total speech output at the lips is

$$s(t) = \frac{1}{2}(s_1(t) + s_2(t)) \tag{5.11}$$

i.e., the total output is average of the individual outputs of the two sources. Figure 5.13 illustrates the result of this averaging for the model of the male vocal tract for the vowel /æ/. The top plot shows the spectrum of the output when the source is located at position 5, while the middle plot displays the spectrum when the source is located at position 6. The shifting of the zeros and the resulting changes in formant amplitudes are visible as the EL source is moved. However, when the two sources are combined, only the low frequency zero is discernable in the spectrum of the resulting output. Because

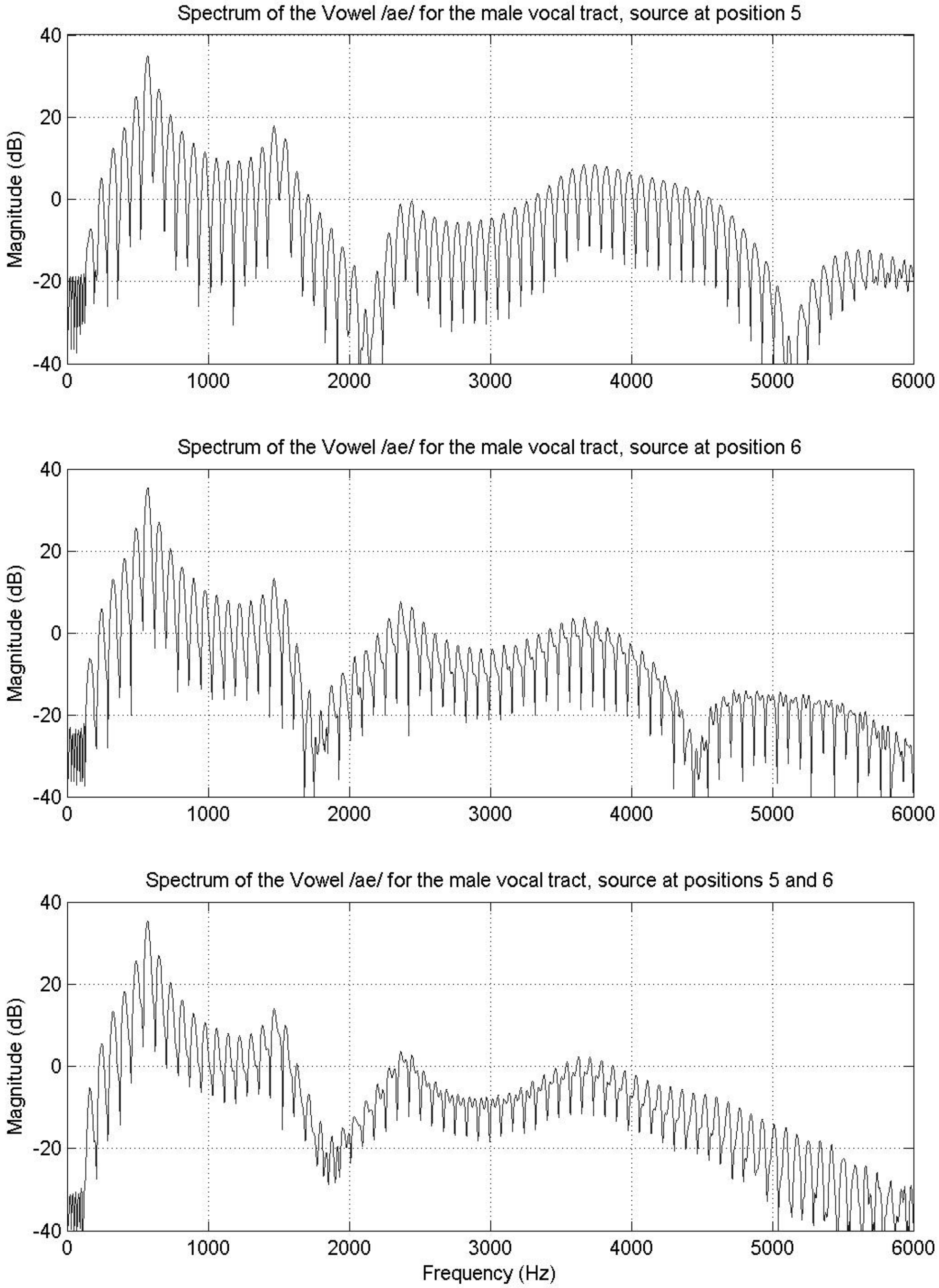


Figure 5.13. The spectra of the modeled vowel /æ/ for the male vocal tract when the source is placed at position 5 (*top*), position 6 (*middle*) and combined at both positions (*bottom*).

the higher frequency zeros of the respective component outputs are at such different frequency locations (5137 Hz and 4425 Hz), adding the two signals together fills in the zeros of the resulting spectrum. Conversely, the lower frequency zeros are only separated by 330 Hz, and are too close together to be completely filled in. Nevertheless, the bandwidth of the zero has increased, producing a formant attenuation that is the average of the formant attenuation found for the individual sources. This averaging effect which occurs for the distributed source could help explain why the point source model overstates the attenuation of the resulting zeros.

5.6. Summary

The purpose of this investigation was to explore the effect of EL source location on the acoustic properties of EL speech using a combination of acoustic analysis and vocal tract modeling. The results indicate that the placement of the EL at a location other than at then terminal end of the vocal tract introduces zeros into the vocal transfer function, whose frequencies are dependent on both the source location and the vowel being produced. Source locations further away from the end of the vocal tract are likely to have a greater impact on EL speech quality as they produce zeros that are located near formant frequencies, and in certain cases significantly reduce the formants' spectral prominence. As such, compensating for the effect of these zeros could potentially improve the quality of EL speech. However, because of the variability of the zero locations, implementing such a correction may prove to be difficult.

6. Discussion

6.1. Enhancement recommendations

Using three different approaches, this study investigated the properties of EL speech that contribute to its unnatural quality with the goal of using the results of the investigation to guide future efforts at improving the quality of EL speech. Based on the results discussed thus far some recommendations can be proposed.

The perceptual study of Chapter 3 explored the relative importance of three established deficits in EL speech: lack of pitch information, a low frequency energy deficit, and corruption by EL self-noise. From the results of this study, it is clear that of these three deficiencies, correcting for the lack of pitch information would provide the most benefit. Pitch information not only adds informational content, but also conveys the emotional state of the speaker, making the speech sound more natural. Consequently, it would seem prudent to devote a significant amount of effort to developing a pitch enhancement scheme. Unfortunately, adding the proper pitch content is probably the most challenging enhancement to implement. Pitch enhancement is a uniquely difficult problem because, unlike other aberrant properties, pitch information is essentially dependent on the thoughts of the speaker and thus no *a priori* information about pitch is available either from the raw EL speech itself or from any normative data. Therefore, it would seem necessary to use signals other than speech to try and estimate the proper pitch.

As has been previously mentioned, Uemi *et al*. (1994) designed a device that used air pressure measurements obtained from a resistive component placed over the stoma to control the fundamental frequency of an EL. Unfortunately, only 2 of the 16 study subjects studied were able to master the control device. Another, more recent effort by Goldstein (2003) used EMG signals measured from specially innervated strap muscles in a processing scheme to provide pitch control. During laryngectomy surgery, the recurrent laryngeal nerve was severed from the laryngeal muscles and sutured into one of the strap muscles which supported the larynx prior to its removal. After waiting the several months required for muscle re-innervation, the EL users who underwent this operation were trained to control the onset/offset and the pitch of an EL device which was connected to an EMG processing device. Of the three laryngectomy subjects that underwent this process, two of them were able to adequately control the pitch of the device. While this system holds promise for future EL users, those current users who have not undergone the re-innervation surgery may not receive much benefit. For this group of users, a device that incorporates a fixed pitch contour may be a useful compromise. Although a fixed contour will not provide any additional information to the speech and may in fact lead listeners to confuse the intent of the speaker (consider for example, the case of asking a question with declarative prosody), it may aid in reducing the unnatural quality of EL speech. The most recent version of the Ultra Voice Plus (www.ultravoice.com) uses such a fixed pitch contour.

As these studies demonstrate, it is not only difficult to provide EL users an effective means of pitch control, but the set of EL users that would benefit may be quite limited. However, the fact that monotonous normal speech was consistently found to approximate the quality of normal natural speech better than pitch enhanced version of EL speech indicates that a great deal of improvement can be achieved without incorporating pitch information into EL speech. Theoretically compensating for the factors that differentiate EL speech from normal monotonous speech would produce speech whose quality approached that of normal natural speech. The results in this document identify at least some of those factors.

The results of the perceptual experiments also demonstrated that, although not as effective as injecting pitch information, removing the self-generated EL noise improved the quality of EL speech. As such, incorporating a noise-reduction scheme into a future EL enhancement system seems prudent. Of the noise reduction methods discussed in Chapter 2, the most effective appears to be the adaptive filtering algorithm suggested by Espy-Wilson *et al.* (1998). However, this algorithm requires the placement of a second microphone at the position of the electrolarynx, something that many EL users may not tolerate (Hillman 1999). Hence it would be valuable to develop a noise reduction scheme that requires less equipment. Simply adding a time-aligned estimate of the direct noise to noise-reduced EL speech served as an acceptable substitute for raw EL, hinting that there may be a simple additive relationship between the self-noise and EL speech. A future enhancement algorithm may be able to exploit this apparent simple additive relationship without resorting to the use of additional equipment. For example, a finite sample of the direct noise could be recorded, stored in memory, and then used to remove the direct noise in a frame by frame manner.

The effectiveness of the low frequency enhancement was mixed. On its own and coupled with the noise reduction enhancement, it produced a limited improvement in speech quality. However, in certain circumstances, it appeared to reduce the effectiveness of the other two enhancements. Nevertheless, it is unlikely that such a substantial dearth of low frequency energy would not contribute to the abnormal quality of EL speech. Thus, it is feasible that the Qi & Weinberg (1991) algorithm used in the perceptual studies was not the most ideal solution to the low frequency deficit. Based on anecdotal evidence, EL speech enhanced using the Qi & Weinberg method tended to have a muffled, unclear quality that seemed to reduce the intelligibility. The non-linear phase of the low pass filter used in this enhancement (See Figure 3.1) may be smearing the resulting speech making it more difficult to understand. Moreover, when designing this filter, Qi & Weinberg minimized the difference between normal and EL speech only for frequencies below 550 Hz. Thus the filter design ignored the situation at higher frequencies, possibly resulting in too much high frequency attenuation. Speech that contains too little high frequency energy often sounds muffled. Consequently, a more effective method to adjust the low frequency content of EL speech is required. Despite the minimal gain in quality, it is worth developing a better low frequency enhancement, as this should be a relatively easy enhancement to enact.

Although it was not directly studied, the perceptual experiments demonstrated that the lack of voice/voiceless information inherent in EL speech is also partially responsible for its unnatural quality. The enhanced versions of the sentences that were comprised of both voiced and voiceless phonemes regularly received significantly worse ratings than their corresponding all voiced counterparts. This suggests that listeners used the lack of voice/voiceless information as another cue that distinguished the enhanced EL speech sentences from the natural normal sentence. Obviously, developing a method to insert this information into EL speech would improve its quality, but currently, the prospects of doing so are remote. In some respects, this enhancement is even more difficult to actualize than the pitch enhancement, as it would require an even finer degree of control. Therefore at this time, the costs of developing a means to correct this deficiency would seem to outweigh the benefits it would entail.

The analysis of the corpus of pre- and post-laryngectomy speech revealed three potential aberrant EL speech properties that could contribute to its unnatural quality: higher formant frequencies, narrower formant bandwidths, and spectral zeros. Unlike the latter two properties, the increase of all the formant frequencies may not make EL speech sound less natural per se; there was no formant shift in the EL sentences used in the perceptual experiments, which were all found to be unnatural. However, the formant shifts do make an EL user's speech deviate from his/her pre-laryngectomy speech and if the ultimate goal of any enhancement effort is to return an EL user's post surgical voice to his/her pre-surgical state, then reversing this change is worth pursuing. As will be discussed later, correcting for this problem should not engender too much difficulty.

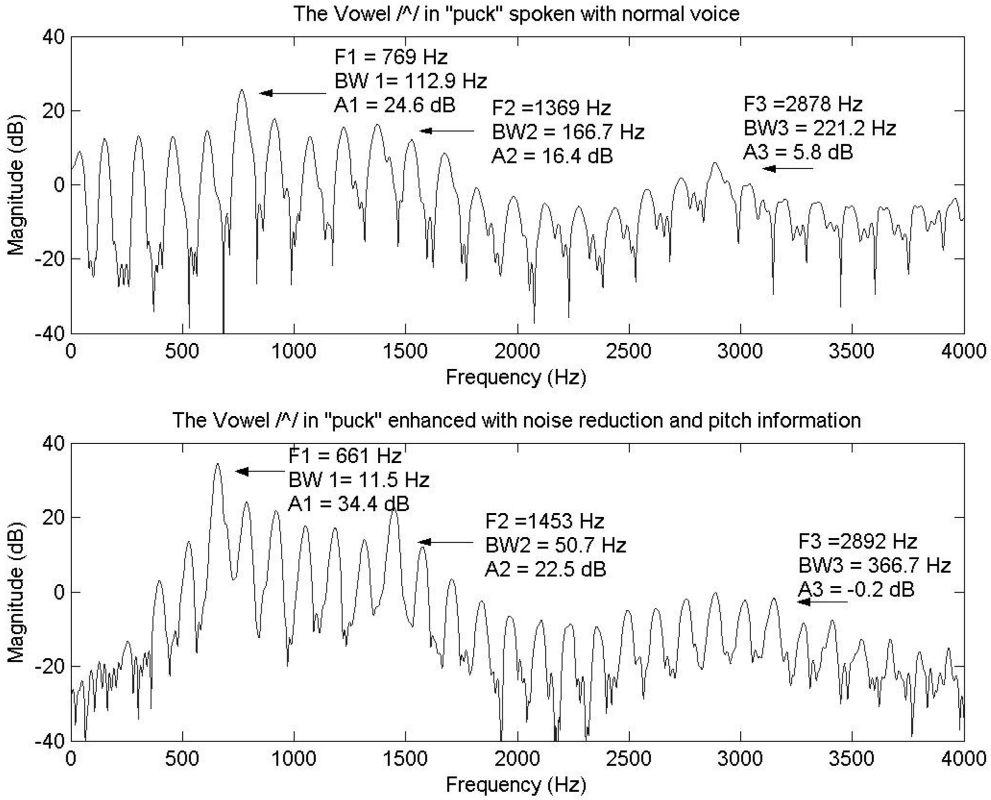


Figure 6.1. The spectra of the vowel /^/ in "puck" in both normal speech (*top*) and in the EL-NP version of the first sentence. The bandwidths of the enhanced EL speech are much narrower than those of normal speech.

The narrower bandwidths found in the speech of the VA database were also found in the EL sentences used in the perceptual experiments. Figure 6.1, which displays the spectra of the vowel /^/ in "puck" found in the EL-NP and normal version sentence 1 for the female speaker, clearly demonstrates this narrowing of the formant bandwidths. The perceptual effect of narrower formant bandwidths has not been well studied, but as the bandwidth of a resonance decreases, the more sinusoidal the resonance becomes. Thus one would expect that speech with narrower formant bandwidths would sound more tonal and perhaps harsher than normal speech. The speech of Subject 7 in the VA Database, which had exceptionally narrow formant bandwidths, was especially unpleasant to listen to. As such, restoring the formant bandwidths to their natural values would be an important goal for any future enhancement scheme to attain.

The effect of the spectral zeros on the quality of EL speech is highly dependent on the frequencies of the zeros and by extension, where the EL user places the device. The speakers in the perceptual experiment held the Servox near position 2, producing zeros that were located at frequencies between the second and third formants. Consequently the effect of these zeros was limited to altering the overall tilt of the spectrum and sometimes interfering with the third formant. Although the change in tilt helps to differentiate the EL speech from the normal speech, the range of spectral tilts for normal voices is quite large (Hanson 1997) so it is unclear whether this change in tilt contributed to the disordered quality of the EL speech. Furthermore, the vocal tract modeling demonstrates that zeros are not stationary and thus a complex algorithm would be required to correct for them. However, no zeros are found in the EL spectrum of Figure 6.1 because it has already been processed by the MELP vocoder. As is discussed in Appendix C, MELP uses linear prediction to model the vocal tract filter. A linear predictive filter is by definition an all pole filter and when forced to model a system that contains both poles and zeros, will use multiple poles to approximate the effects of the zero. As such, while an analysis of MELP processed speech will not reveal any spectral zeros, it will demonstrate that any alterations of the relative formant amplitudes have been maintained. Figure 6.2 displays the spectra of the vowel /^/ in "puck" from the raw EL speech sentence and again from the sentence enhanced with noise reduction and added pitch information. It is clear that the zero at 2406 Hz in the spectrum of the unprocessed speech is not visible in the enhanced speech. However, MELP does faithfully reproduce the relative formant amplitudes (within 2 dB) which are affected by the zero. This result suggests that in certain cases, compensating for the zeros may be most effectively accomplished by properly adjusting the formant frequencies and bandwidths in the encoded vocal tract transfer function to achieve the desired relative amplitudes.

However, many EL users place the EL further away from the terminal end of the vocal tract, resulting in zeros that are more likely to interfere with the formants. As such, simply adjusting the formant attributes would probably not sufficiently counteract the effect of the zeros, especially if a formant is almost completely attenuated. Yet, developing a more complex zero-compensation system in these situations would be of great benefit to the EL users because of the zeros would have an adverse effect on both the resulting speech quality and the intelligibility. Furthermore, although they were not the focus of this thesis, intra-oral EL devices such as the Cooper-Rand and the Ultra Voice would be well served by a zero-compensation algorithm since the source is placed only a few centimeters from the lips when these devices are used. In sum, any enhancement algorithm that corrects the effects of the EL source location must be tailored to the needs of specific EL users.

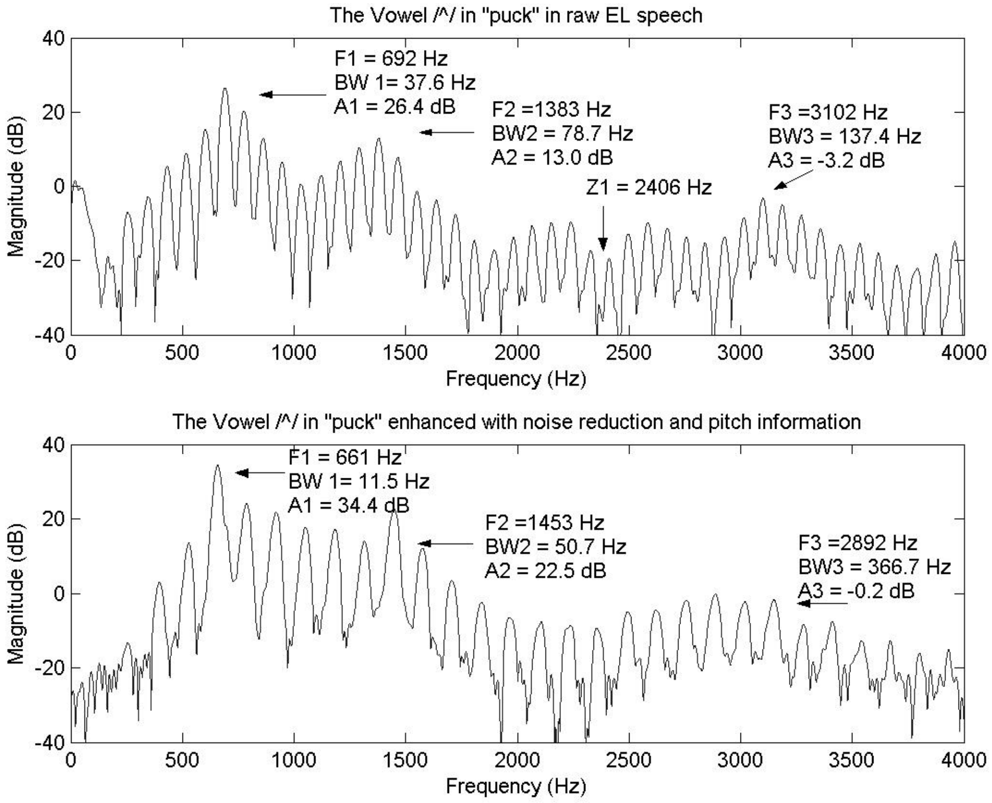


Figure 6.2. Spectra of the vowel /^/ in "puck" in raw EL speech (*top*) and enhanced EL speech (*bottom*). Because the enhanced speech has been processed using the MELP vocoder, it cannot reproduce the zero in the EL spectrum. However, it does faithfully reproduce the formant amplitudes which are affected by the zero. The harmonics are different between the two spectra because as detailed in Chapter 3, the pitch of the processed speech was altered to match that of the normal version of the sentence.

6.2. A Framework for Enhancement

The investigation discussed in this document was conducted with the aim of using the results to guide a future enhancement effort to improve the quality of EL speech. Specifically, it was envisioned that this enhancement effort would take the form of the enhancement component of the improved EL communication device being developed by the Voice Project group in the W.M. Keck Neural Prosthesis Research Center in Boston (Houston *et al.* 1999). In this configuration, the enhancement would be a post-processing scheme that would operate on EL speech recorded at the lips. In practice, the speech would have to first be analyzed, then altered, and finally resynthesized. This suggests that a vocoder would make for a perfect platform upon which the enhancement module can be based.

The MELP vocoder (see Appendix C), a version of which was used in to generate the sentences in the perceptual study, could serve in this capacity. Like all linear predictive vocoders, MELP separates the excitation source from the vocal tract filter. Linear predictive (LP) coefficients which are then converted to line spectral frequencies (LSFs) are used to represent the vocal tract filter. Both the LP and LSF representations are easily manipulable and would allow for the required alterations needed for EL speech enhancements. For example, increasing the formant bandwidths can be accomplished by multiplying the LP coefficients by r^n where r is a number less than 1, and n is the index of the coefficient. Shifting the formant frequencies downwards by a constant value can be implemented by multiplying the LSFs by a scaling factor that is less than unity. If a more complex modification is desired, different pairs of LSFs can be multiplied by differently weighted scaling factors. Finally, a zero correction algorithm can be performed on either the LP coefficients or the LSFs as needed.

The excitation source of the MELP vocoder also lends itself to modification. MELP encodes the source using its pitch period, the amplitude of its first 10 harmonics, the degree of voicing in 5 frequency bands, and a voiced/voiceless flag. By parameterizing the voicing source in this fashion, MELP is flexible enough to accurately synthesize different sorts of voiced speech (e.g. breathy speech) as well as unvoiced speech. Moreover, this flexibility lends itself to be used for modifying the source component of EL speech as was demonstrated by using MELP to implement the insertion of a natural pitch contour into EL speech. If the ability to give EL users an effective means of pitch control becomes available or if it is desired to give EL users a fixed but more natural sounding pitch contour, realizing the pitch change through the vocoder may be a better option than doing so through the EL source because it allows for a simpler source to be used. More likely, however, the excitation source can be modified to compensate for the shortcomings of EL source. For example, the amplitudes of the 10 harmonics can be adjusted to correct the low frequency energy deficit without affecting the phase of the speech. Additionally, low levels of noise can be added to the source to simulate a more breathy voice, or if the technology arises to give EL users voice/voiceless control, a voiceless excitation source (i.e. noise) can be appropriately synthesized.

The output of the enhancement module would then be transmitted either by an external speaker worn by the EL user or over the telephone. The enhanced speech would have to be greatly superior to raw EL speech to convince many EL users to wear extra equipment. However, the telephone is an excellent application for an EL enhancement algorithm since listeners on the other end will be unaware of any processing delay. Improving the state of EL telephone speech would be an important advance since EL users have great difficulty communicating over the telephone. Moreover, the increasing popularity of mobile phones heightens the need for better EL telephone speech.

Fortunately, today's digital mobile phones already have vocoders built into them, making them ideal platforms for an enhancement algorithm. For example, the 3G CDMA protocol employs a Selectable Mode Vocoder (SMV) (Greer and DeJaco, 2001) which makes use of a version of a code-excited linear predictive (CELP) vocoder called eX-CELP (Gao *et al.* 2001). Meanwhile, the current GSM protocol uses an Adaptive-Multirate (AMR) codec that uses a different CELP (algebraic code-excited linear predictive, ACELP) vocoder (Bessette *et al.* 2002). CELP vocoders are similar to the MELP vocoder in that they separate the speech into an excitation source and vocal tract filter. However, the manner in which CELP vocoders encode the excitation is quite different from MELP and may not be as flexible. Nevertheless, adapting one or more of these vocoders to include and EL speech enhancement algorithm appears to be a practical and viable method of improving the state of EL speech.

6.3. Future work

Although this study has identified and explored many properties of EL speech that contribute to its unnatural quality, there are other properties may also play a role in this matter. The EL voicing source consists of only a periodic excitation whereas normal speech contains a noisy component as well (Klatt and Klatt 1990). This deficiency could potential be corrected using the MELP vocoder described earlier, but the nature of the correction must be thoroughly explored. Moreover, the amplitude of a normal glottal excitation is modulated during normal speech while the amplitude of the EL excitation source is fixed at a constant value when activated. The perceptual effect of correcting for these shortcomings should be examined in the fashion described in Chapter 3. If applying the full combination of potential enhancements does not produce EL speech that has a quality similar to that of normal monotonous speech, then work should continue identifying and testing the effects of other abnormal EL speech properties.

Although intelligibility was not dealt with in this thesis, it is an important property that must be addressed; an enhancement that improves the quality of EL speech but also degrades its intelligibility may not be beneficial to an EL user. As work progresses on improving EL speech quality, the intelligibility of the improved EL speech should be examined as well.

While the investigation into other aberrant EL speech properties is being conducted, an enhancement algorithm based on the recommendations discussed in this document can be developed and tested. These would include implementing a new method of low frequency enhancement, adapting Espy Wilson *et al.*'s (1998) method of noise reduction or developing a simpler one, enacting a downward formant shift, widening the formant bandwidths, and designing an algorithm to correct for the presence of zeros in the vocal tract transfer function. Again, the procedure for the perceptual studies can be applied to testing these enhancement components as are they realized. Once a working prototype enhancement algorithm has been developed, it would need to be mated to a hardware platform, tested and revised accordingly.

7. Conclusion

The results of this work have advanced the state of knowledge about the deficiencies of electrolarynx speech. The perceptual experiments discussed in Chapter 3 determined that of the three best characterized aberrant EL speech properties, the lack of pitch information was most detrimental to the quality of EL speech. The other two properties that were studied, the presence of the EL self-noise and a low frequency energy deficit were also found to reduce the quality of EL speech but to a lesser degree. However, it was also found that normal-monotonous speech sounds more like normal natural speech than any form of enhanced EL speech. This implies that 1) EL speech quality can be vastly improved without the very difficult task of adding pitch information, and 2) there must be other properties of EL speech that are contributing to its unnatural quality.

The analysis of the VA corpus of pre- and post-laryngectomy speech sought to identify these other aberrant properties. The results indicated that the formant bandwidths of EL speech are narrower than those of normal speech. Moreover, several spectral zeros were observed to be altering the formant amplitudes and in some cases canceling out formants. It was believed that these zeros were a product of the location of the EL excitation source and because of the potentially negative effect on EL speech quality, the relationship between source location and zero frequencies were examined.

The vocal tract modeling and speech recording experiment demonstrated that placing the EL further away from the terminal end of the vocal tract decreased the frequencies of the resulting zeros. As such, the effect of these zeros will vary among EL users; for those that place the device near the end of the vocal tract, the impact will be minimal. However, for those that place the device closer to the chin (and for those using intra-oral ELs), the zeros will likely interfere with the formants, thereby reducing both speech quality and intelligibility.

Based on these findings, an enhancement algorithm that corrects for the low frequency deficit, the interference of the EL self-noise, the narrower formant bandwidths, and the effect of the source location, should produce EL speech whose quality surpasses what is currently available. Additionally, adding a fixed but more natural sounding pitch contour may be a useful compromise for EL users for whom effective pitch control cannot be provided. As an enhancement system is developed it will also be important to explore other abnormal properties of EL speech that may adversely affect its quality. The lack of breathiness and amplitude modulation of the currently used EL excitation source are two potential examples of such properties. If their effects are found to be significant and the ability to compensate for those effects is feasible, then the enhancement algorithm should be expanded to correct those deficiencies.

A future enhancement system was envisioned as a post-processing scheme that would operate on the speech recorded at the lips of an EL user. This formulation makes the enhancement of telephone speech an ideal application, especially since currently available digital mobile phones already contain the means to separate speech into its components. Because of the difficulties EL users experience when using the phone, improving EL telephone speech should improve their quality of life.

8. Acknowledgements

The following people are owed a great deal of thanks for helping me complete this work:

The members of my committee: Bob Hillman, Ken Stevens, Joe Perkell and Tom Quatieri for their support, guidance, and expertise.

My fellow researchers at the Voice and Speech Lab: Harold Cheyne, Ehab Goldstein, James Heaton, and Jim Kobler, for their ideas and assistance in running my experiments.

Glenn Bunting and Janice Eng for spending several hours being my human guinea pigs.

Brad Story for his assistance with the vocal tract modeling work.

Yingyong Qi for giving me the opportunity to learn signal processing techniques in a corporate environment.

My parents and my sister for willingly listen to me moan and groan when things didn’t always go as planned.

And most importantly, Jennifer Friedberg, for supporting me in this endeavor and for doing everything she could to help me finish this work.

9. Appendix A. Vowel Notation

Because of the limitations of the software used to generate some figures and tables, it was impossible to use the proper phonetic notation for certain vowels. Therefore, when needed, the following substitute notations were used:

1. /ae/ = /æ/ (in “bat.”)

2. /eh/ = /ɛ/ (in “bet.”)

3. /au/ = /ɔ/ (in “bought.”)

4. /U/ = /ʊ/ (in “put.”)

5. /er/ = /e^{r}/ (in “Bert.”)

6. /^/ = /ʌ/ (in “but.”)

10. Appendix B. Attenuating EL Self-Noise

A method to effectively reduce the presence of EL self-noise was needed in order to measure its perceptual effect on the quality of speech and to allow for the measurement of the effect of source location. Espy-Wilson *et al.* developed an adaptive filtering algorithm that was reasonably effective at attenuating the direct noise but at the cost of reducing the intelligibility of nasal consonants. However, because of the need to eliminate any perceptual cues (other than the ones being studied) that could be used to differentiate between the versions of the EL speech sentences, it was decided that another noise reduction method was needed. To meet this need, the Door was developed.

The Door was constructed by fastening three ½ inch thick boards of plywood that were sized to fit snugly into the doorway of the acoustic chamber of the Voice and Speech Lab. At about 4 feet from the bottom, a hole was cut into the door to allow for a form-fitting mask (Intertech non-conductive face mask, Smiths Industries Medical Systems) to be sealed into it. When in use, an EL user would place his or her face in the port in the door while keeping the EL outside the booth thus only allowing sound from the lips (and nose) to enter into the acoustic chamber. To further seal the acoustic chamber from the outside environment, 5 clamps were fitted on the inside of the door to pull it securely against the door frame of the chamber. Figure 10.1 shows three views of the Door.

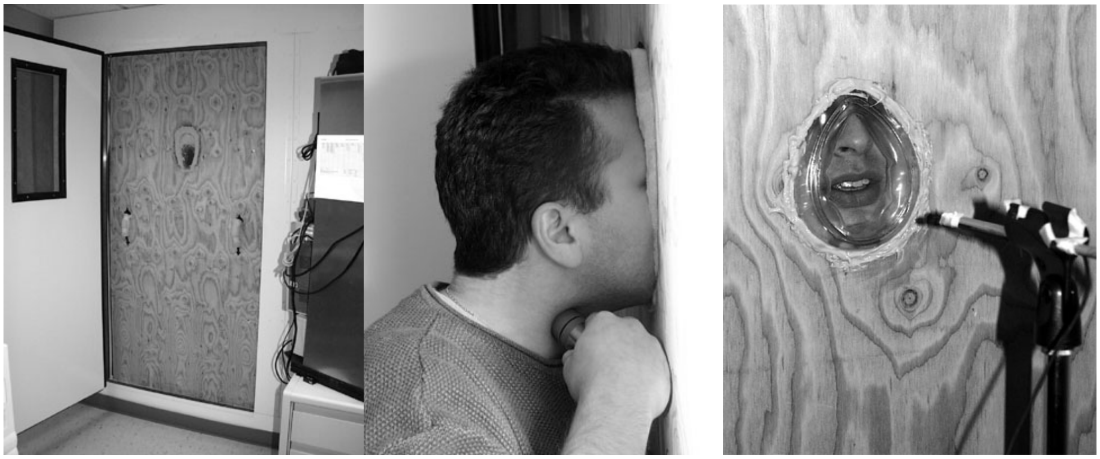


Figure 10.1. Three views of the door. *Left.* The Door sealed in place in the doorway of the acoustic chamber. *Middle.* A speaker using the door while speaking with an electrolarynx. Keeping the EL outside of the acoustic chamber reduces the amount of self-noise in the resulting speech. *Right.* A view from inside the acoustic chamber. The plastic mask seals around the speaker's nose and mouth while a microphone placed inside the booth records the speech.

10.1. Door effectiveness

To measure the effectiveness of the door, the two speakers used in the experiments described in Section 5 were asked to keep their mouths closed while activating the EL. This task was repeated at the 3 positions on the neck both with and without the door (i.e. inside the acoustic chamber). Figure 10.2 shows the long-term average spectra of the direct noise estimates at all 3 positions for the male speaker and Figure A.3 shows the same for the female speaker.

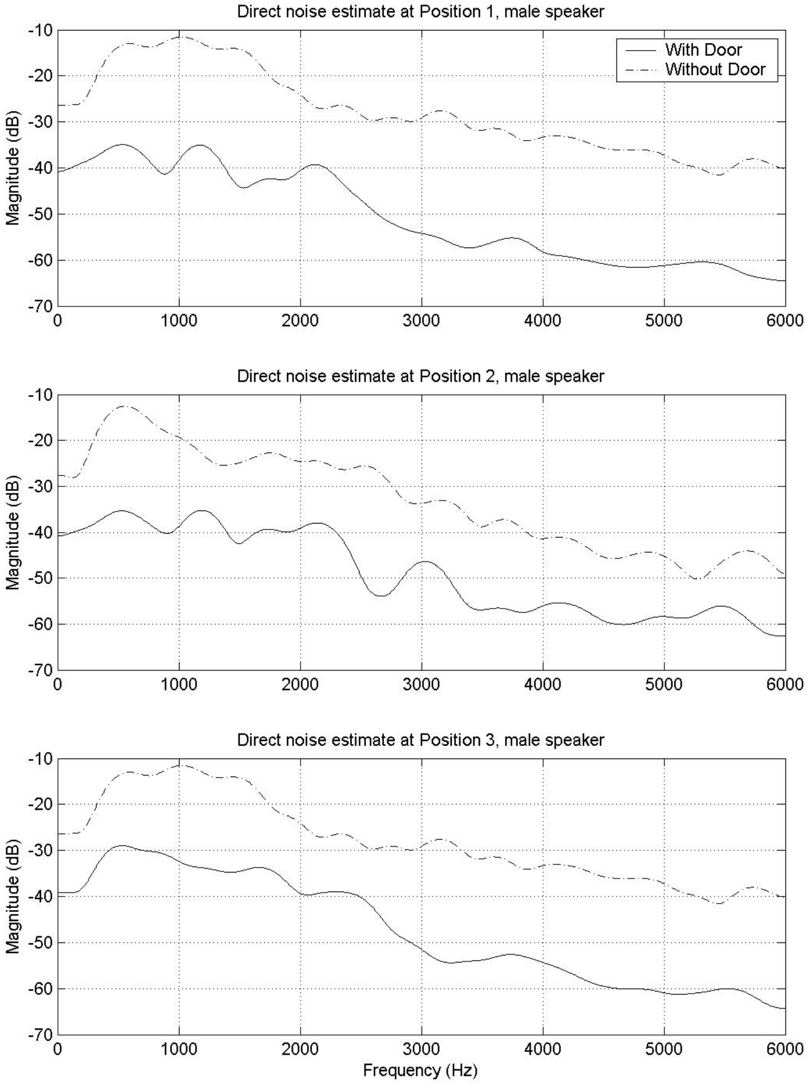


Figure 10.2. The long-term average spectra of the direct noise estimates at all 3 positions for the male speaker. These spectra demonstrate that the Door is very effective at reducing the amount of self-noise in EL speech for the male speaker. On average, the Door reduced the amount of self –noise by 45.4, 32.4, and 38.8 dB for positions 1, 2, and 3 respectively.

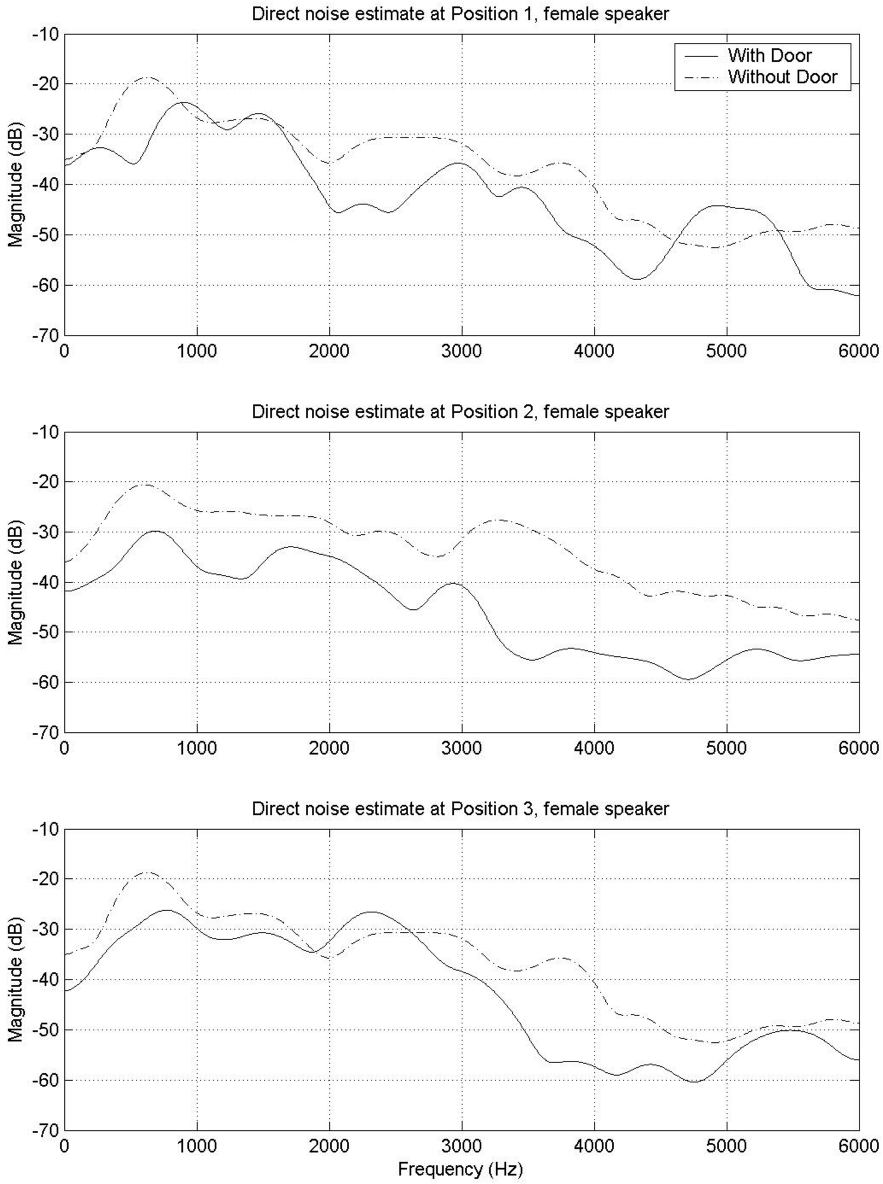


Figure 10.3. The long-term average spectra of the direct noise estimates at all 3 positions for the female speaker. Although in general, the Door did attenuate the EL self-noise, it was not as effective for the female speaker as it was for the male. On average, the Door reduced the amount of self –noise by 11.1, 32.4, and 24.4 dB for positions 1, 2, and 3 respectively.

Figures 10.2 and 10.3 both demonstrate that the Door reduced the amount of EL self-noise in the recording environment and that it was more effective for the male speaker than the female. Averaging over frequency, for the male speaker, the self-noise was reduced by 45.4, 32.4, and 38.8 dB at positions 1, 2, and 3 respectively while for the female, the reduction was 11.1, 32.4, and 24.4 dB for the same three positions.

10.2. Radiation from the door

Because the outputs of the models developed in Section 5 were evaluated against recorded data, it was important to know the radiation characteristic of the Door. The radiation characteristic is defined as the relationship between the volume velocity at the lips and the pressure measured at a distance from the lips. In a typical environment, for distances, *r*, which are greater than a few centimeters (in this case, *r* = 15 cm.), the open mouth can be considered a simple source radiating in all directions. In such cases, the radiation characteristic can be approximated as

$$R(f) = \frac{j2\pi f\rho}{4\pi r} \cdot e^{-j\frac{2\pi fr}{c}}, \tag{10.1}$$

where *c* is the speed of sound, and ρ is the density of air. Equation (10.1) is an accurate approximation (within a few decibels) for frequencies up to 4000 Hz. (Stevens 1998). The effect of the radiation characteristic can be approximated by differentiator. However, the situation of speaking through the Door is more akin to a cylinder in an infinite baffle than a simple source and thus it was essential to determine how the Door radiation character differed from the simple source approximation.

The following experiment was performed to estimate the radiation characteristic of the Door. A large funnel was shaped so that it fit securely into the port of the Door; the opening of the funnel was approximately the same size as the opening of the port. The funnel was attached to a loudspeaker placed into the port of the Door and driven with broadband noise with bandwidth of 25 kHz. Using the SYSID software package (Sysid Labs), the transfer function between the speaker driving signal and the pressure was computed. A similar measurement was made by placing the funnel-speaker combination inside the acoustic chamber (i.e. the free field) and measuring the same transfer function.

Photos of the funnel attached to the speaker are shown in Figure 10.4.

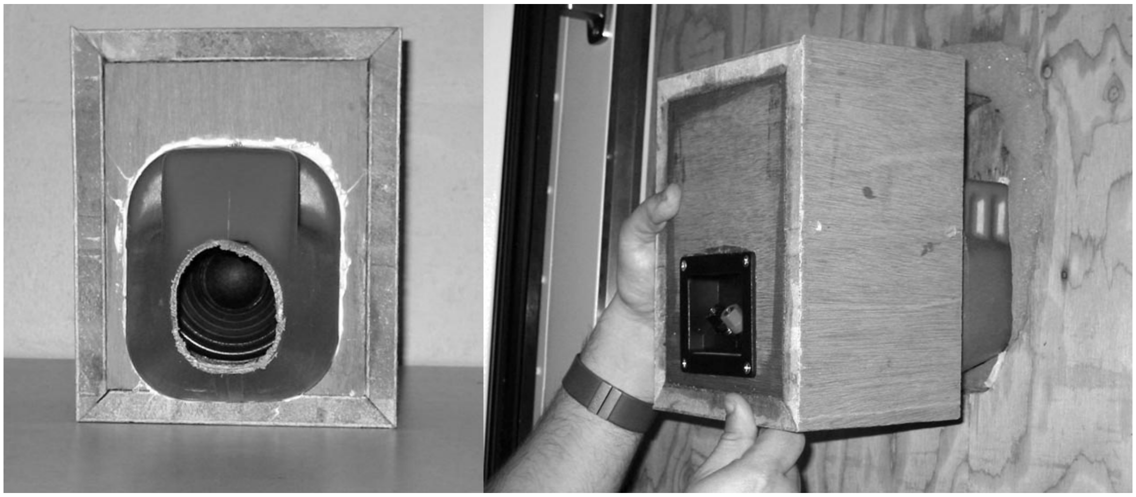


Figure 10.4. ***Left.*** **The funnel attached to the speaker. The cone of the speaker is visible through the opening of the funnel.** ***Right.*** **The funnel-speaker combination placed in the Door.**

The computed transfer functions are not true radiation characteristics; rather they relate the driving signal voltage to the pressure at the microphone. This relation can be written as:

$$\frac{P(f)}{V(f)} = H(f)R(f) = G(f) \tag{10.2}$$

where $P(f)$ is the pressure measured at the microphone, $V(f)$ is the voltage of the driving signal, $H(f)$ is the transfer function between the driving signal voltage and the volume velocity at the funnel opening, $R(f)$ is the radiation characteristic and $G(f)$ is the product of the two transfer functions. Because $H(f)$ should remain constant measuring $G(f)$ provides a useful estimate of any changes in the radiation characteristic that occur between the two situations. Figure 10.5 displays $G(f)$ for both the Door and inside the acoustic chamber.

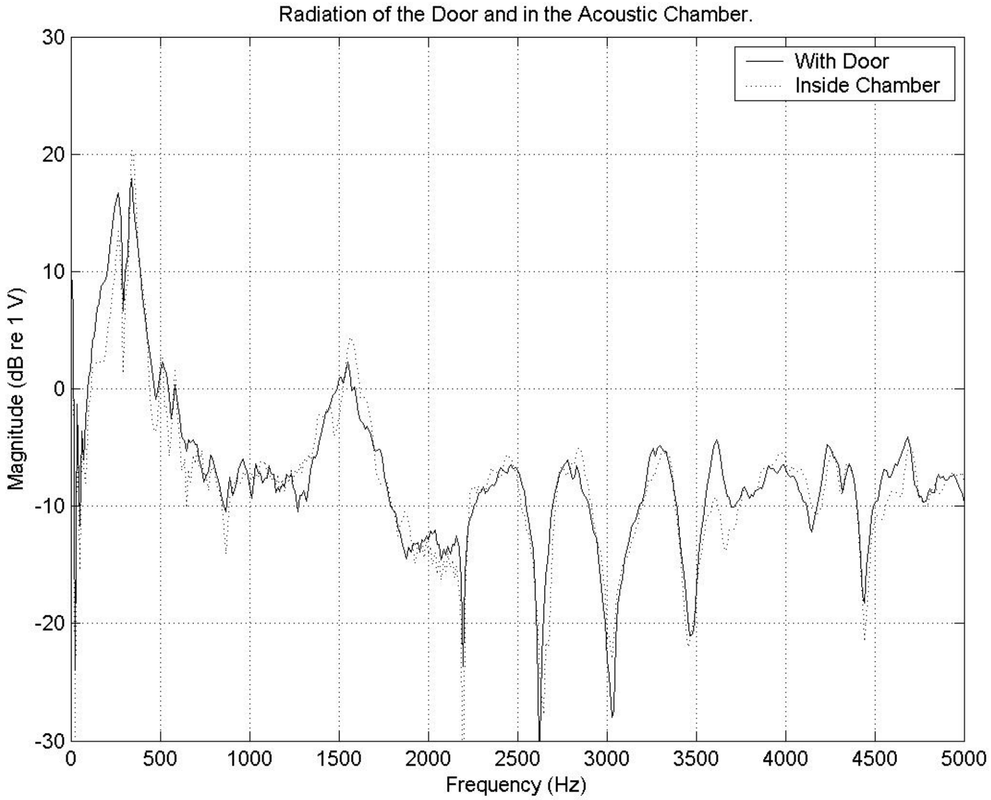


Figure 10.5. The spectra of the radiation from the Door and inside the acoustic chamber are very similar demonstrating the radiation characteristics of the two situations are almost identical.

As Figure 10.5 shows, the spectra of the two measured transfer functions are very similar, thus indicating that the radiation characteristics are practically identical. As such, it is safe to use a differentiator as an approximation for the Door radiation characteristic as was done in the modeling described in Section 5.

11. Appendix C. MELP

The **M**ixed **E**xcitation **L**inear **P**redictive (MELP) vocoder was developed as a means of providing high quality speech at low bit rates. Originally proposed by McCree *et al.* (1995), it was later formalized as a Federal Information Processing Standard (1999) for voice transmission at 2.4 kb/s. Based on a standard linear predictive (LP) vocoder, MELP contains several additional features that improve the quality of its outputted speech. Moreover, because it separates the voicing source into periodic and noisy components in addition to separating the voicing source from the vocal tract, the MELP lends itself to be used as means of modifying speech. It is for this reason that MELP was chosen as the means of changing the pitch contour of the perceptual sentences described in Chapter 3.

This appendix summarizes the MELP algorithm detailed in the Federal Standard and discusses modifications made in the version used in Chapter 3. Although the MELP standard specifies the specifics of the quantization and bit encoding, these will not be described here as they were not used in this study. Those interested in the encoding and quantization should refer to the MELP standard.

11.1. The Encoder

The MELP encoder works on speech that has been sampled at 8000 Hz. The following encoding operations are done every 22.5 ms or 180 points.

11.1.1. Step One: High Pass Filtering

Energy contained at frequencies 60 Hz and below are removed by pre-filtering the speech with a 4th order Chebyshev type II filter with a 60 Hz cutoff frequency and a stop band rejection of 30 dB. The output of this filter will be referred to as the input speech throughout the rest of this document.

11.1.2. Step Two: Initial Pitch determination

Although technically, it is the pitch *period* that is being calculated in this section, to be consistent with the MELP federal standard document, the terminology that document will be employed here.

11.1.2.1 Integer pitch calculation

Prior to the pitch calculation, the input speech is low pass filtered at 1 kHz using a 6th order Butterworth filter. The integer pitch, P_1 is then defined as the value of τ, $\tau = 40$, 41, …, 160, which maximizes the normalized autocorrelation function $r(\tau)$ where

$$r(\tau) = \frac{c_\tau(0,\tau)}{\sqrt{c_\tau(0,0)c_\tau(\tau,\tau)}} \qquad (11.1)$$

and

$$c_\tau(m,n) = \sum_{k=-\left\lfloor \frac{\tau}{2} \right\rfloor -80}^{-\left\lfloor \frac{\tau}{2} \right\rfloor +79} s_{k+m} s_{k+n} \qquad (11.2)$$

and $\left\lfloor \frac{\tau}{2} \right\rfloor$ represents truncation to an integer value. The autocorrelation is centered on sample s_0 which is defined as the last sample in the current frame.

11.1.2.2 Bandpass Voicing Analysis

This part of the encoder determines voicing strengths of five distinct frequency bands, Vbp_i, $i = 1, 2, \ldots, 5$ and refines the integer pitch estimate and its corresponding normalized autocorrelation value. To begin this analysis, the input speech is filtered into 5 frequency bands using 6th order Butterworth filters with passbands of 0-500 Hz, 500-1000 Hz, 1000-2000 Hz, 2000-3000 Hz, and 3000-4000 Hz.

The integer pitch refinement is conducted on the output of the lowest passband filter. The measurement is centered on the filter output sample that corresponds to when its input is the last sample of the current frame. Eq (1) is used to conduct an integer pitch search on lags from 5 samples shorter to 5 samples longer of the values of P_1 from both the current frame and the previous frame. A fractional pitch estimate (see section 11.1.2.3) and corresponding autocorrelation value are computed for both pitch candidates. The pitch estimate candidate that has the highest autocorrelation value is chosen as the fractional pitch estimate, P_2. The autocorrelation value, $r(P_2)$ is used as the voicing strength of the lowest band, Vbp_1.

The remaining bandpass voicing strengths are found by choosing the larger of $r(P_2)$ as computed from the fractional pitch procedure performed on the bandpass signal and the time envelope of the bandpass signal. To compensate for an experimentally observed bias, $r(P_2)$ is decreased by 0.1 for the time envelope. The envelopes are generated by passing a full-wave rectification of the bandpass signal through a smoothing filter. The smoothing filter consists of a zero at DC cascaded with a complex pole par at 150 Hz and a radius of 0.97.

11.1.2.3 Fractional Pitch Refinement

This procedure attempts to improve the accuracy of the integer pitch estimate by interpolating between successive pitch periods to find a fractional offset. If it is assumed that the integer pitch has a value of T samples, then the interpolation formula assumes that the true maximum of $r(\tau)$ falls between T and $T+1$ or T and $T-1$ samples. Therefore

$c_T(0,T+1)$ and $c_T(0,T-1)$ are computed to determine if the maximum is more likely to fall between T and $T+1$ or T and $T-1$. If $c_T(0,T-1) > c_T(0,T+1)$ then the maximum is most likely between T and $T+1$, and the pitch is decremented by one prior to interpolation. The factional offset, Δ, is determined by

$$\Delta = \frac{c_T(0,T+1)c_T(T,T) - c_T(0,T)c_T(T,T+1)}{c_T(0,T+1)[c_T(T,T) - c_T(T,T+1)] + c_T(0,T)[c_T(T+1,T+1) - c_T(T,T+1)} \qquad (11.3)$$

with a normalized autocorrelation function of

$$r(T+\Delta) = \frac{(1-\Delta)c_T(0,T+1)}{\sqrt{c_T(0,0)[(1-\Delta)^2 c_T(T,T) + 2\Delta(1-\Delta)c_T(T,T+1) + \Delta^2 c_T(T+1,T+1)}} \qquad (11.4)$$

The offset, Δ, is clamped between –1 and 2 while the fractional pitch estimate is clamped between 20 and 160.

11.1.3. Aperiodic Flag

The aperiodic flag is set to 1 if $Vbp_1 < 0.5$ and set to 0 otherwise. This flag is used in the decoder to make the pulse component of the excitation aperiodic.

11.1.4. Linear Predictive Analysis

A 10th order linear predictive (LP) analysis is performed on a 200 sample Hamming windowed segment centered around the last sample in the frame. The LP coefficients were computed using Levinson-Durbin recursion. The coefficients were then multiplied by 0.994^i, $i=1,2,\ldots 10$, to implement a 15 Hz bandwidth expansion. The LP residual was then computed by filtering the input signal by a prediction filter comprised of the LP coefficients. Again, the window for this computation is centered on the last sample in the frame and has a width great enough to be used by the final pitch calculation. The linear predictive coefficients are then converted into line spectral frequencies (LSFs) which are then sorted into ascending order and are checked to ensure that there is at least a 50 Hz separation between adjacent LSFs.

11.1.5. Peakiness Calculation

The peakiness of the residual signal, r_n, is calculated over 160 sample window centered on the last sample of the current frame, and is defined as the ratio of the L2 norm to the L1 norm, i.e.

$$peakiness = \frac{\sqrt{\frac{1}{160}\sum_{n=1}^{160} r_n^2}}{\frac{1}{160}\sum_{n=1}^{160} |r_n|} \tag{11.5}$$

If the peakiness is greater than 1.34, than the lowest band voicing strength, Vbp_1 is set to 1.0. If the peakiness exceeds 1.6, then the lowest three voicing strengths are set to 1.0.

11.1.6. Final Pitch Calculation

The final pitch calculation is performed on the residual signal that has been low pass filtered using a 6th order Butterworth filter with a 1000 Hz cutoff. Using Eq (1), an integer pitch search is conducted over lags from 5 samples shorter to 5 samples longer than P_2, rounded to the nearest integer. A fractional pitch estimate is then computed on the optimal pitch lag, producing a candidate value for the final pitch estimate, P_3 and its corresponding autocorrelation value, $r(P_3)$.

If $r(P_3) \geq 0.6$, a pitch doubling check is performed (see Section 11.1.6) on the low pass filtered residual using a doubling threshold, $D_{th} = 0.75$ if $P_3 \leq 100$ or $D_{th} = 0.5$ otherwise. The doubling check procedure may produce new values of P_3 and $r(P_3)$.

If $r(P_3) \leq 0.6$ then a pitch refinement is performed around P_2 using the input speech signal, producing new values of P_3 and $r(P_3)$. If $r(P_3) < 0.55$ then P_3 is replaced by P_{avg}, the longer term average pitch (see Section 11.1.9). Otherwise the pitch doubling procedure is performed on P_3 using $D_{th} = 0.9$ if $P_3 \leq 100$ or $D_{th} = 0.7$ otherwise. Again the doubling check procedure may produce new values of P_3 and $r(P_3)$.and once more, if $r(P_3) < 0.55$ then P_3 is replaced by P_{avg}.

11.1.7. Pitch Doubling Check

The pitch doubling check procedure searches for pitch estimates that are multiples of the true pitch. The procedure starts by conducting a fractional pitch refinement around a candidate pitch value, P, produce tentative values for the checked pitch, P_c and the corresponding value of the autocorrelation, $r(P_c)$. Then, the largest value of k is found where $r(P_c/k) > D_{th}r(P_c)$, where $(P_c/k) \geq 20$ and $k = 8,7,\ldots,2$. If such a value of k is exists then a fractional pitch refinement is conducted around P_c/k producing new values of P_c and $r(P_c)$. If $P_c/k < 30$ then a double verification is performed.

11.1.8. Gain Calculation

The gain of the input speech signal is measured twice per frame using a pitch adaptive window length. The window length is identical for both pitch measurements within a frame. If $Vbp_1 > 0.6$, the window length is the shortest multiple of P_2 which is longer than 120 samples. If this value is greater than 320 samples, then it is divided by 2. If $Vbp_1 < 0.6$ then the window length is fixed at 120 samples. The first measurement is centered at 90 samples before the last sample of the frame and produces gain, G_1. The second gain estimate, G_2 is computed using a window centered around the last sample of the frame. The gain is the RMS value in dB, of the signal in the window, s_n is:

$$G_i = 10\log_{10}\left(0.01 + \frac{1}{L}\sum_{n=1}^{L} s_n^2\right) \tag{11.6}$$

where L is the window length. If the gain measurement is less than 0.0 then it is fixed at 0.0.

11.1.9. Average Pitch Update

The long term pitch average, P_{avg} is updated as follows. If $r(P_3) > 0.8$ and $G_2 > 30$ dB then P_3 is placed into a buffer of the three most recently found strong pitch values, p_i, i= 1, 2, 3. Otherwise, all the pitch values in the buffer are moved to a pitch default, $P_{default}$ = 50 samples as follows:

$$p_i = 0.95 p_i + 0.05 P_{default}\ ,\ i = 1, 2, 3 \tag{11.7}$$

P_{avg} is then the median of the values in the pitch buffer and is used in the final pitch calculation.

11.1.10. Bandpass Voicing Quantization

If $Vbp_1 \leq 0.6$ (i.e. unvoiced) then the remaining voicing strengths are set to zero. If $Vbp_1 > 0.6$ then the remaining voicing strengths are set to 1 if their values exceed 0.6. Otherwise, they are set to 0.

11.1.11. Fourier Magnitude Calculation

The amplitudes of the first 10 harmonics of the residual signal are measured by first computing the magnitude of a 512 point FFT on a 200 sample window centered on the last sample of the frame. A spectral peak picker is used to find the amplitudes of the harmonics. The peak picker first finds the maximum within a $512/P_3$ bin centered around the initial estimate for each pitch harmonic. The initial estimate of the *ith* harmonic is $512i/P_3$. The smaller of 10 or $P_3/4$ harmonics are measured and are normalized so that the they have an RMS value of 1.0. If fewer than 10 harmonics are found then the remaining magnitudes are set to 1.0.

11.1.12. Encoder Summary

The analysis performed by the encoder produces the following parameters for each analysis frame which are "transmitted" to the decoder: the final pitch, P_3, the aperiodic flag, the line spectral frequencies, the 10 residual harmonics, the 5 voicing strengths, $Vbp_{i,}$ and the two gain values, G_1 and G_2. In the standard MELP implementation these values are all quantized and encoded. However for the work described in this thesis, the encoding was unnecessary and not performed. Thus the output of the version of the encoder used in Chapter 3 was a set of arrays and matrices of the analysis parameters.

11.2. The Decoder

11.2.1. Voiced/Unvoiced decision

The MELP decoder works on a frame-by-frame basis and pitch synchronously interpolates the received parameters between frames. Prior to the interpolation, however, the decoder decides whether the synthesis will occur in the voiced mode or the unvoiced mode. In the unvoiced mode, (i.e. $Vbp_1 = 0$), the default parameter values are used for the pitch, jitter, bandpass voicing, and Fourier magnitudes. The pitch value is set to 50 samples, the jitter is set to 25%, all of the bandpass voicing strengths are set to 0, and the Fourier magnitudes are set to 1. In the voiced mode, *Vbp1* is set to 1; jitter is set to 25% if the aperiodic flag is a 1; otherwise jitter is set to 0%. The bandpass voicing strength for the upper four bands is set to 1 if the corresponding bit is a 1; otherwise the voicing strength is set to 0.

11.2.2. Noise Attenuation

A small amount of gain attenuation is applied to both gain parameters and is performed as follows. A background noise estimate is updated as follows. If $G_1 > G_n + C_{up}$ then $G_n = G_n + C_{up}$. If $G_2 > G_n + C_{up}$ then $G_n = G_n + C_{down}$. Otherwise, $G_n = G_1$. $C_{up} = 0.0337435$ and $C_{down} = 0.135418$ so that estimator moves up at 3 dB/second and down at 12 dB/second. The noise estimate is initialized at 10 and clamped between 10 and 20 dB. The gain, G_1 is then modified by subtracting G_{att} from it where

$$G_{att} = -10\log_{10}\left(1-10^{0.1[G_n+3-G_1]}\right) \tag{11.8}$$

G_{att} is clamped to a maximum value of 6dB. The noise estimation and gain modification steps are repeated for the second gain estimate, G_2.

11.2.3. Parameter Interpolation

All MELP parameters (except for the aperiodic flag) are interpolated pitch-synchronously for each synthesized pitch period. If the starting point of the synthesis, t_0, $t_0 = 0,1,\ldots,179$, is less than 90, then *t*he gain (in dB) is linearly interpolated between the second gain of the previous frame, G_{2p} and the first gain of the current frame. Otherwise the gain is interpolated between the first gain of the current frame, G_1 and the second gain, G_2. The other parameters are interpolated between the previous and current frame values using the following interpolation factor:

$$\text{int} = \frac{t_0}{180}. \tag{11.9}$$

The are two exceptions to this interpolation rule. First, if G_1 is more that 6 dB greater than G_2 and the pitch period of the current frame is less than half of that of the previous frame, the pitch interpolation is disabled and the current pitch period is used. Second, if G_2 is 6 dB greater than G_{2p} then the LSFs, spectral tilt, and pitch are interpolated using:

$$\text{int} = \frac{G_{\text{int}} - G_{2p}}{G_2 - G_{2p}} \tag{11.10}$$

where G_{int} is the interpolated gain. The interpolation factor is clamped between 0 and 1.

11.2.4. Mixed Excitation Generation

The MELP mixed excitation signal is comprised of a periodic pulse excitation and a noise excitation. The pulse excitation, $e_p(n)$, n = 0, 1, …, T-1 is computed by performing an inverse Discrete Fourier Transform (DFT) of one pitch period in length:

$$e_p(n) = \frac{1}{T}\sum_{k=0}^{T-1} M(k)e^{\frac{j2\pi nk}{T}} \tag{11.11}$$

The pitch period, T, is the interpolated pitch value plus the jitter time the interpolated pitch value where jitter is the interpolated jitter strength times the output of a random number generator between -1 and 1. This pitch period is rounded to the nearest integer and held between 20 and 160.

Because the phases of $e_p(n)$ are set to zero, the $M(k)$ are real and because $e_p(n)$ is real, the $M(k)$ are symmetric and obey:

$$M(T-k) = M(k), \quad k = 1,2, \ldots, L \tag{11.12}$$

where $L = T/2$ is T is even, and $L = (T-1)/2$ if T is odd. The DC term, $M(0)$, is set to zero, while $M(k)$, k = 1, 2, …, 10, are set to the interpolated Fourier magnitude values. The magnitudes not specified are set to 1. After the inverse DFT is performed, the excitation pulse is circularly shifted by 10 samples so that the main excitation pulse occurs at the 10th sample of the period. The pulse is then multiplied by the square-root of the pitch and by 1000 to give the proper signal level.

The noise excitation of length T is produced by a uniform random noise generator with an RMS value of 1000 and range of -1732 to 1732.

The pulse excitation and noise components are then filtered through a bandpass filter bank analogous to the one used in the encoding. The filter coefficients are pitch-synchronously interpolated. The bandpass filter coefficients are multiplied by the corresponding interpolated voicing strength, Vbp_i, and summed prior to filtering the pulse excitation. For the noise excitation the filter coefficients are multiplied by ($1-Vbp_i$). The filtered outputs are then summed to produce the mixed excitation signal. The filter coefficients can be found in the Federal Standard document.

11.2.5. Adaptive Spectral Enhancement

The mixed excitation signal is filtered through a tenth order pole/zero adaptive spectral enhancement filter with an additional first-order tilt compensation. The filter coefficients are obtained from a bandwidth expansion of the linear prediction transfer function, $A(z)$, which is obtained from the interpolation of the LSFs. The enhancement filter, $H_{ase}(z)$ is given by:

$$H_{ase}(z) = \frac{A(\alpha z^{-1})}{A(\beta z^{-1})} \cdot \left(1 + \mu z^{-1}\right) \tag{11.13}$$

where $\alpha = 0.5p$, $\beta = 0.8p$, and the spectral tilt coefficient μ which is first calculated as $min(0.5k_1,0)$, interpolated, and then multiplied by p, the signal probability. The first reflection coefficient, k_1, is obtained from the LSFs and is typically negative during voiced frames. The signal probability, p, is computed by

$$p = \frac{G_{\text{int}} - G_n - 12}{18} \tag{11.14}$$

and is clamped between 0 and 1.

11.2.6. Linear Prediction Synthesis

The spectrally enhanced excitation signal is filtered with a direct form synthesis filter whose coefficients are obtained from the interpolated LSFs.

11.2.7. Gain Adjustment

The synthesized speech is multiplied by a gain scaling factor, S_{gain} which is computed for each pitch period as

$$S_{gain} = \frac{10^{\frac{G_{\text{int}}}{20}}}{\sqrt{\frac{1}{T}\sum_{n=1}^{T} \hat{s}_n^2}} \tag{11.15}$$

where $\hat{s}_n$ is the synthesized speech signal. The prevent discontinuities in the speech, this scale factor is linearly interpolated between the previous and current values for the first ten samples of the pitch period.

11.2.8. Pulse Dispersion

The synthesized speech is filtered with a 65th order FIR filter derived from a spectrally-flattened triangle pulse. The coefficients can be found in the appendix of the Federal Standard document.

11.2.9. Synthesis Loop Control

After the pitch period is processed, the decoder updates the next synthesis starting point, t_0 by adding T (i.e. $t_0 = t_o + T$). If $t_o < 180$, then the synthesis of the current frame continues from the parameter interpolation step. Otherwise, the remainder of the current period which extends beyond the end of the current frame is buffered and 180 is subtracted from t_0 to set the starting point for the next frame.

11.3. Modifications to MELP

Although standard MELP produces reasonably good quality synthesized speech at a low bit rate (2.4 kbs), it was easy to distinguish between the original and MELP speech. As such, modifications were made to the MELP algorithm to improve the quality of the synthesized speech. The modifications are as follows.

1. Because the processing was not done in real time, maintaining a low bit rate was not a concern. Thus, to improve the time resolution of the coder and to help reduce the number of discontinuities between frames, the parameter estimates were made every 5 ms instead of every 22.5 ms.
2. To improve the frequency resolution of the LP estimate, the LP estimation window was increased from 200 samples to 320 samples.
3. The standard MELP vocoder makes a rough pitch period estimate on a low pass filtered version of the input speech and then proceeds to make two finer pitch period estimates based on the initial estimate. This scheme often resulted errors from the first estimate being propagated through to the second and third pitch period estimates. These errors were not always detected by the pitch correction measures such as the pitch doubling check and as such, they degraded the quality of the synthesized speech. The initial pitch estimate was especially erroneous in EL speech because of its lack of low frequency energy Thus, only a single pitch period estimate was made as described in section 11.1.6 except that lag values from 20 to 160 were used in the search.

4. Standard MELP quantizes the bandpass voicing strengths to 1 or 0 in order to efficiently encode them. However, because no encoding was required in this situation, the voicing strengths were left unquantized. This allowed for a more

nuanced mixing of the pulsed and noise components of the synthesized excitation signals.

Implementing these changes produced resynthesized speech that was informally judged by experiences listeners of EL speech to be of a higher quality than that produced by standard MELP. Although it was difficult to differentiate between the synthesized speech produced by the modified MELP vocoder and the original speech, all tokens that were presented in the perceptual experiments described in Chapter 3 were processed using the modified MELP algorithm.

12. Appendix D. The Zoo Passage.

The following passage was used in the VA speech recordings discussed in Section 4.

The Trip to the Zoo.

Last Sunday Bob went to the zoo with his mother and father. His sister Mary and his brother George went along too. Mother packed a big basket full of good things to eat. Father took the car to the service station to get gas and have the oil checked. The family left the house at eleven o’clock and got to the zoo at twelve o’clock. As you can see, they didn’t have far to go.

13. Appendix E. Data from analysis of VA Database

Table 13.1. Mean Formant Frequencies of 9 Vowels for Pre- and Post-Laryngectomy Speech

		Post-Laryngectomy			Pre-Laryngectomy		
Vowel		**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**
/i/	**F1 (Hz)**	418	363	481	361	281	434
	F2 (Hz)	2362	1917	2544	2054	1538	2569
	F3 (Hz)	2964	2654	3266	2473	2014	2879
/I/	**F1 (Hz)**	550	435	677	452	383	534
	F2 (Hz)	1899	1573	2089	1542	1257	1767
	F3 (Hz)	2815	2336	3427	2449	2066	2895
/ɛ/	**F1 (Hz)**	534	444	588	405	272	535
	F2 (Hz)	1974	1341	2441	1668	1399	1993
	F3 (Hz)	2751	2287	3279	2294	1889	2624
/æ/	**F1 (Hz)**	754	605	881	608	549	674
	F2 (Hz)	1918	1721	2095	1617	1504	1725
	F3 (Hz)	2998	2412	3423	2460	2014	3016
/a/	**F1 (Hz)**	838	667	1113	650	552	776
	F2 (Hz)	1367	1143	1667	1170	1015	1377
	F3 (Hz)	2783	2497	3255	2283	1807	2720
/U/	**F1 (Hz)**	555	479	674	452	395	513
	F2 (Hz)	1541	1259	1828	1435	1363	1613
	F3 (Hz)	2624	1970	3052	2265	1897	2973
/u/	**F1 (Hz)**	838	415	578	650	300	449
	F2 (Hz)	1367	986	1831	1170	888	1731
	F3 (Hz)	2783	1885	3082	2283	1995	2619
/Λ/	**F1 (Hz)**	670	556	723	549	420	537
	F2 (Hz)	1528	1245	1965	1326	1138	1526
	F3 (Hz)	2476	2058	2871	2383	2212	2539
/e^r/	**F1 (Hz)**	621	556	726	479	420	537
	F2 (Hz)	1562	1338	1710	1464	1294	1731
	F3 (Hz)	2400	1658	2903	2005	1675	2510

Table 13.2. Mean Bandwidths of 9 Vowels for Pre- and Post-Laryngectomy Speech

		Post-Laryngectomy			Pre-Laryngectomy		
Vowel		**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**
/i/	**BW1 (Hz)**	39	7	109	45	26	67
	BW2 (Hz)	171	43	305	152	75	282
	BW3 (Hz)	194	43	145	193	75	132
/I/	**BW1 (Hz)**	40	10	81	87	46	128
	BW2 (Hz)	81	31	211	158	51	484
	BW3 (Hz)	208	68	551	200	76	304
/ɛ/	**BW1 (Hz)**	42	9	96	83	53	120
	BW2 (Hz)	150	41	277	189	48	309
	BW3 (Hz)	221	100	346	214	63	441
/æ/	**BW1 (Hz)**	38	18	89	71	20	130
	BW2 (Hz)	156	53	389	127	74	194
	BW3 (Hz)	252	111	600	228	90	431
/a/	**BW1 (Hz)**	59	19	137	86	17	158
	BW2 (Hz)	91	28	162	86	39	133
	BW3 (Hz)	226	84	409	209	123	395
/U/	**BW1 (Hz)**	53	18	117	102	26	150
	BW2 (Hz)	94	33	204	104	63	193
	BW3 (Hz)	180	52	363	257	89	369
/u/	**BW1 (Hz)**	59	13	66	86	35	170
	BW2 (Hz)	91	37	229	86	39	261
	BW3 (Hz)	226	53	288	209	56	572
/^/	**BW1 (Hz)**	50	9	167	94	29	138
	BW2 (Hz)	80	16	171	130	63	362
	BW3 (Hz)	157	66	238	216	61	403
/e^{r}/	**BW1 (Hz)**	57	9	167	88	29	138
	BW2 (Hz)	83	38	179	132	60	275
	BW3 (Hz)	153	63	281	278	66	651

Table 13.3 Mean Relative Formant Amplitudes for Pre- and Post-Laryngectomy Speech

		Post- Laryngectomy			Pre-Laryngectomy		
Vowel		Mean	Min	Max	Mean	Min	Max
/i/	A1-A2 (dB)	11.7	-8.6	28.9	19.6	8.6	32.7
	A2-A3 (dB)	1.4	-7.3	21.6	2.4	-7.0	31.8
	A1-A3 (dB)	13.2	-7.3	-8.3	21.9	-7.0	28.8
/I/	A1-A2 (dB)	9.4	-5.9	24.9	15.5	6.3	27.5
	A2-A3 (dB)	11.7	2.9	17.7	9.1	-2.5	22.8
	A1-A3 (dB)	21.1	7.0	36.6	24.6	16.6	34.8
/ɛ/	A1-A2 (dB)	12.7	-9.5	30.7	16.8	5.2	26.5
	A2-A3 (dB)	6.9	-8.6	32.3	3.5	-6.0	16.7
	A1-A3 (dB)	19.6	5.5	26.4	20.3	11.2	34.8
/æ/	A1-A2 (dB)	14.8	7.1	21.2	11.3	-2.0	21.0
	A2-A3 (dB)	10.9	2.7	19.4	10.1	-1.0	21.4
	A1-A3 (dB)	25.7	16.0	34.2	21.3	12.9	33.7
/ a /	A1-A2 (dB)	5.4	-4.1	14.7	5.9	-1.6	14.6
	A2-A3 (dB)	19.9	7.6	29.1	17.8	10.9	35.0
	A1-A3 (dB)	25.3	14.8	33.6	23.8	18.5	33.4
/U/	A1-A2 (dB)	11.3	-3.3	20.2	11.7	1.5	19.6
	A2-A3 (dB)	10.1	-2.9	16.6	14.9	4.3	25.4
	A1-A3 (dB)	21.4	6.7	33.5	26.6	18.6	35.4
/u/	A1-A2 (dB)	5.4	1.5	20.8	5.9	3.3	22.7
	A2-A3 (dB)	19.9	3.9	21.2	17.8	-3.5	32.5
	A1-A3 (dB)	25.3	10.3	36.9	23.8	14.3	48.7
/^/	A1-A2 (dB)	10.0	-7.4	20.2	9.6	3.1	20.5
	A2-A3 (dB)	13.9	3.0	23.2	14.3	-5.5	27.7
	A1-A3 (dB)	23.9	17.3	34.1	23.9	15.8	29.3
/e^r/	A1-A2 (dB)	8.5	-7.4	20.2	11.3	3.1	20.5
	A2-A3 (dB)	9.7	2.0	16.1	9.3	0.6	26.6
	A1-A3 (dB)	18.1	-3.0	32.0	20.5	12.2	30.7

14. Appendix F. EL Speech Modeling Data

14.1. Cross-sectional areas used for both vocal tract models

These cross-sectional areas were generated by scaling the outputs of the Story & Titze (1998) algorithm by a factor of 2 in order to produce proper total vocal tract volumes.

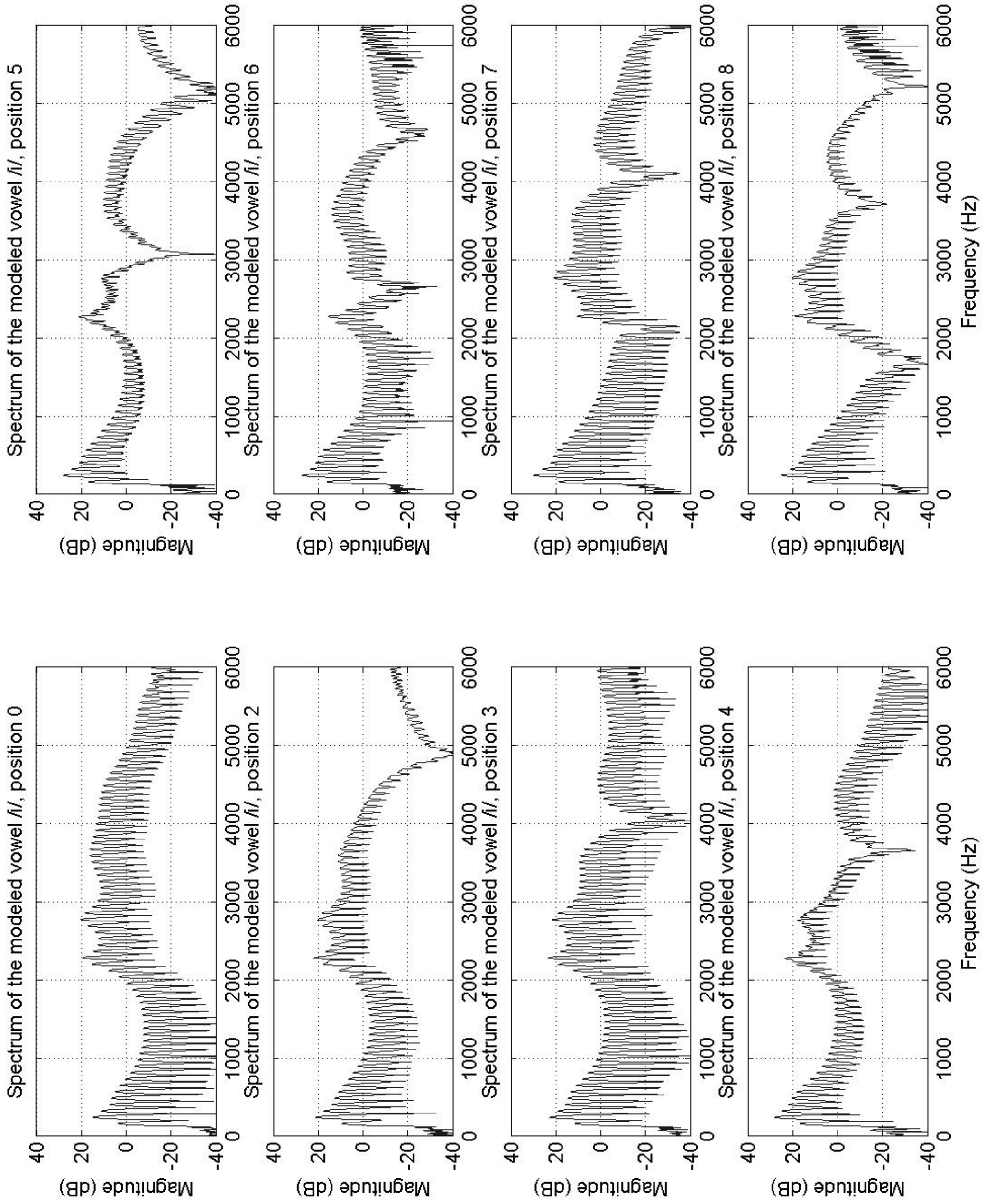
Table 14.1. Cross-sectional Areas (in cm^2) of Each Segment of the Male Vocal Tract Model

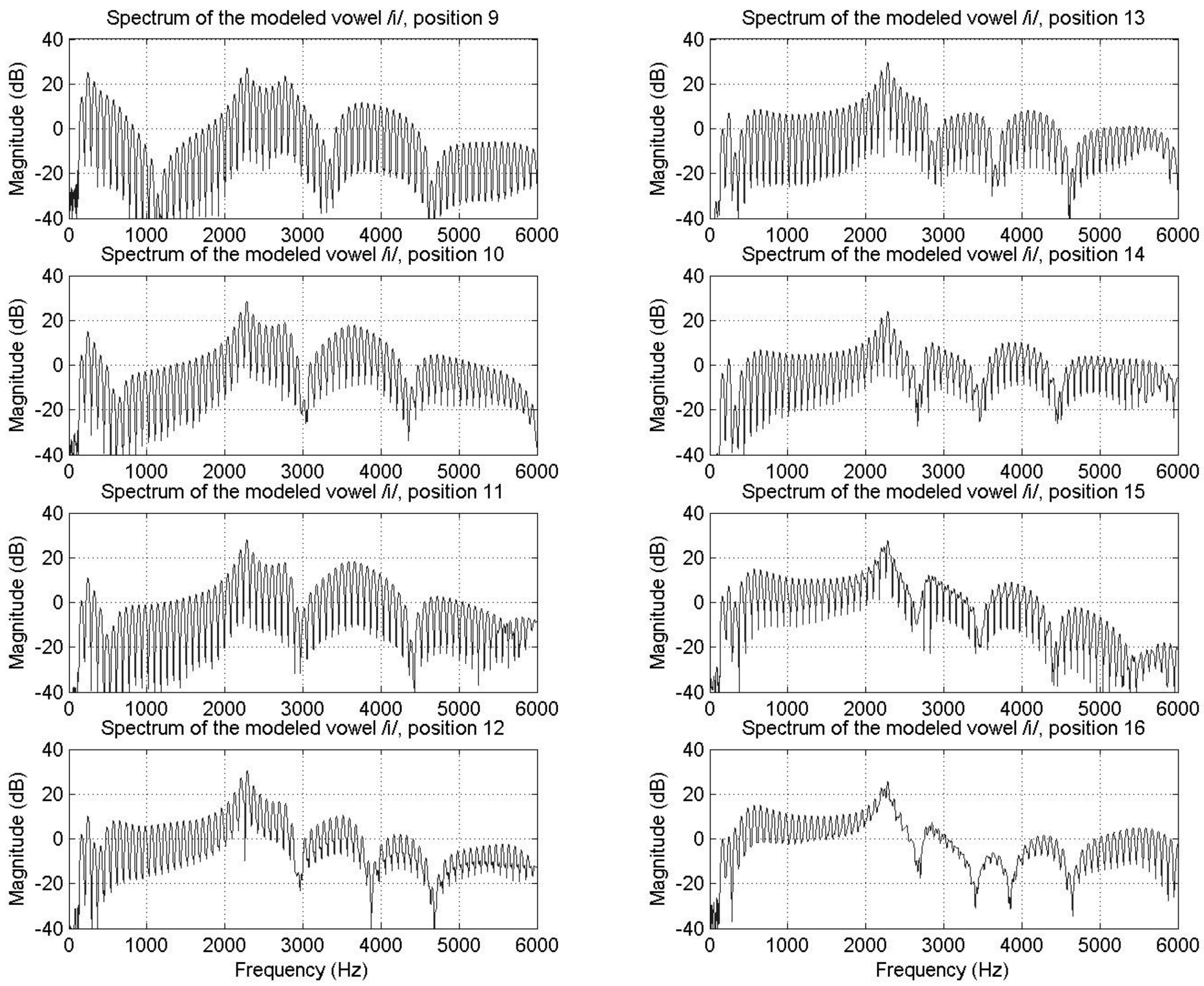
	Segment																
Vowel	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
/i/	0.52	0.80	2.48	5.56	8.08	10.06	10.60	7.96	3.60	0.94	1.22	0.30	0.24	1.14	3.06	3.20	1.92
/I/	0.52	1.22	3.88	4.86	5.84	6.54	6.68	6.28	4.58	3.74	1.50	1.54	2.40	3.94	4.74	3.20	1.64
/ε/	1.08	1.50	3.88	4.02	5.56	6.26	6.40	6.00	4.58	2.90	1.50	0.98	2.68	4.50	4.74	3.20	2.20
/æ/	1.22	1.50	3.04	2.62	3.48	3.74	3.88	4.32	3.94	3.46	3.10	4.04	5.48	7.02	6.42	4.04	2.48
/a/	1.08	1.22	3.04	1.54	1.54	1.34	1.34	1.76	2.68	3.18	4.78	8.06	11.64	12.90	9.50	4.04	2.20
/ɔ/	1.08	1.22	3.04	1.82	1.54	1.34	1.06	1.76	1.84	2.90	4.42	8.36	12.48	14.28	10.06	3.48	1.64
/U/	1.36	1.50	4.58	4.36	5.72	5.98	5.82	5.12	3.52	1.96	1.62	1.64	3.24	5.34	5.30	2.62	0.80
/u/	1.48	2.92	4.58	4.64	5.22	4.58	4.08	4.00	3.24	1.12	0.78	0.80	0.72	5.06	6.42	2.34	0.24
/^/	0.88	1.26	3.74	3.26	3.72	3.74	3.54	3.44	2.96	2.52	2.18	2.48	5.64	8.14	6.82	2.40	1.36
/e^r/	0.56	0.52	2.76	1.08	2.48	1.08	3.04	3.88	5.00	6.96	7.80	5.28	1.64	0.52	10.04	10.60	2.76

Table 14.2. Cross-sectional Areas (in cm^2) of Each Segment of the Female Vocal Tract Model

	Segment													
Vowel	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
/i/	0.58	4.44	6.40	9.48	10.04	8.64	5.84	2.76	0.52	0.52	1.08	3.04	3.32	2.48
/I/	0.60	4.42	5.56	7.80	6.68	4.16	4.16	2.20	0.52	1.08	2.48	4.44	2.20	1.70
/ε/	0.60	3.30	4.16	5.56	5.84	5.68	4.58	3.32	4.44	1.36	3.88	5.44	2.88	2.54
/æ/	0.74	3.02	2.48	2.76	3.04	3.44	3.46	3.60	3.88	3.88	7.80	8.52	4.52	2.54
/a/	0.74	2.46	1.64	1.36	1.34	1.54	2.06	3.04	3.88	7.24	13.12	12.72	5.64	1.70
/ɔ/	2.06	2.48	1.36	0.84	0.66	1.08	1.64	2.80	3.72	6.40	16.20	14.52	6.12	1.92
/U/	0.66	3.88	4.16	4.46	4.30	5.00	2.84	4.22	1.76	3.32	6.40	7.80	3.32	0.88
/u/	1.12	5.56	6.40	6.72	7.10	5.56	4.94	4.50	0.46	0.80	3.62	5.56	2.48	0.60
/^/	0.84	3.60	3.26	3.34	3.18	3.06	2.84	2.54	2.70	4.44	7.82	8.92	3.88	1.44
/e^r/	0.52	0.52	4.38	0.94	0.94	3.06	3.96	5.90	9.70	4.80	0.56	1.92	11.16	1.44

14.2. Vowel spectra generated from the model of the male vocal tract





Spectrum of the modeled vowel /i/, position 9

Spectrum of the modeled vowel /i/, position 10

Spectrum of the modeled vowel /i/, position 11

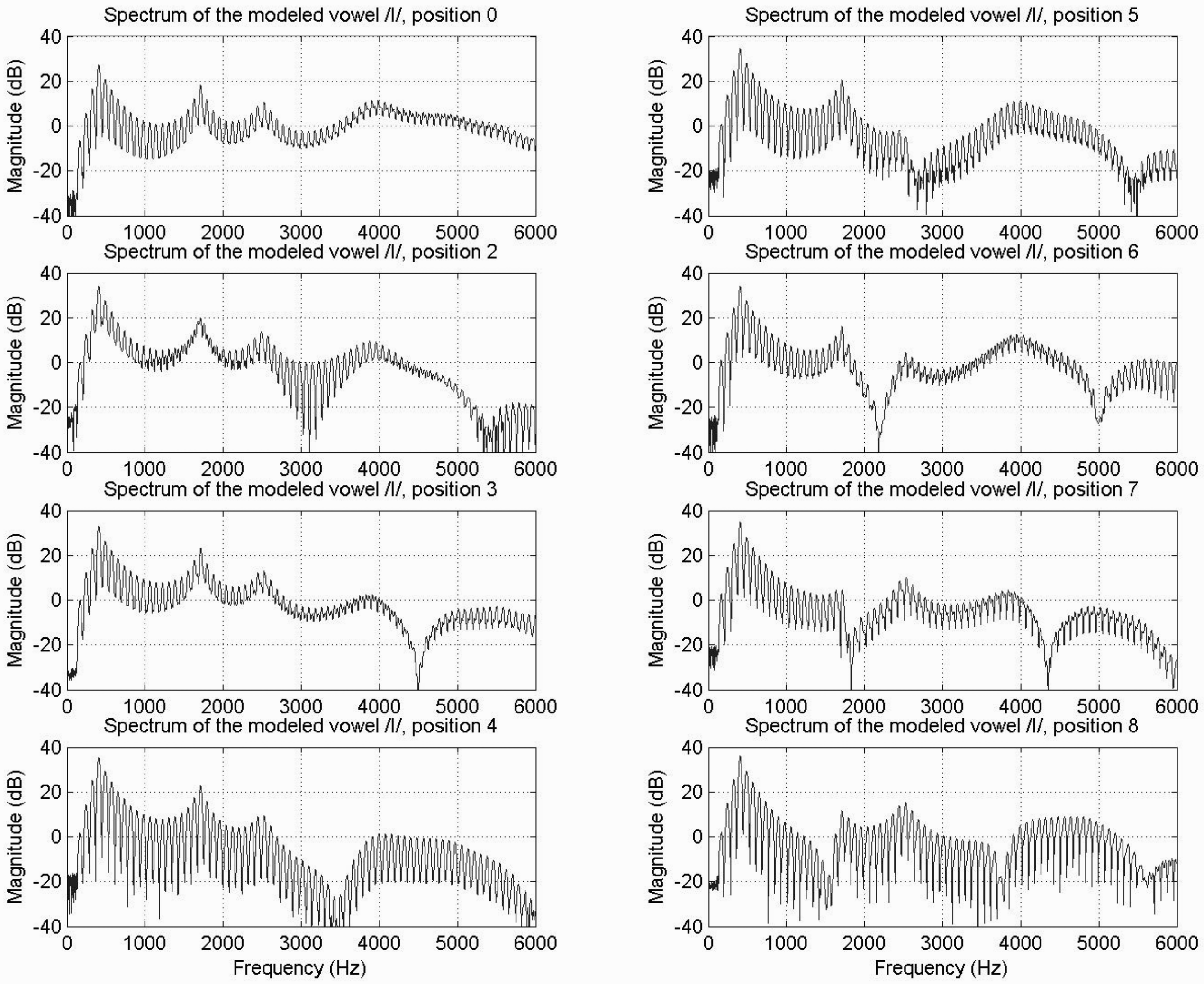
Spectrum of the modeled vowel /i/, position 12

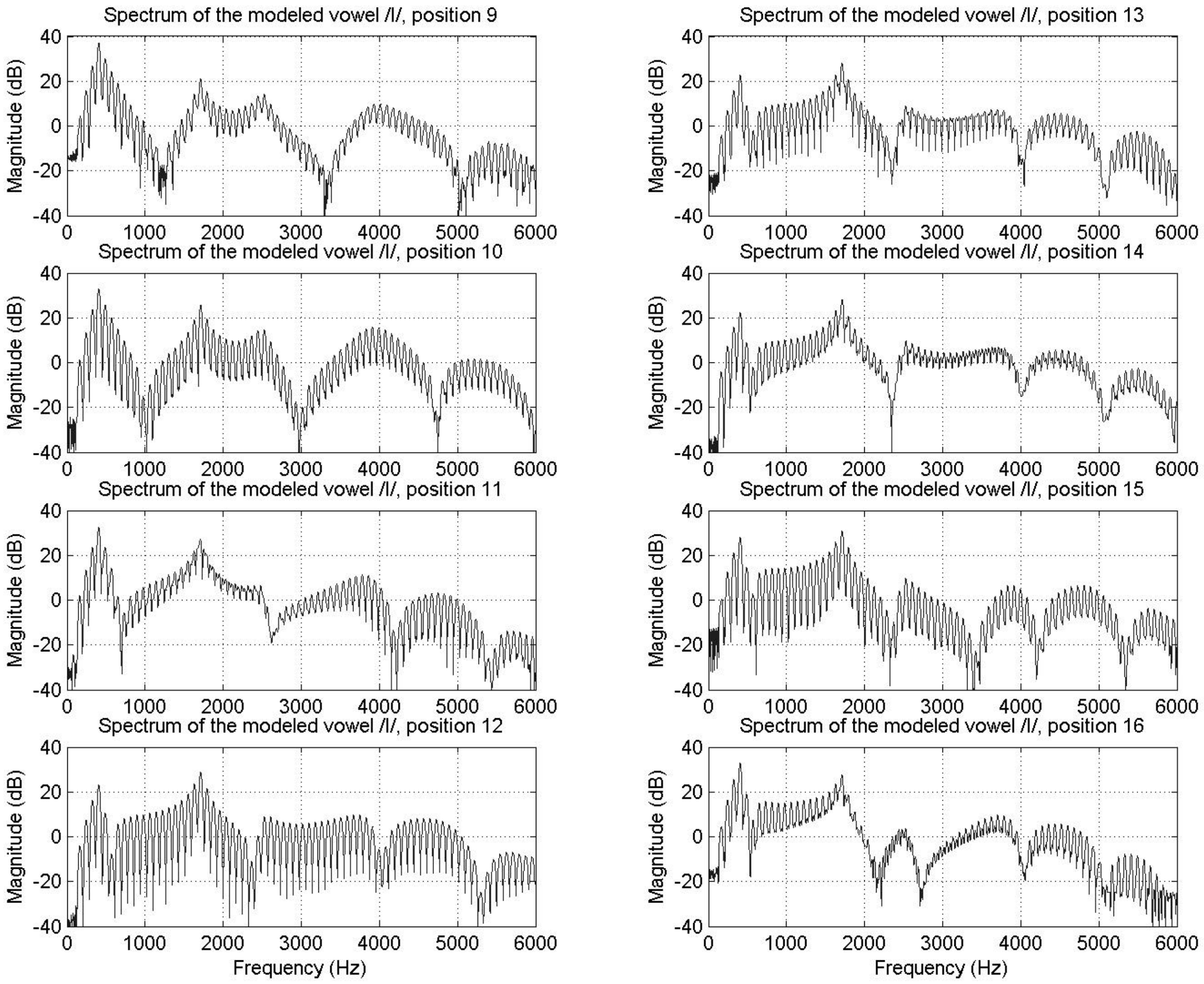
Spectrum of the modeled vowel /i/, position 13

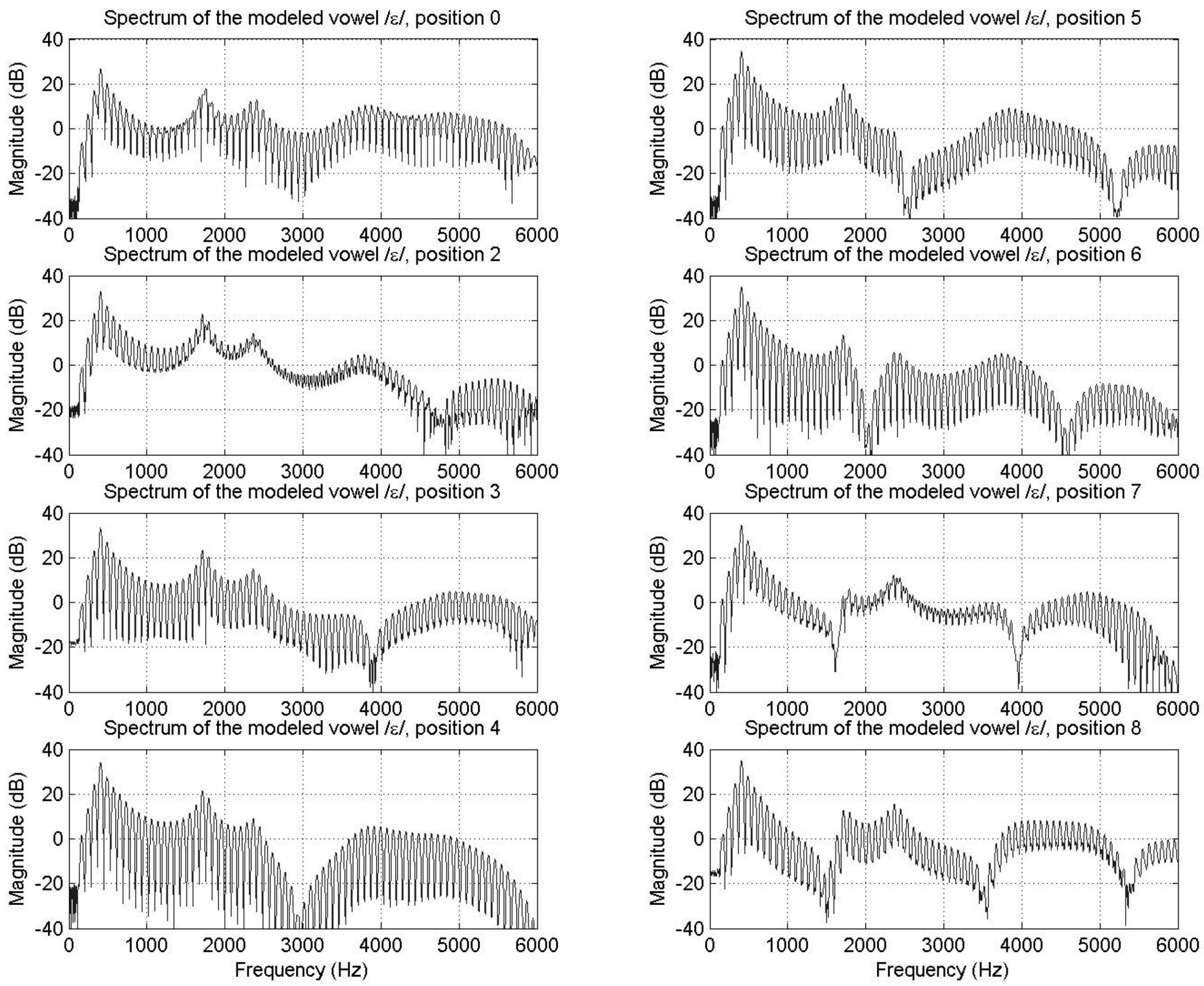
Spectrum of the modeled vowel /i/, position 14

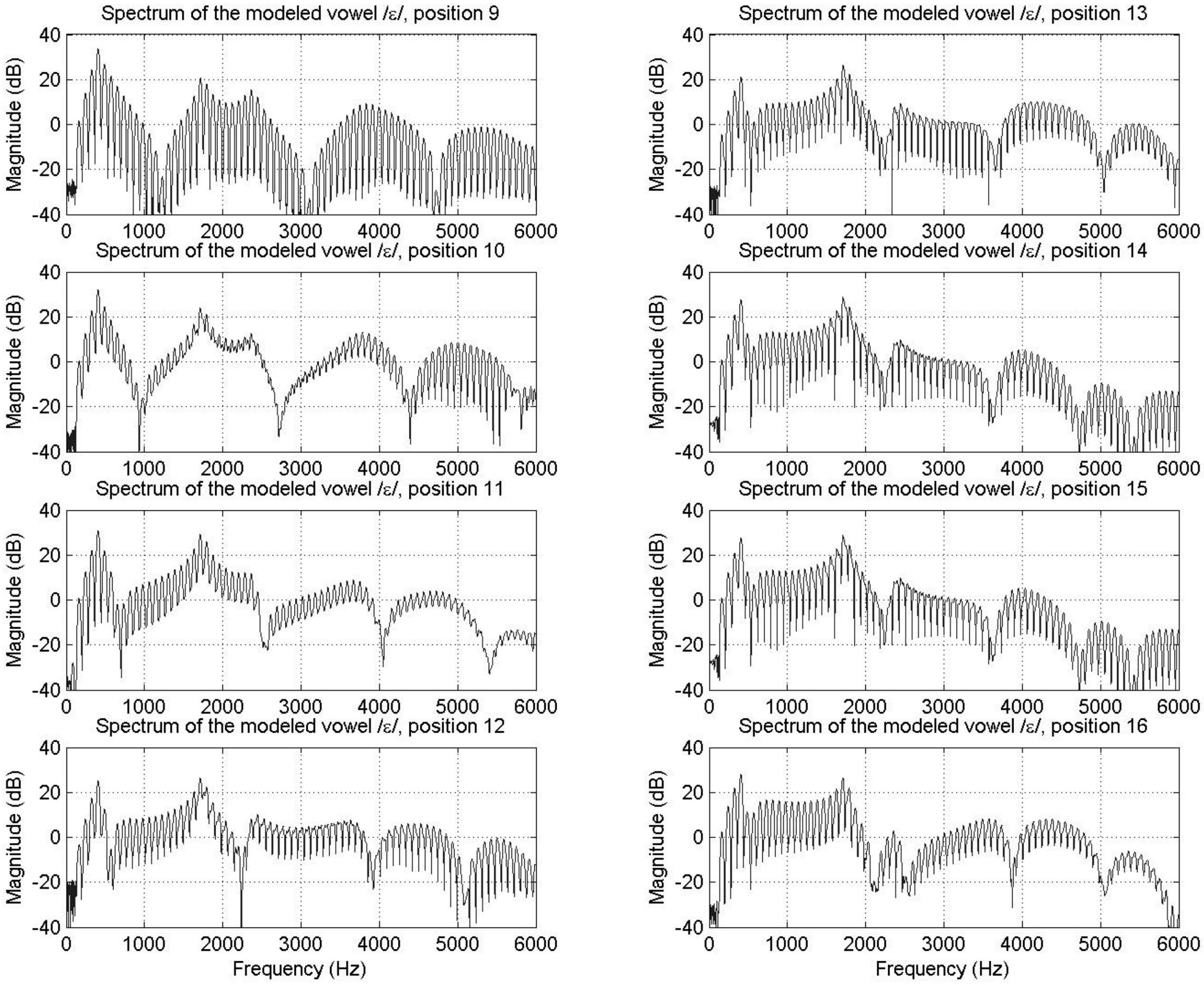
Spectrum of the modeled vowel /i/, position 15

Spectrum of the modeled vowel /i/, position 16


Spectrum of the modeled vowel /I/, position 0
Spectrum of the modeled vowel /I/, position 2
Spectrum of the modeled vowel /I/, position 3
Spectrum of the modeled vowel /I/, position 4
Spectrum of the modeled vowel /I/, position 5
Spectrum of the modeled vowel /I/, position 6
Spectrum of the modeled vowel /I/, position 7
Spectrum of the modeled vowel /I/, position 8
Magnitude (dB)
40
20
0
-20
-40
0
1000
2000
3000
4000
5000
6000
Frequency (Hz)


Spectrum of the modeled vowel /I/, position 9
Spectrum of the modeled vowel /I/, position 10
Spectrum of the modeled vowel /I/, position 11
Spectrum of the modeled vowel /I/, position 12
Spectrum of the modeled vowel /I/, position 13
Spectrum of the modeled vowel /I/, position 14
Spectrum of the modeled vowel /I/, position 15
Spectrum of the modeled vowel /I/, position 16
Magnitude (dB)
40
20
0
-20
-40
Frequency (Hz)
0
1000
2000
3000
4000
5000
6000


Spectrum of the modeled vowel /ε/, position 0
Spectrum of the modeled vowel /ε/, position 2
Spectrum of the modeled vowel /ε/, position 3
Spectrum of the modeled vowel /ε/, position 4
Spectrum of the modeled vowel /ε/, position 5
Spectrum of the modeled vowel /ε/, position 6
Spectrum of the modeled vowel /ε/, position 7
Spectrum of the modeled vowel /ε/, position 8
Magnitude (dB)
40
20
0
-20
-40
0
1000
2000
3000
4000
5000
6000
Frequency (Hz)



Spectrum of the modeled vowel /ε/, position 9

Spectrum of the modeled vowel /ε/, position 10

Spectrum of the modeled vowel /ε/, position 11

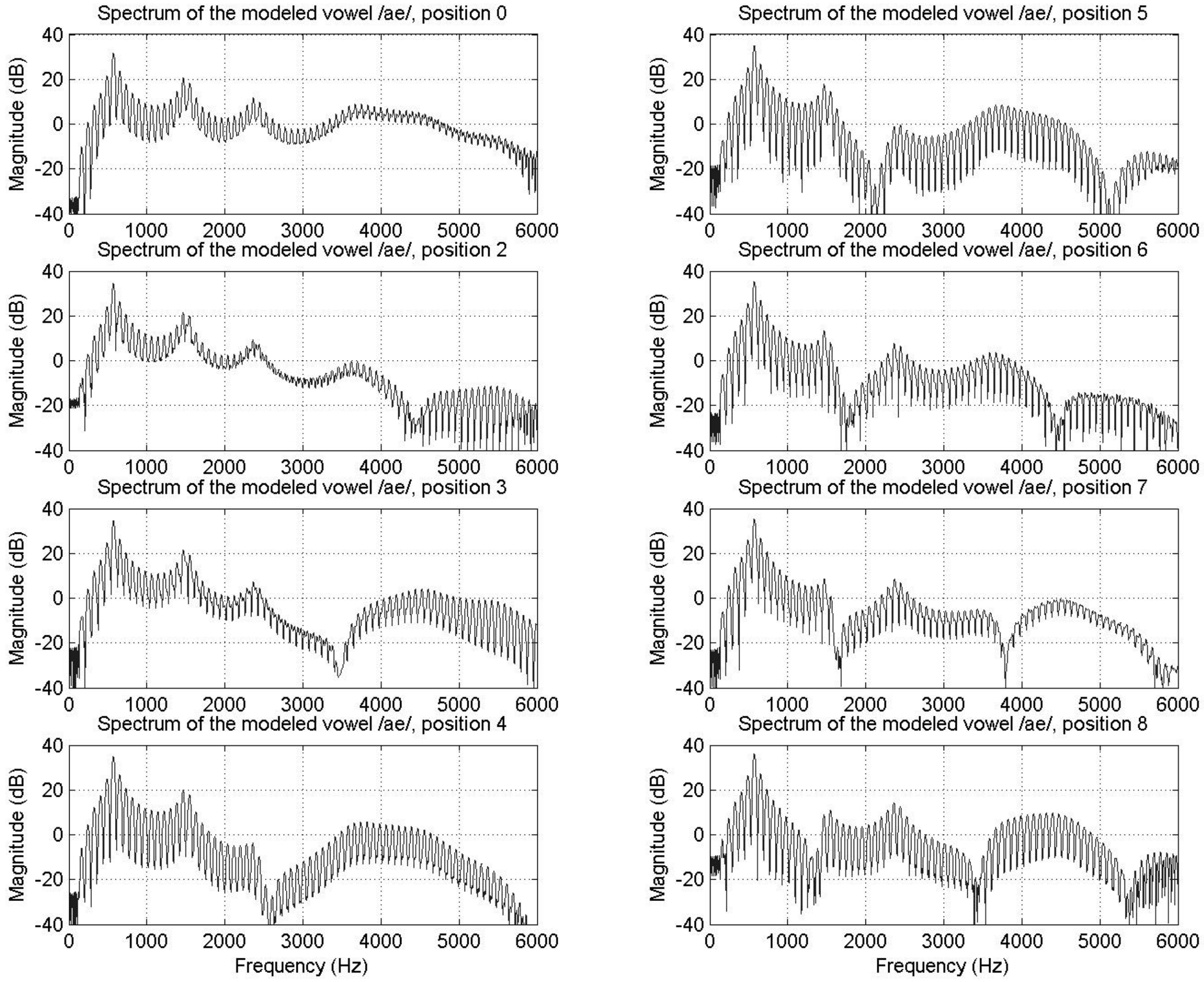
Spectrum of the modeled vowel /ε/, position 12

Spectrum of the modeled vowel /ε/, position 13

Spectrum of the modeled vowel /ε/, position 14

Spectrum of the modeled vowel /ε/, position 15

Spectrum of the modeled vowel /ε/, position 16



Spectrum of the modeled vowel /ae/, position 0

Spectrum of the modeled vowel /ae/, position 2

Spectrum of the modeled vowel /ae/, position 3

Spectrum of the modeled vowel /ae/, position 4

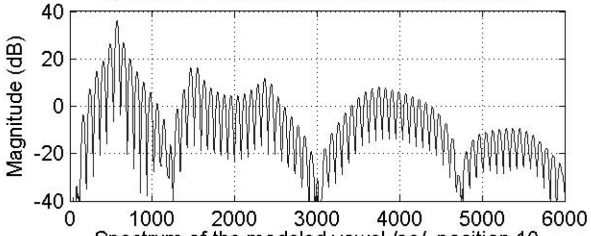
Spectrum of the modeled vowel /ae/, position 5

Spectrum of the modeled vowel /ae/, position 6

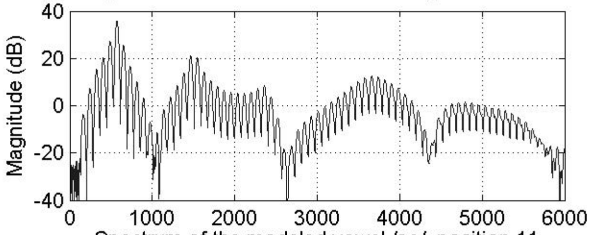
Spectrum of the modeled vowel /ae/, position 7

Spectrum of the modeled vowel /ae/, position 8

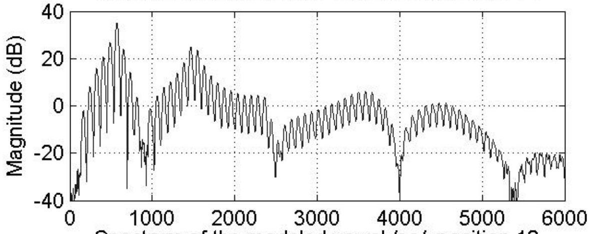
Spectrum of the modeled vowel /ae/, position 9



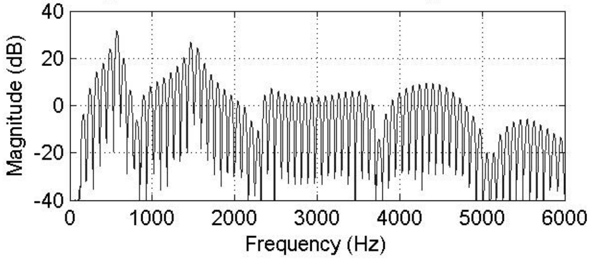
Spectrum of the modeled vowel /ae/, position 10



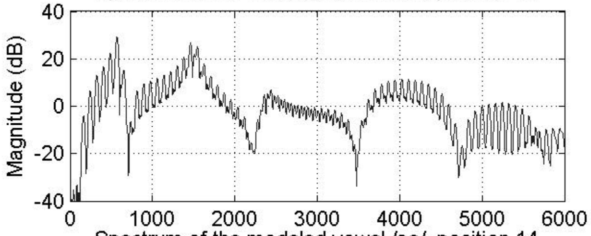
Spectrum of the modeled vowel /ae/, position 11



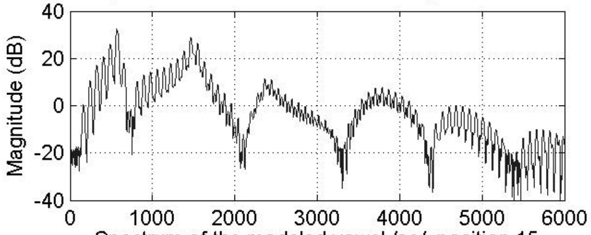
Spectrum of the modeled vowel /ae/, position 12



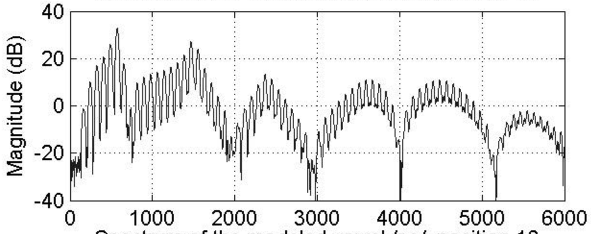
Spectrum of the modeled vowel /ae/, position 13



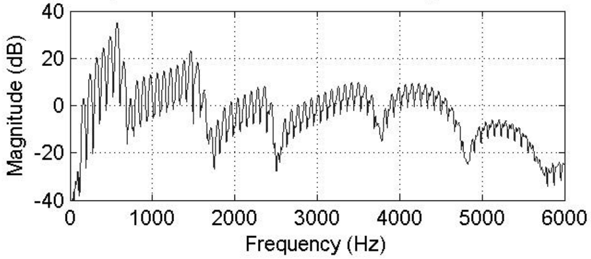
Spectrum of the modeled vowel /ae/, position 14

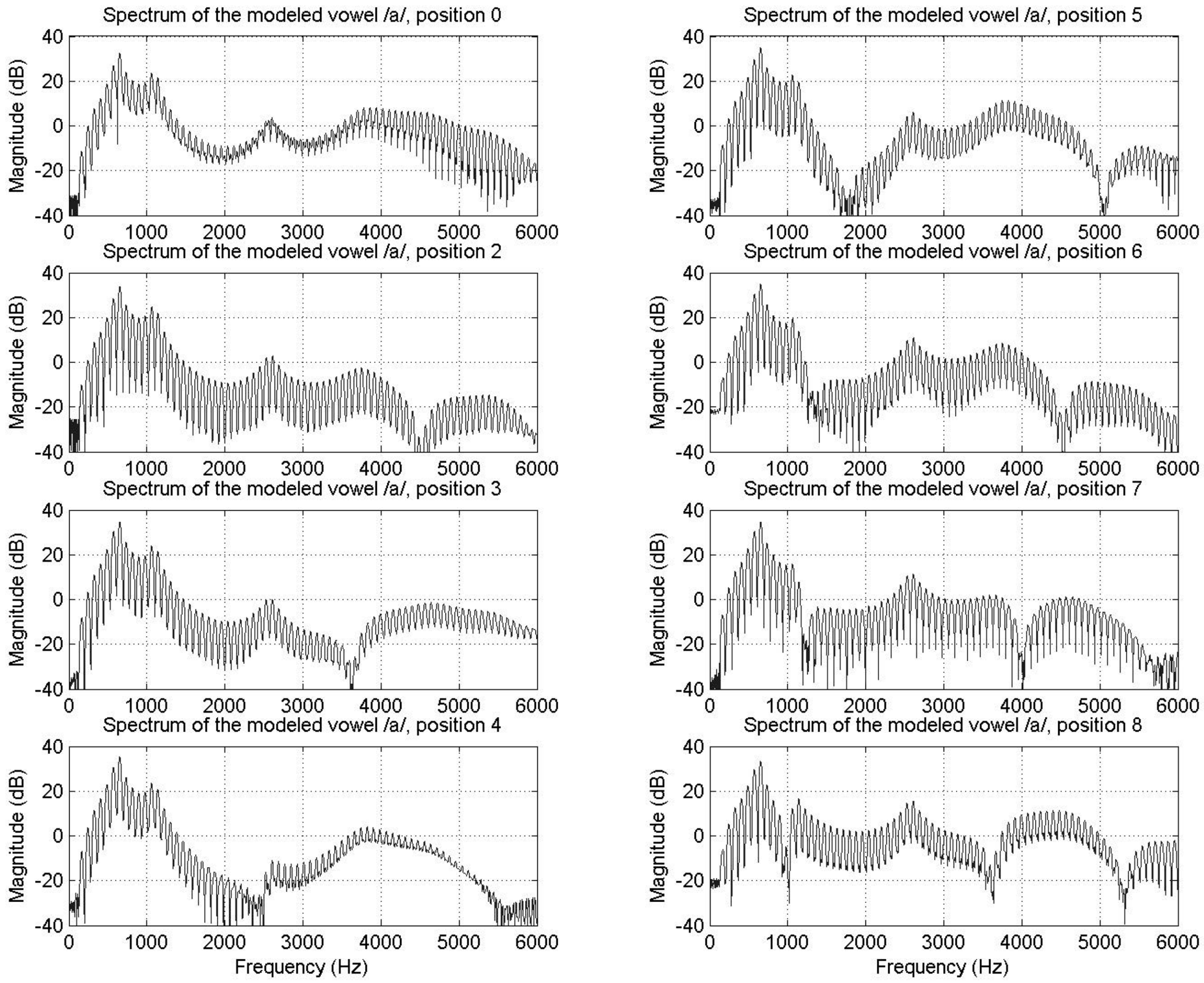


Spectrum of the modeled vowel /ae/, position 15



Spectrum of the modeled vowel /ae/, position 16





Spectrum of the modeled vowel /a/, position 0

Spectrum of the modeled vowel /a/, position 2

Spectrum of the modeled vowel /a/, position 3

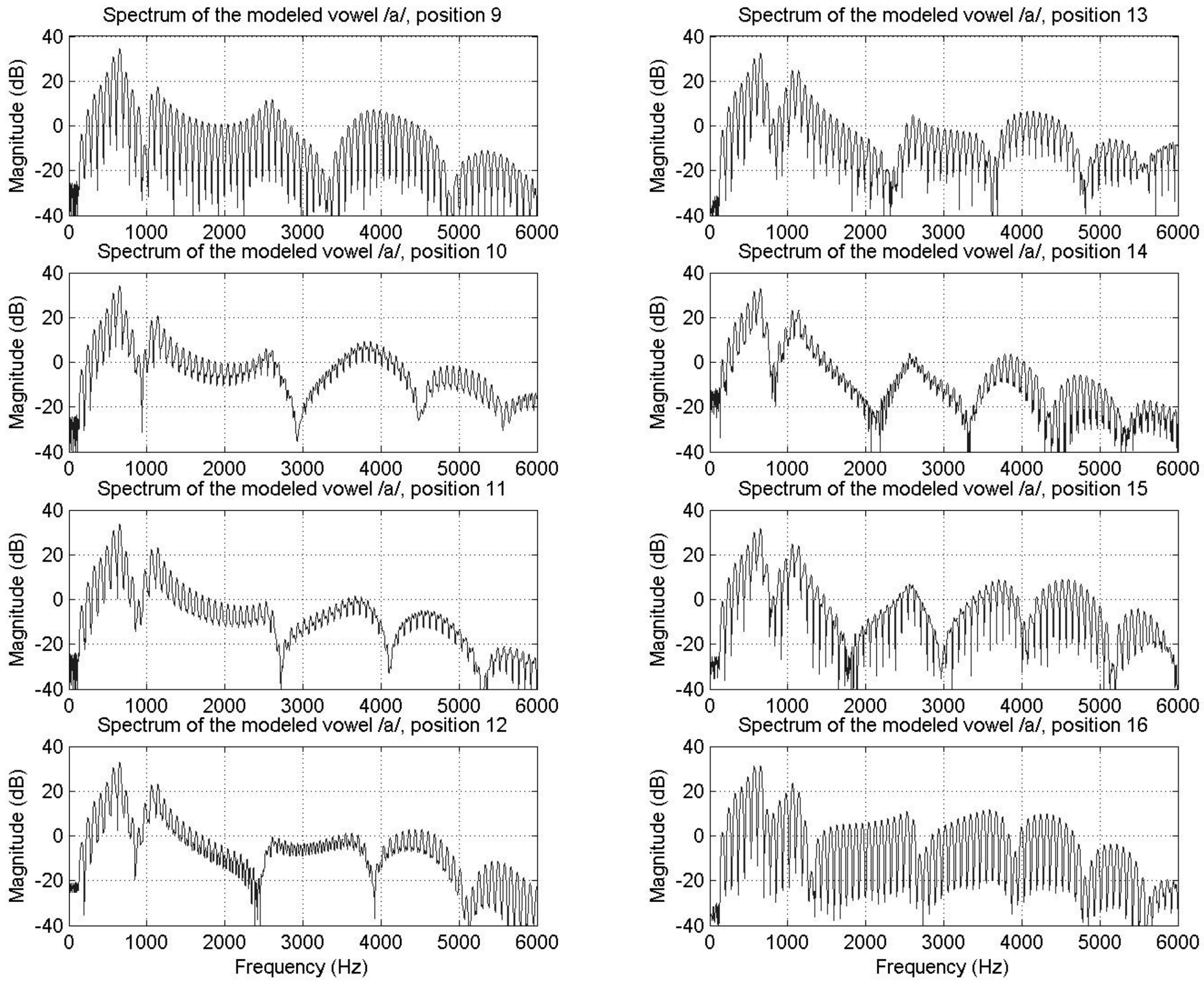
Spectrum of the modeled vowel /a/, position 4

Spectrum of the modeled vowel /a/, position 5

Spectrum of the modeled vowel /a/, position 6

Spectrum of the modeled vowel /a/, position 7

Spectrum of the modeled vowel /a/, position 8


Spectrum of the modeled vowel /a/, position 9

Spectrum of the modeled vowel /a/, position 10

Spectrum of the modeled vowel /a/, position 11

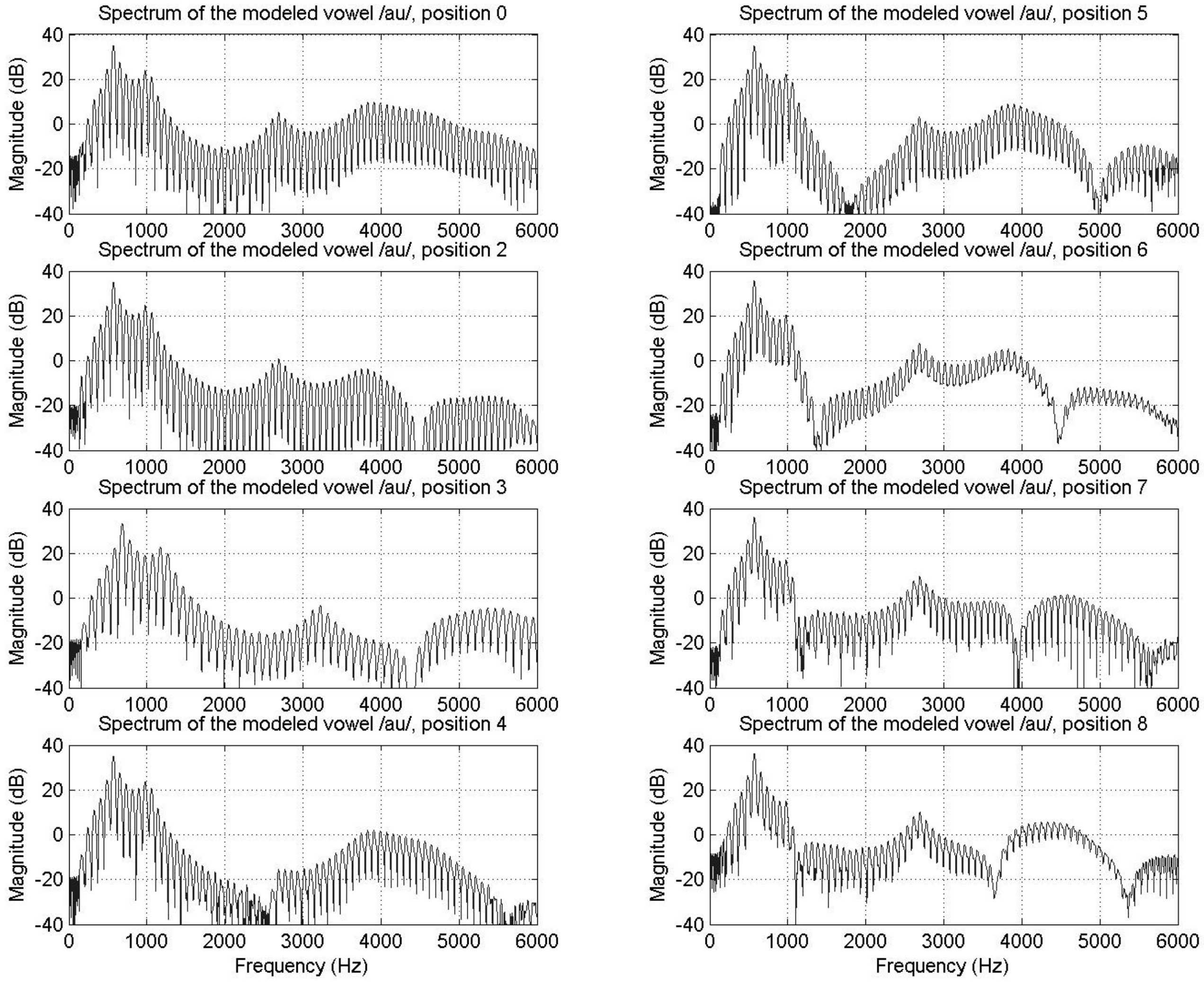
Spectrum of the modeled vowel /a/, position 12

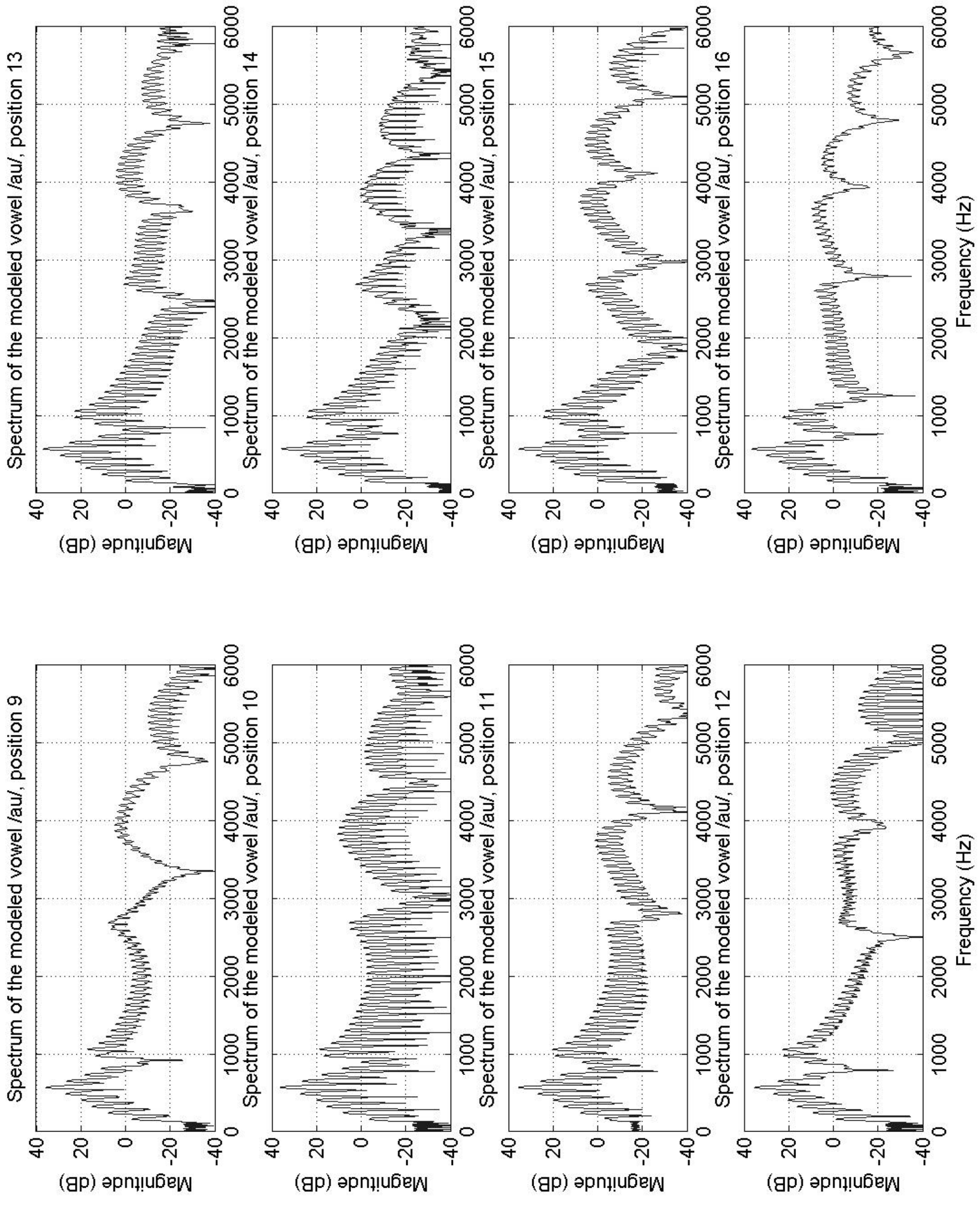
Spectrum of the modeled vowel /a/, position 13

Spectrum of the modeled vowel /a/, position 14

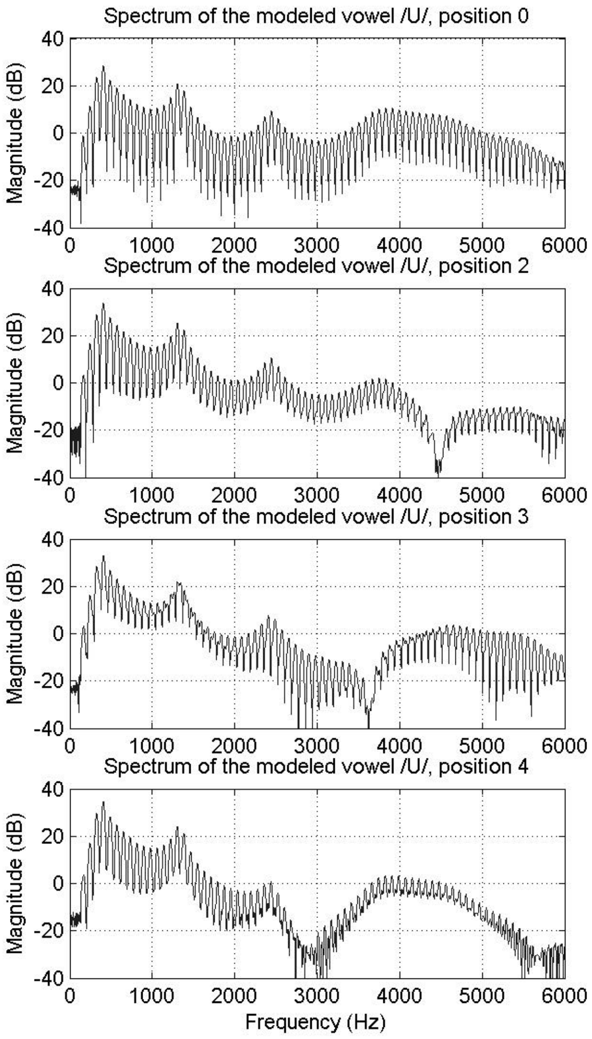
Spectrum of the modeled vowel /a/, position 15

Spectrum of the modeled vowel /a/, position 16


Spectrum of the modeled vowel /au/, position 0
Spectrum of the modeled vowel /au/, position 2
Spectrum of the modeled vowel /au/, position 3
Spectrum of the modeled vowel /au/, position 4
Spectrum of the modeled vowel /au/, position 5
Spectrum of the modeled vowel /au/, position 6
Spectrum of the modeled vowel /au/, position 7
Spectrum of the modeled vowel /au/, position 8
Magnitude (dB)
40
20
0
-20
-40
0
1000
2000
3000
4000
5000
6000
Frequency (Hz)



Spectrum of the modeled vowel /au/, position 9
Spectrum of the modeled vowel /au/, position 10
Spectrum of the modeled vowel /au/, position 11
Spectrum of the modeled vowel /au/, position 12
Spectrum of the modeled vowel /au/, position 13
Spectrum of the modeled vowel /au/, position 14
Spectrum of the modeled vowel /au/, position 15
Spectrum of the modeled vowel /au/, position 16
Magnitude (dB)
40
20
0
-20
-40
0
1000
2000
3000
4000
5000
6000
Frequency (Hz)

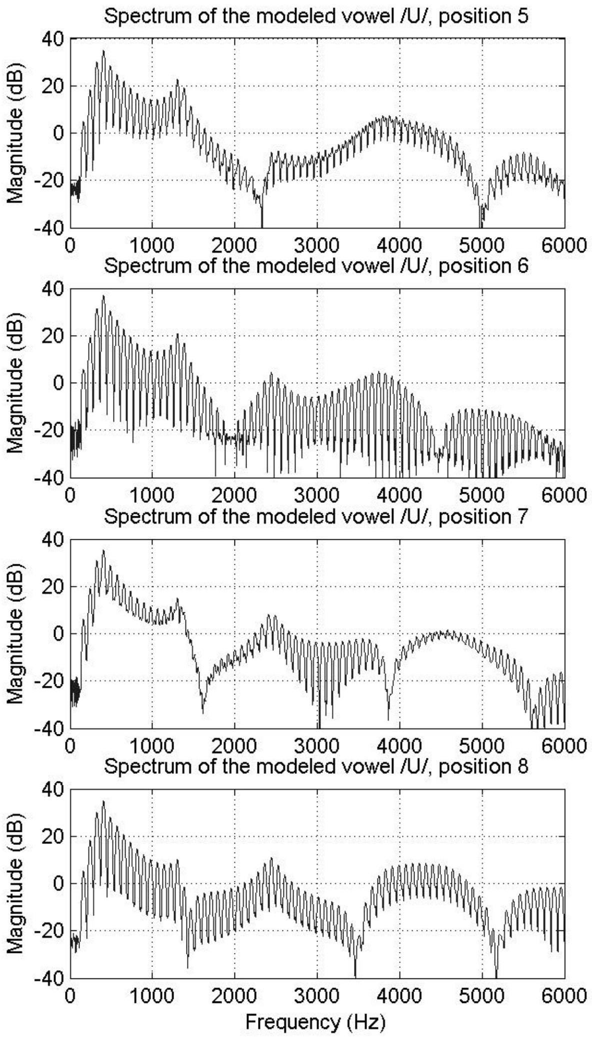


Spectrum of the modeled vowel /U/, position 0

Spectrum of the modeled vowel /U/, position 2

Spectrum of the modeled vowel /U/, position 3

Spectrum of the modeled vowel /U/, position 4

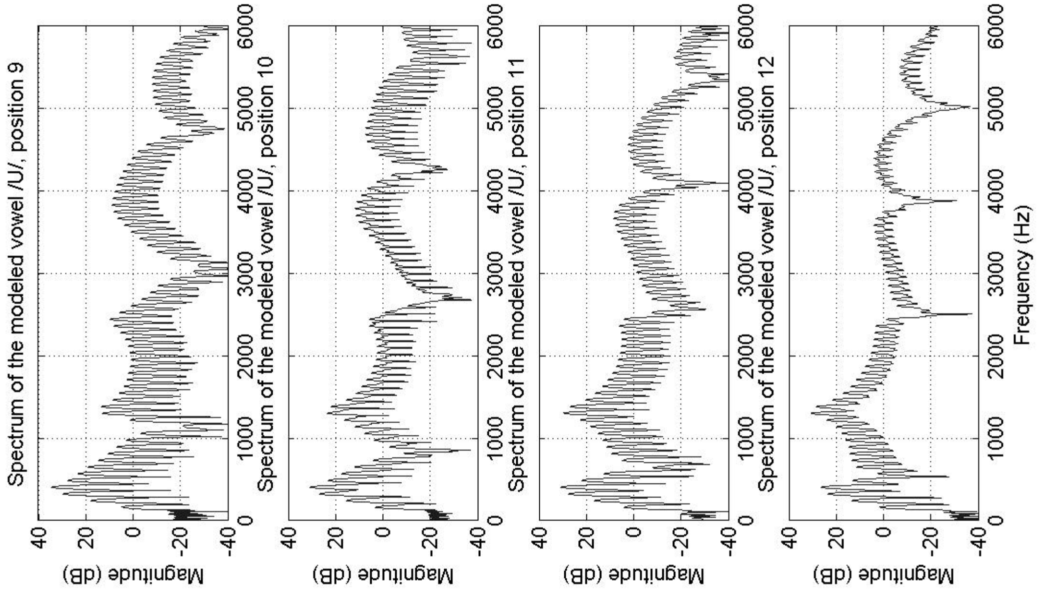


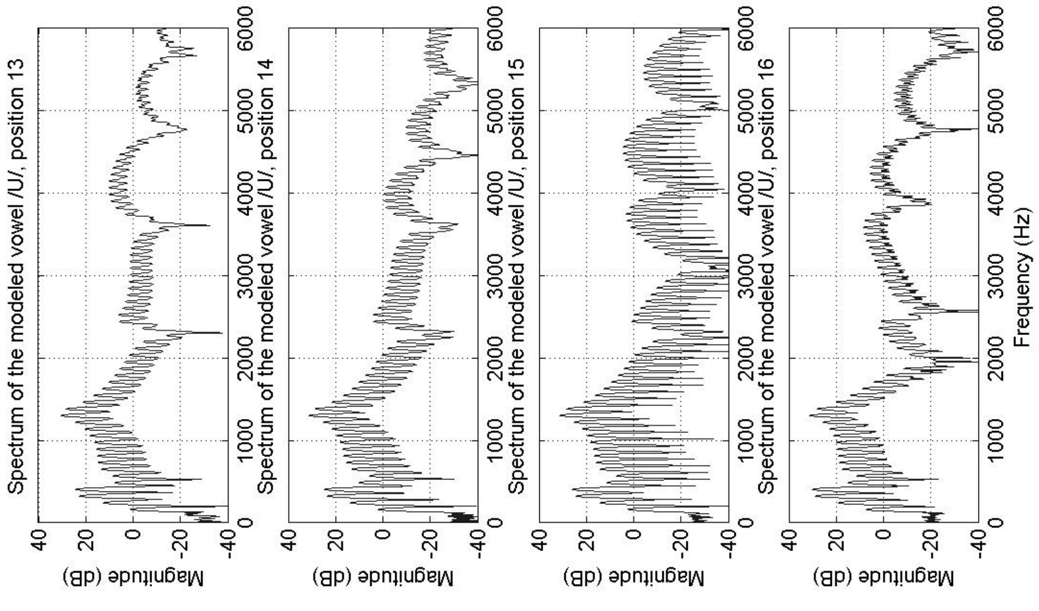
Spectrum of the modeled vowel /U/, position 5

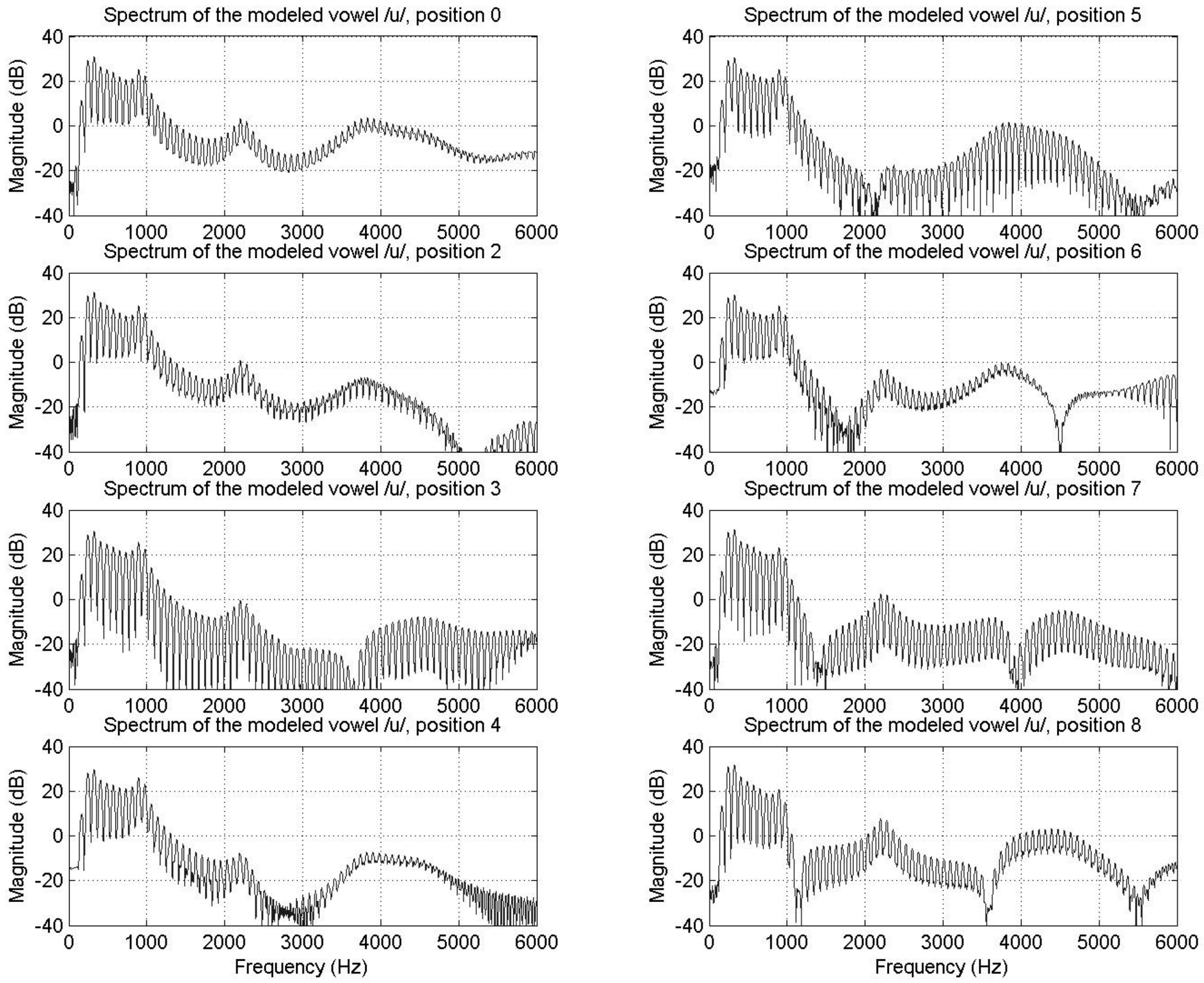
Spectrum of the modeled vowel /U/, position 6

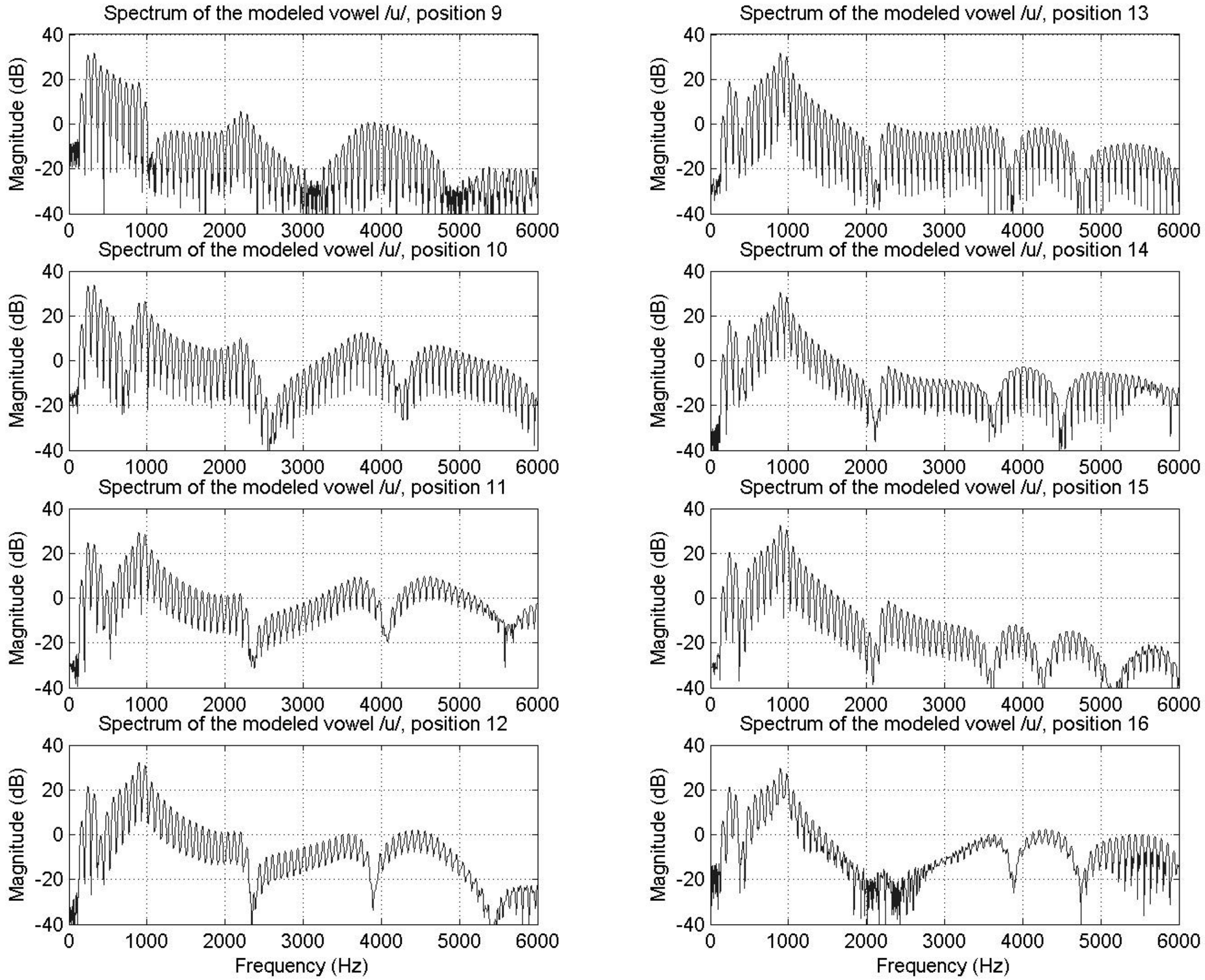
Spectrum of the modeled vowel /U/, position 7

Spectrum of the modeled vowel /U/, position 8


Spectrum of the modeled vowel /U/, position 9
Spectrum of the modeled vowel /U/, position 10
Spectrum of the modeled vowel /U/, position 11
Spectrum of the modeled vowel /U/, position 12
Magnitude (dB)
40
20
0
-20
-40
0
1000
2000
3000
4000
5000
6000
Frequency (Hz)


Spectrum of the modeled vowel /U/, position 13
Spectrum of the modeled vowel /U/, position 14
Spectrum of the modeled vowel /U/, position 15
Spectrum of the modeled vowel /U/, position 16
Magnitude (dB)
40
20
0
-20
-40
0
1000
2000
3000
4000
5000
6000
Frequency (Hz)


Spectrum of the modeled vowel /u/, position 0
Spectrum of the modeled vowel /u/, position 2
Spectrum of the modeled vowel /u/, position 3
Spectrum of the modeled vowel /u/, position 4
Spectrum of the modeled vowel /u/, position 5
Spectrum of the modeled vowel /u/, position 6
Spectrum of the modeled vowel /u/, position 7
Spectrum of the modeled vowel /u/, position 8
Magnitude (dB)
40
20
0
-20
-40
0
1000
2000
3000
4000
5000
6000
Frequency (Hz)


Spectrum of the modeled vowel /u/, position 9

Spectrum of the modeled vowel /u/, position 10

Spectrum of the modeled vowel /u/, position 11

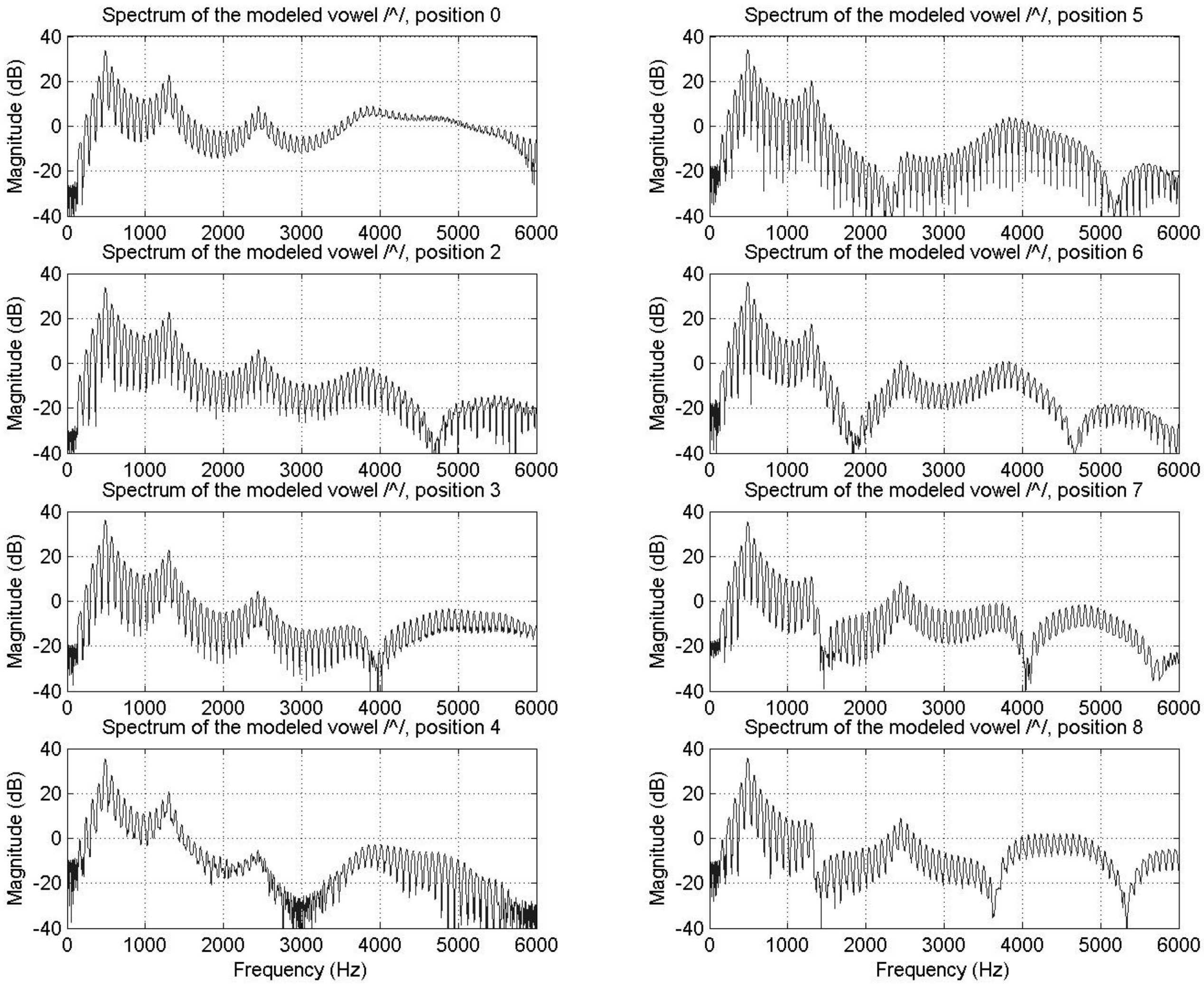
Spectrum of the modeled vowel /u/, position 12

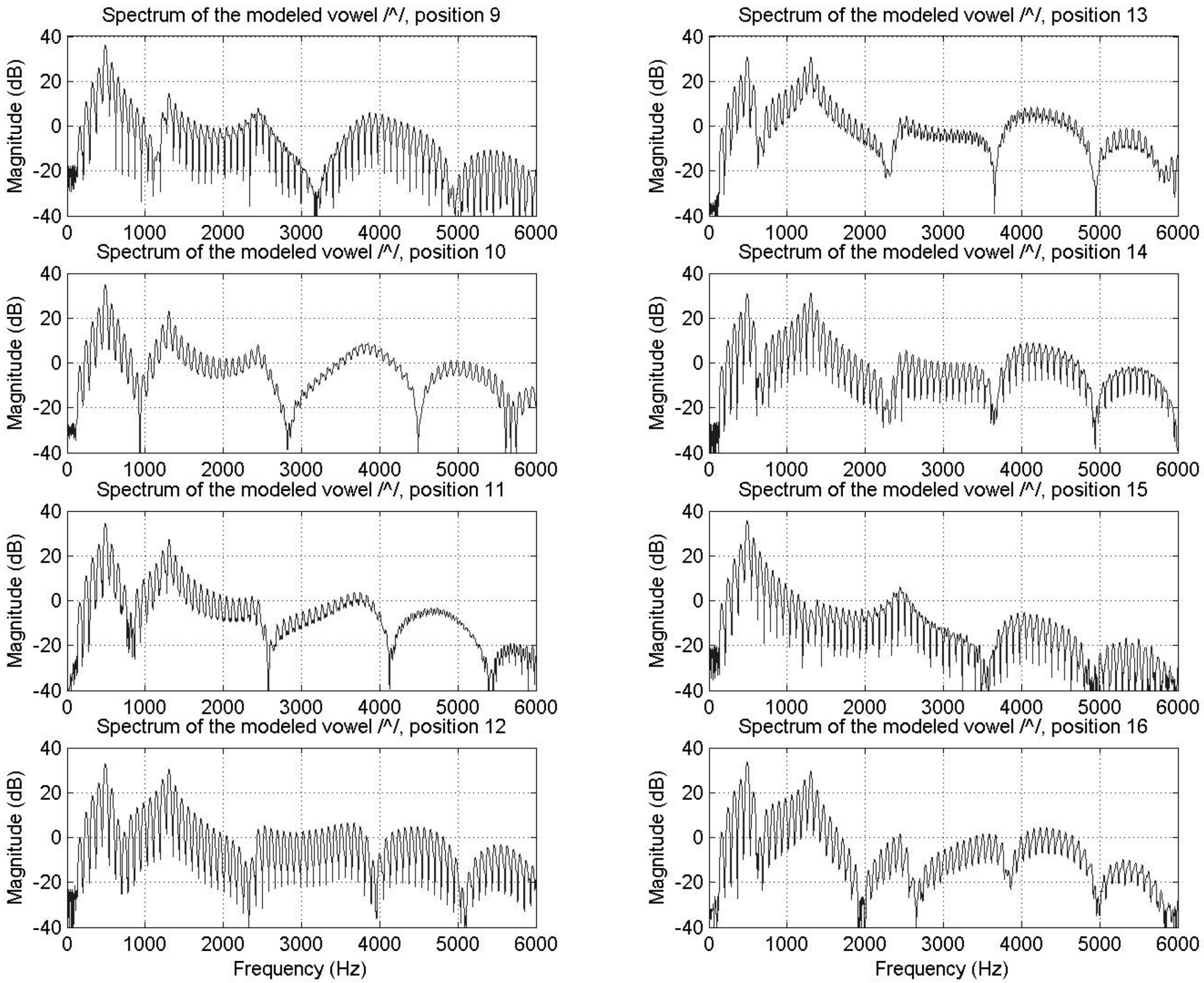
Spectrum of the modeled vowel /u/, position 13

Spectrum of the modeled vowel /u/, position 14

Spectrum of the modeled vowel /u/, position 15

Spectrum of the modeled vowel /u/, position 16


Spectrum of the modeled vowel /^/, position 0
Spectrum of the modeled vowel /^/, position 2
Spectrum of the modeled vowel /^/, position 3
Spectrum of the modeled vowel /^/, position 4
Spectrum of the modeled vowel /^/, position 5
Spectrum of the modeled vowel /^/, position 6
Spectrum of the modeled vowel /^/, position 7
Spectrum of the modeled vowel /^/, position 8
Magnitude (dB)
40
20
0
-20
-40
0
1000
2000
3000
4000
5000
6000
Frequency (Hz)



Spectrum of the modeled vowel /^/, position 9

Spectrum of the modeled vowel /^/, position 10

Spectrum of the modeled vowel /^/, position 11

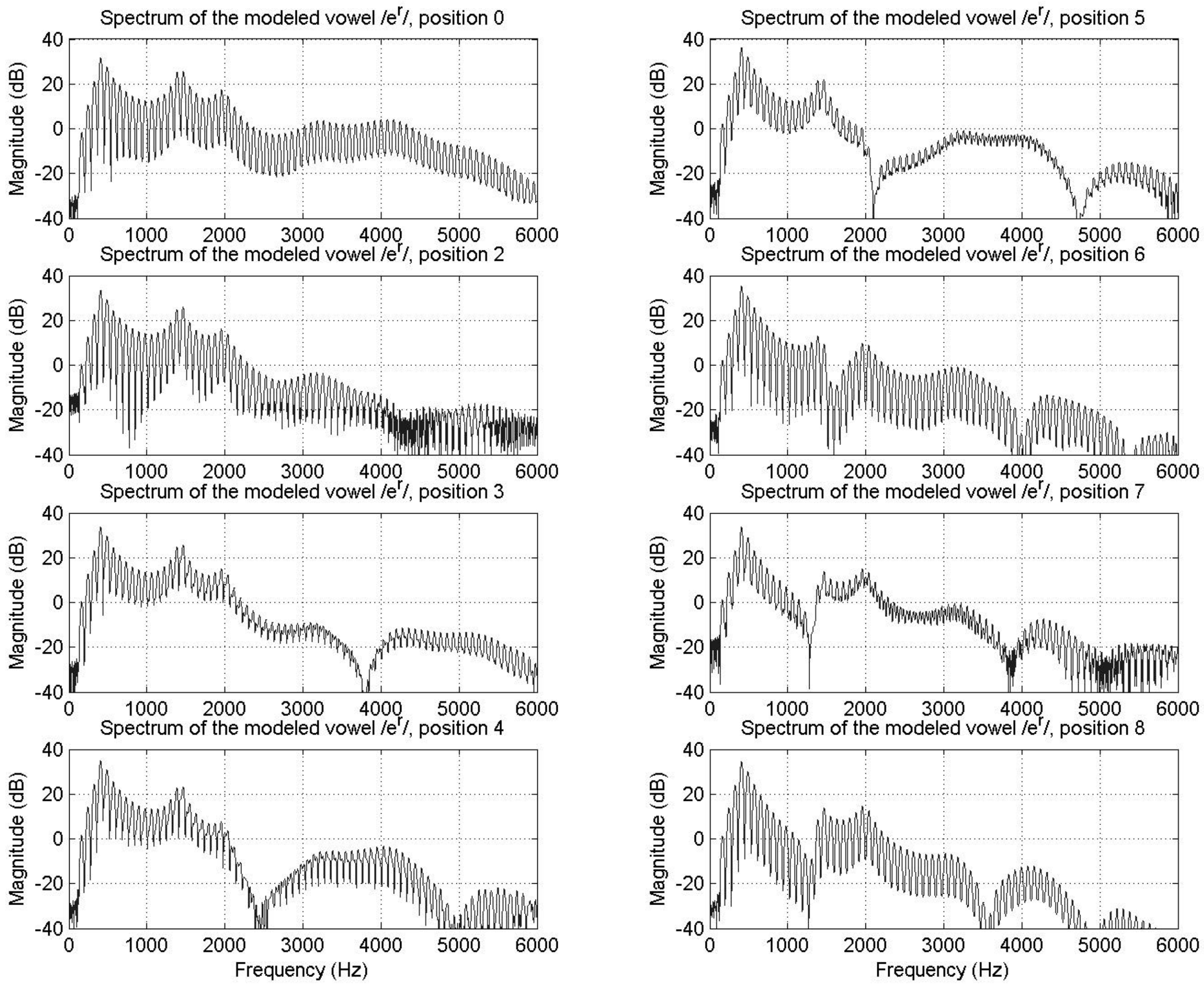
Spectrum of the modeled vowel /^/, position 12

Spectrum of the modeled vowel /^/, position 13

Spectrum of the modeled vowel /^/, position 14

Spectrum of the modeled vowel /^/, position 15

Spectrum of the modeled vowel /^/, position 16



Spectrum of the modeled vowel /e^r/, position 0

Spectrum of the modeled vowel /e^r/, position 2

Spectrum of the modeled vowel /e^r/, position 3

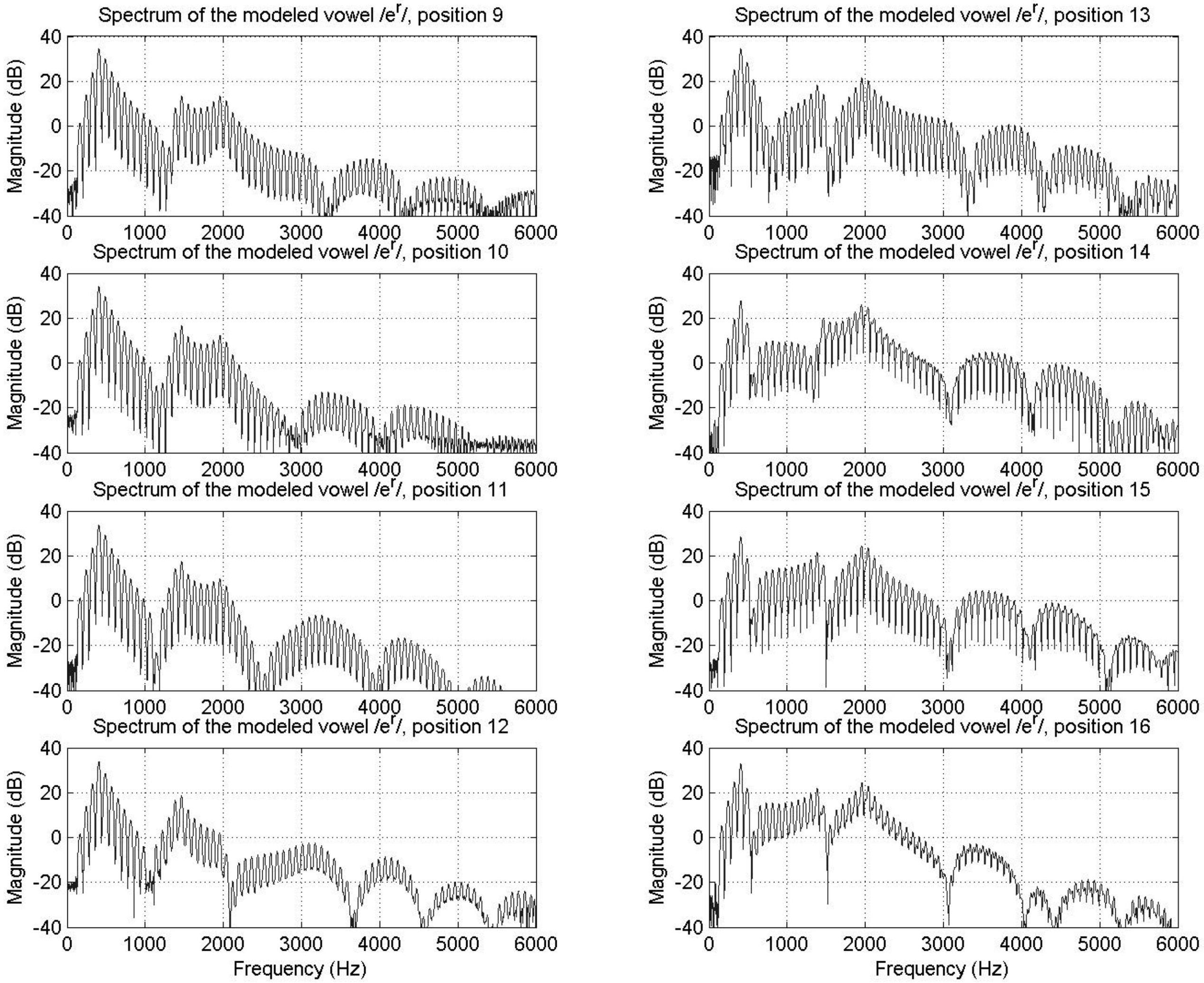
Spectrum of the modeled vowel /e^r/, position 4

Spectrum of the modeled vowel /e^r/, position 5

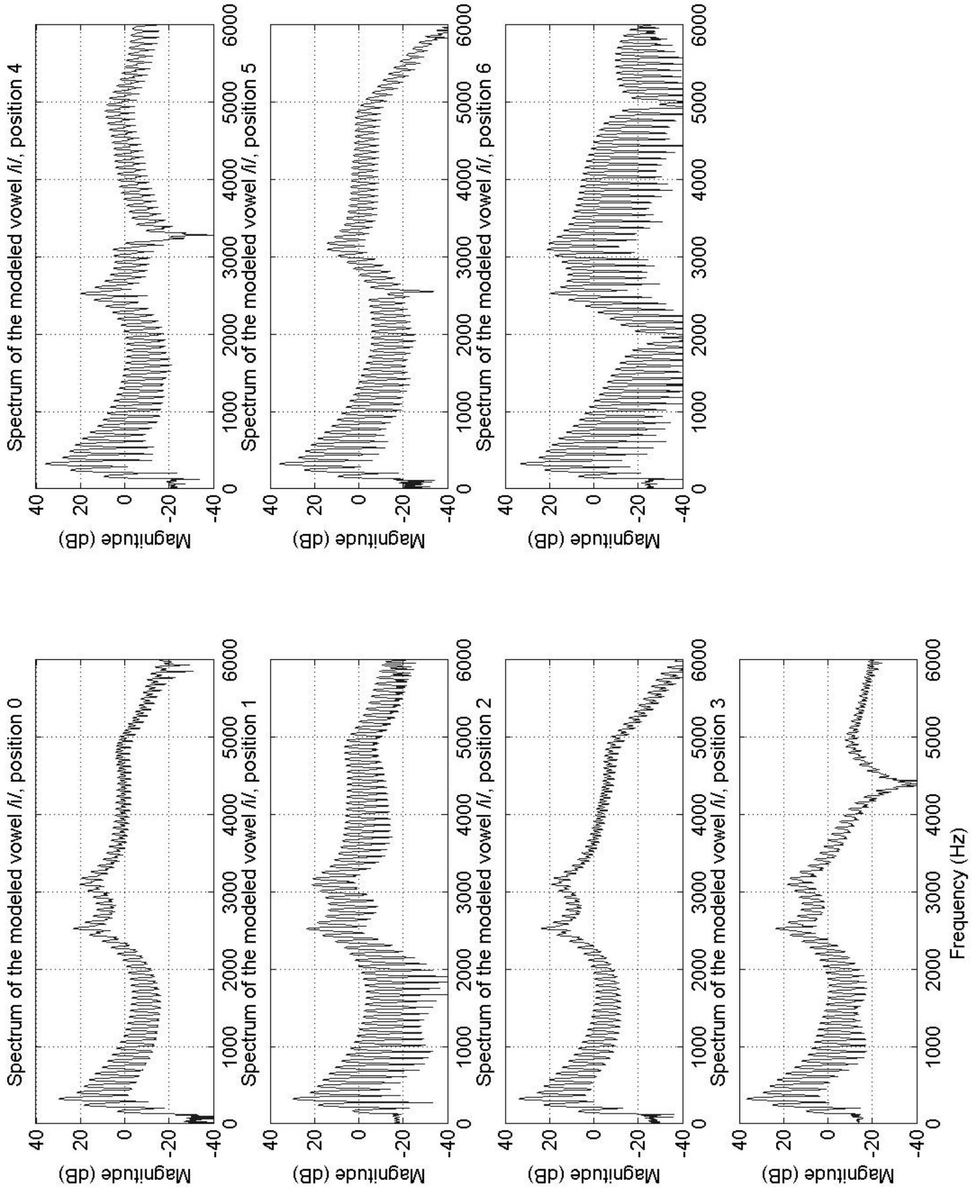
Spectrum of the modeled vowel /e^r/, position 6

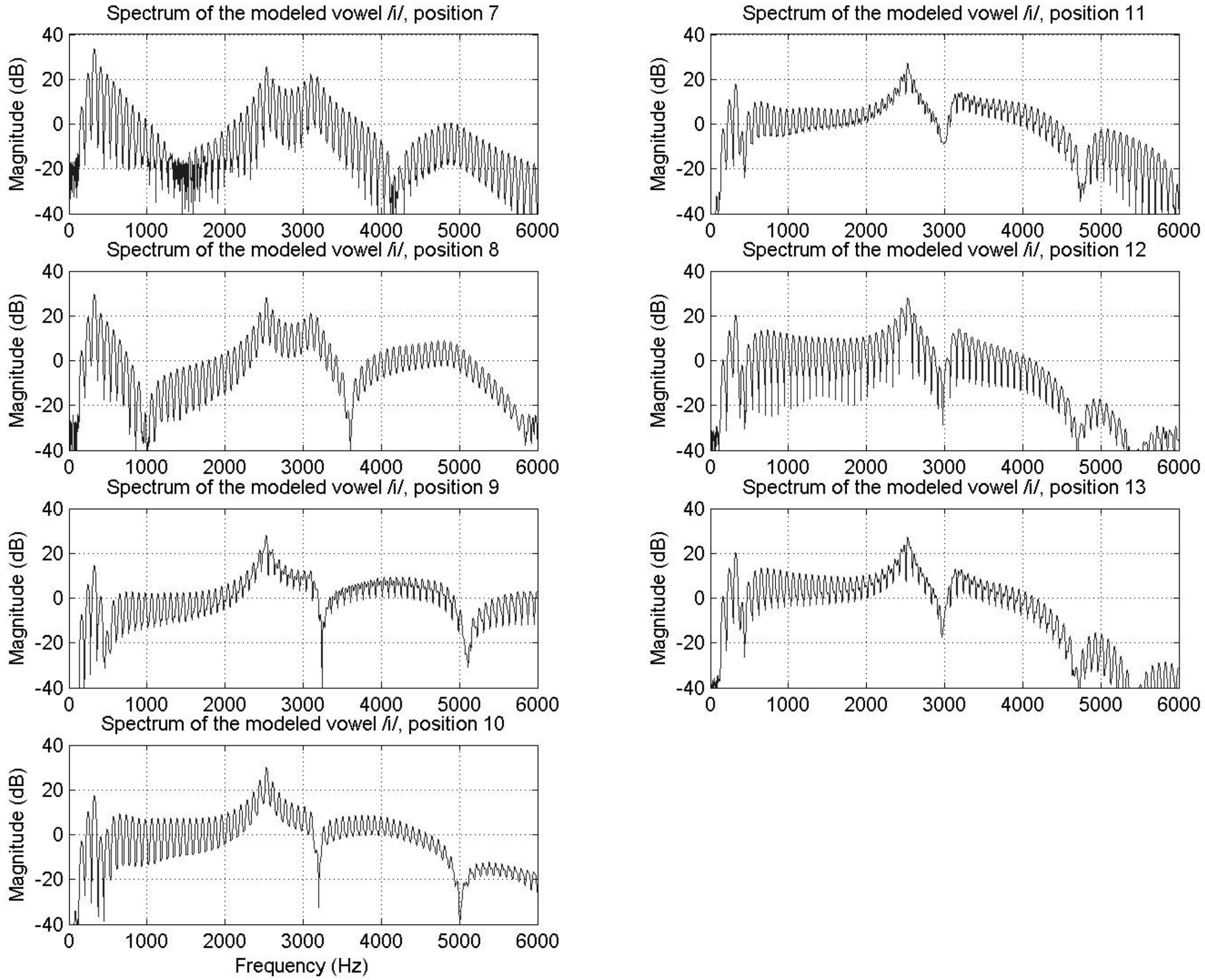
Spectrum of the modeled vowel /e^r/, position 7

Spectrum of the modeled vowel /e^r/, position 8


Spectrum of the modeled vowel /e^r/, position 9
Spectrum of the modeled vowel /e^r/, position 10
Spectrum of the modeled vowel /e^r/, position 11
Spectrum of the modeled vowel /e^r/, position 12
Spectrum of the modeled vowel /e^r/, position 13
Spectrum of the modeled vowel /e^r/, position 14
Spectrum of the modeled vowel /e^r/, position 15
Spectrum of the modeled vowel /e^r/, position 16
Magnitude (dB)
40
20
0
-20
-40
0
1000
2000
3000
4000
5000
6000
Frequency (Hz)

14.3. Vowel spectra generated from the model of the female vocal tract





Spectrum of the modeled vowel /i/, position 7

Spectrum of the modeled vowel /i/, position 8

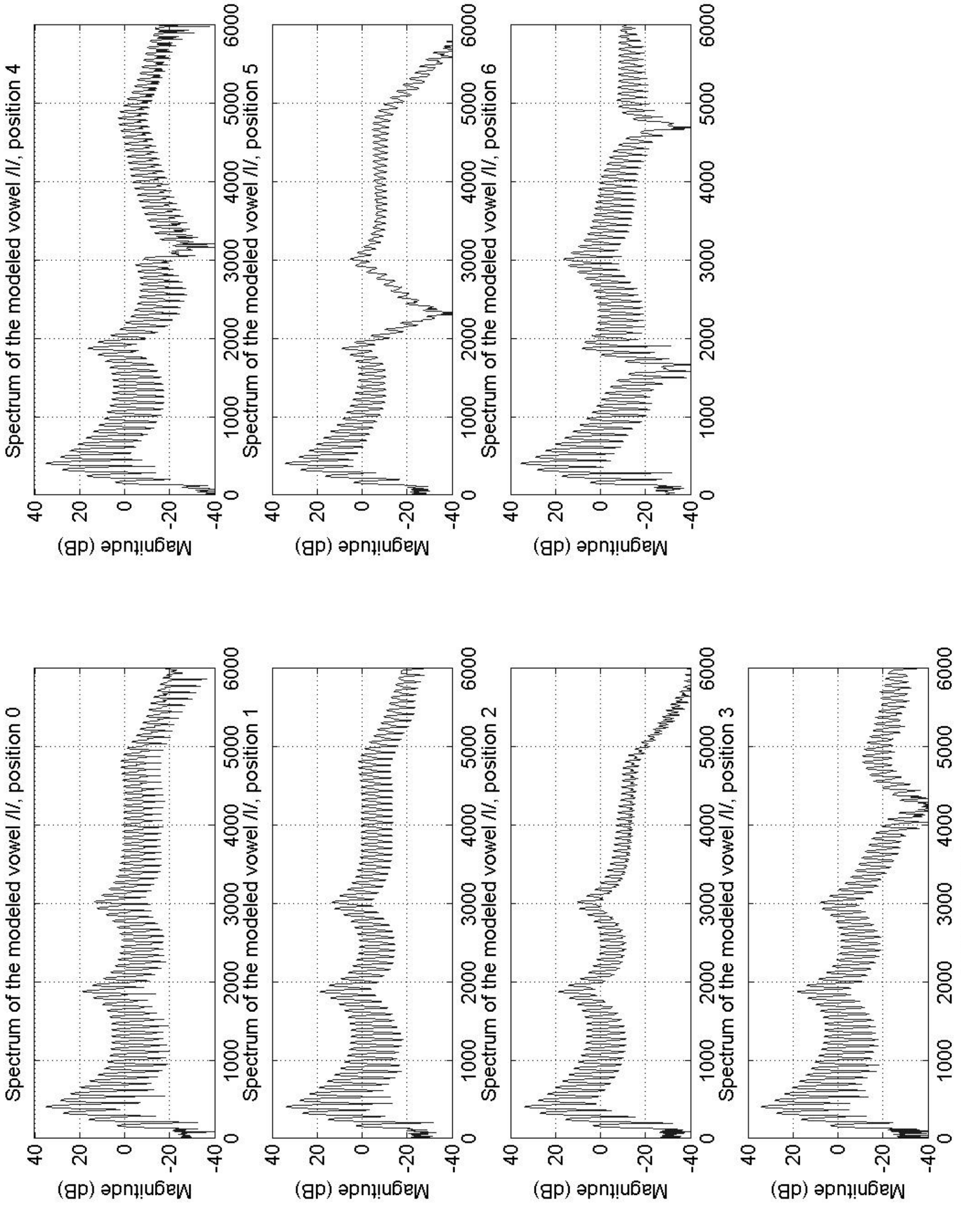
Spectrum of the modeled vowel /i/, position 9

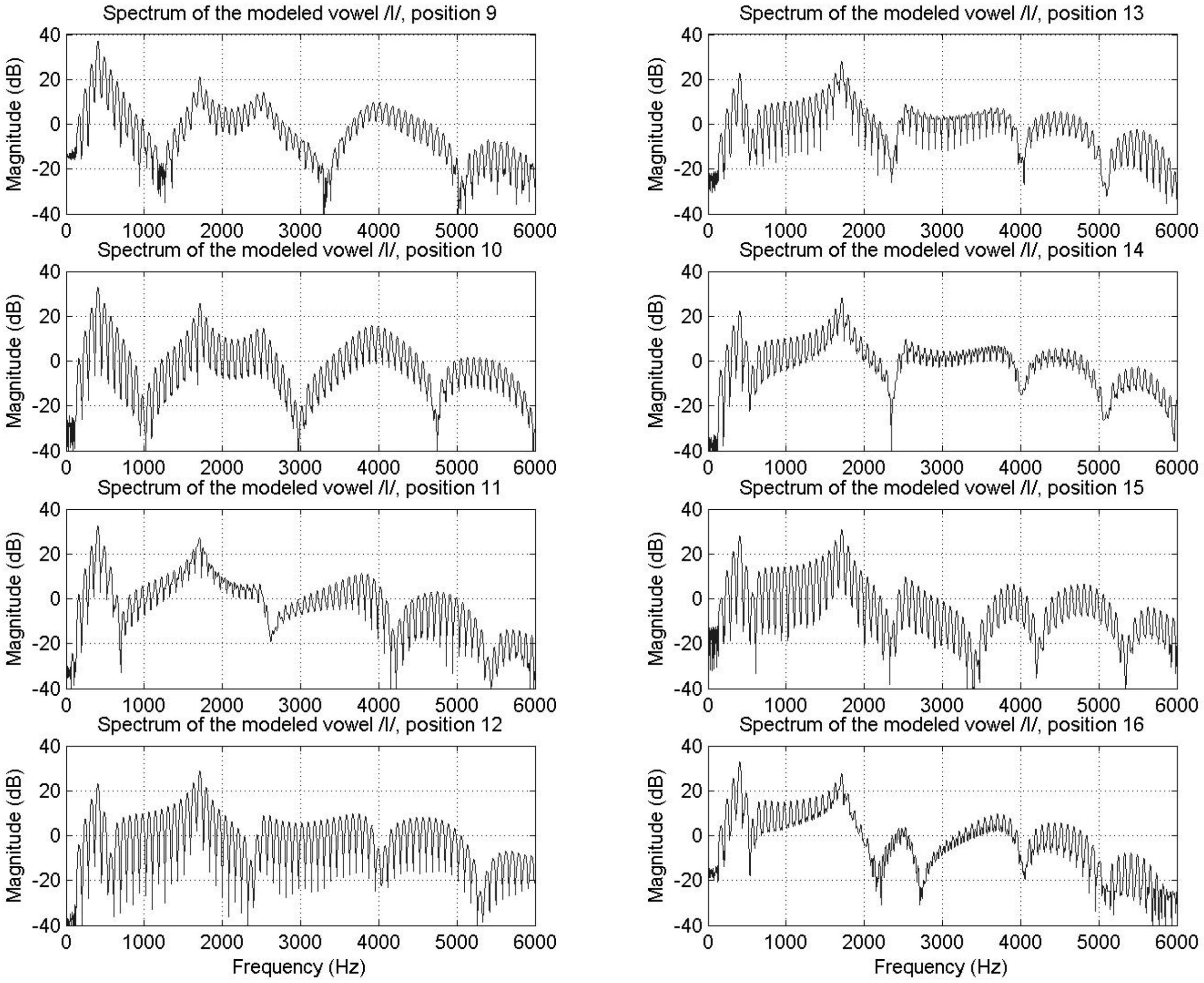
Spectrum of the modeled vowel /i/, position 10

Spectrum of the modeled vowel /i/, position 11

Spectrum of the modeled vowel /i/, position 12

Spectrum of the modeled vowel /i/, position 13


Spectrum of the modeled vowel /I/, position 0
Spectrum of the modeled vowel /I/, position 1
Spectrum of the modeled vowel /I/, position 2
Spectrum of the modeled vowel /I/, position 3
Spectrum of the modeled vowel /I/, position 4
Spectrum of the modeled vowel /I/, position 5
Spectrum of the modeled vowel /I/, position 6
Magnitude (dB)
40
20
0
-20
-40
0
1000
2000
3000
4000
5000
6000
Frequency (Hz)



Spectrum of the modeled vowel /l/, position 9

Spectrum of the modeled vowel /l/, position 10

Spectrum of the modeled vowel /l/, position 11

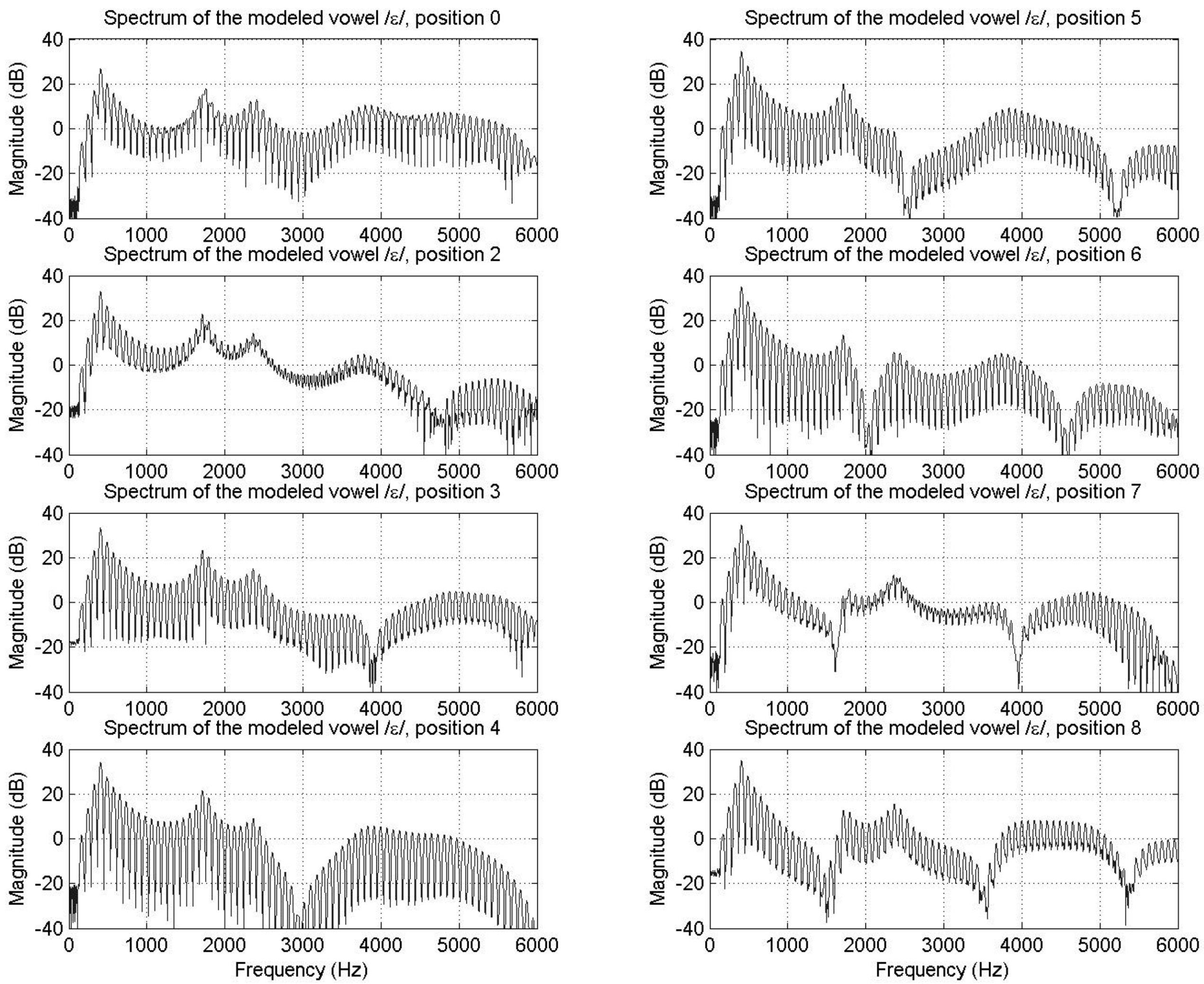
Spectrum of the modeled vowel /l/, position 12

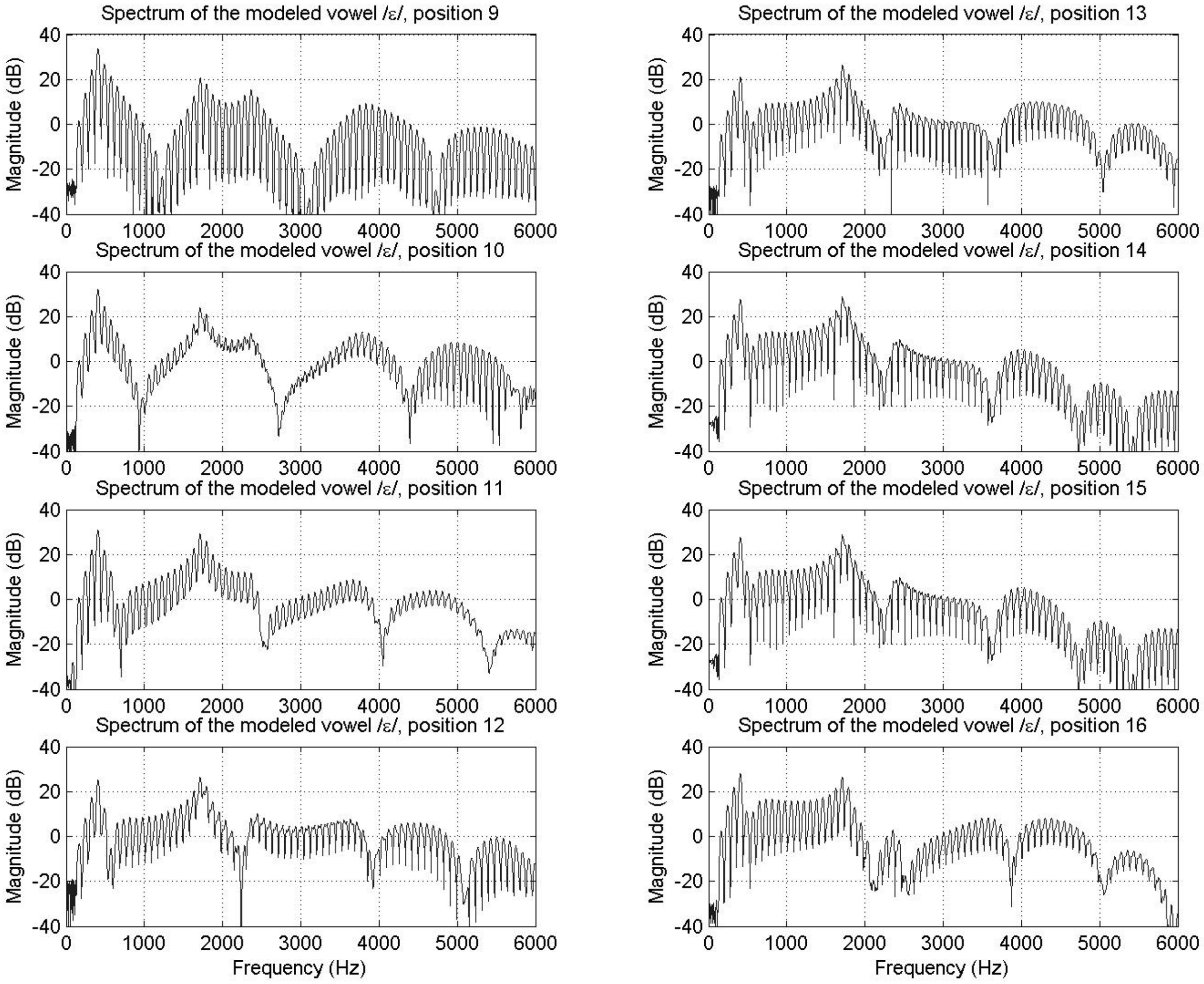
Spectrum of the modeled vowel /l/, position 13

Spectrum of the modeled vowel /l/, position 14

Spectrum of the modeled vowel /l/, position 15

Spectrum of the modeled vowel /l/, position 16


Spectrum of the modeled vowel /ε/, position 0
Spectrum of the modeled vowel /ε/, position 2
Spectrum of the modeled vowel /ε/, position 3
Spectrum of the modeled vowel /ε/, position 4
Spectrum of the modeled vowel /ε/, position 5
Spectrum of the modeled vowel /ε/, position 6
Spectrum of the modeled vowel /ε/, position 7
Spectrum of the modeled vowel /ε/, position 8
Magnitude (dB)
40
20
0
-20
-40
0
1000
2000
3000
4000
5000
6000
Frequency (Hz)



Spectrum of the modeled vowel /ε/, position 9

Spectrum of the modeled vowel /ε/, position 10

Spectrum of the modeled vowel /ε/, position 11

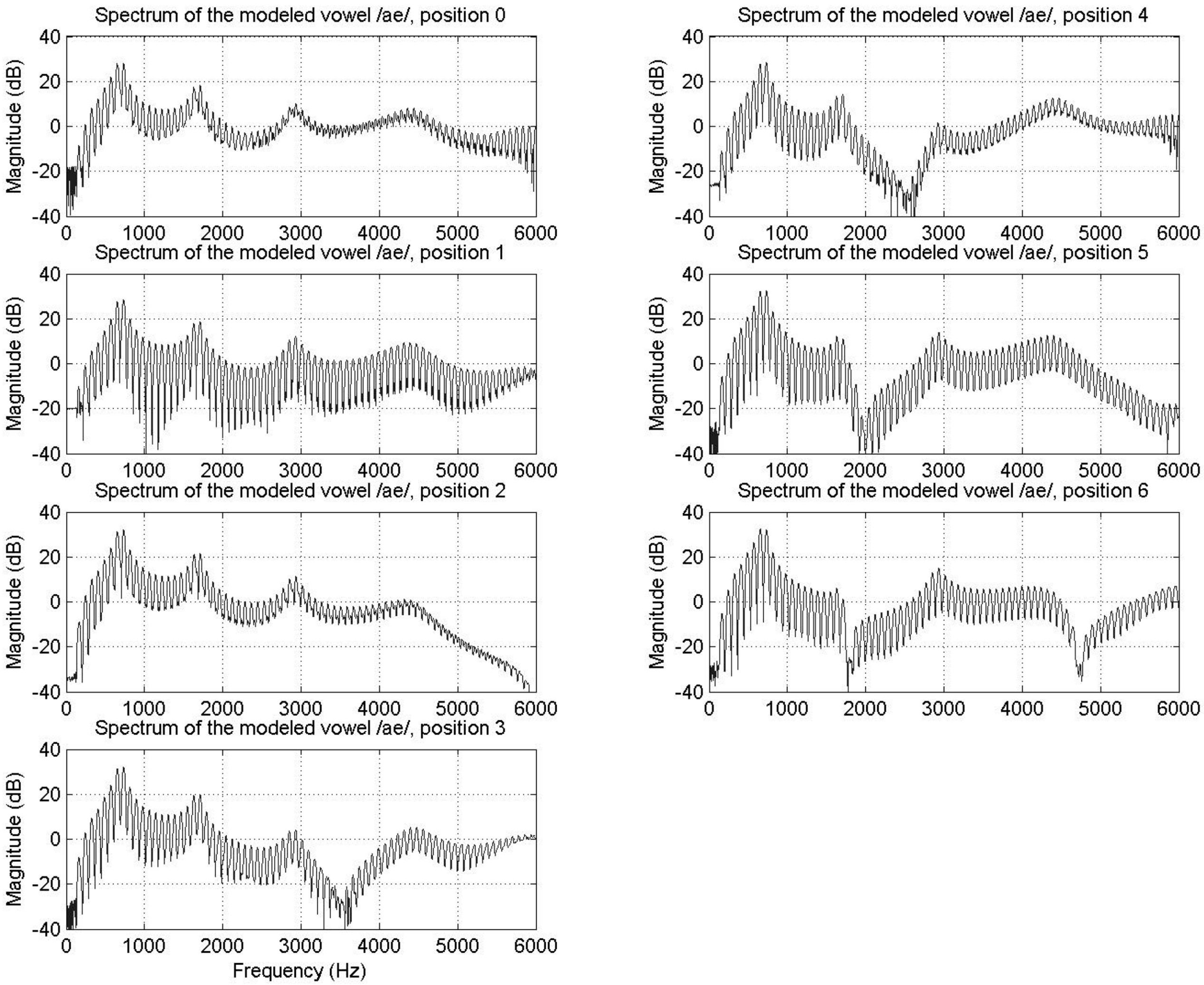
Spectrum of the modeled vowel /ε/, position 12

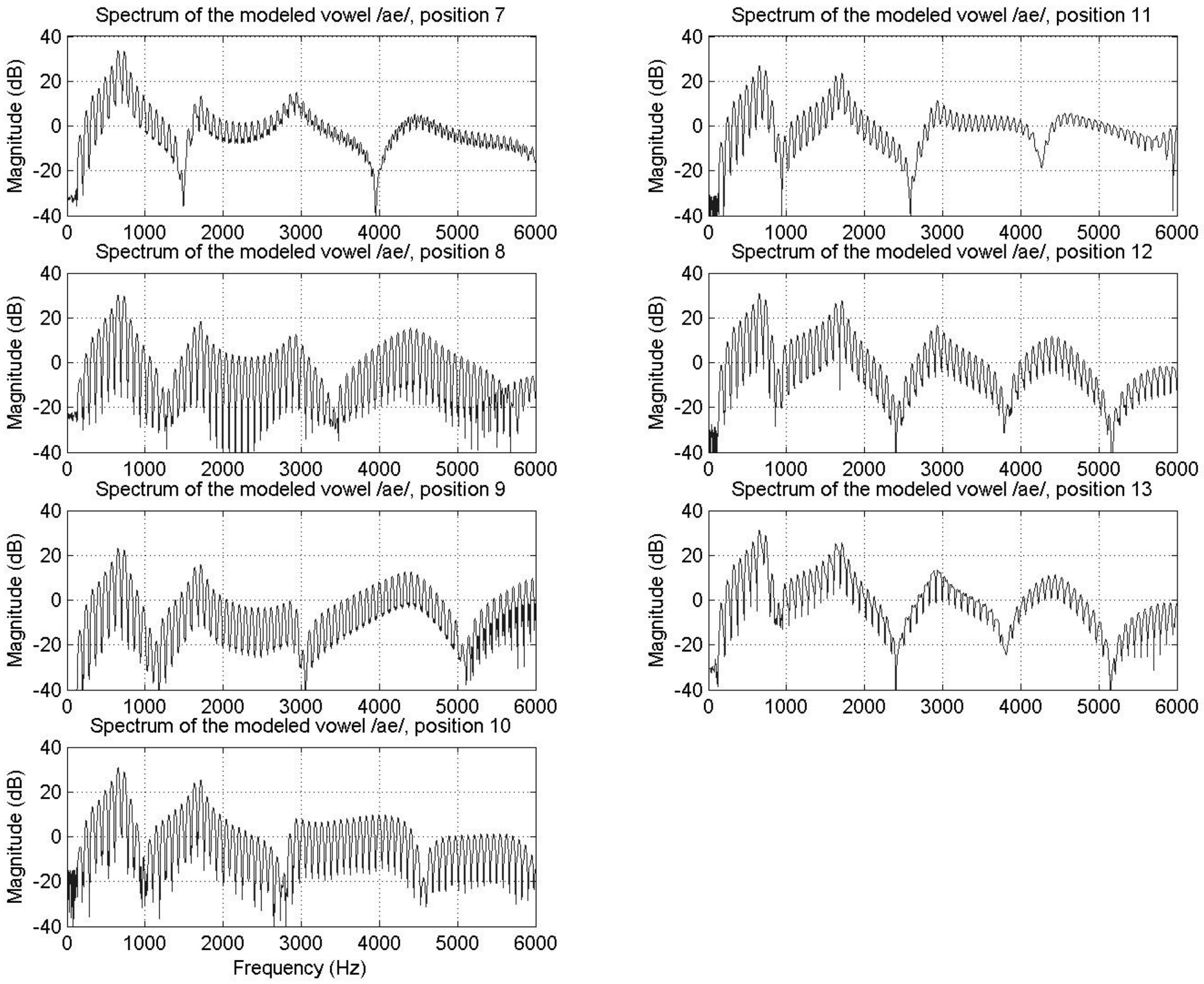
Spectrum of the modeled vowel /ε/, position 13

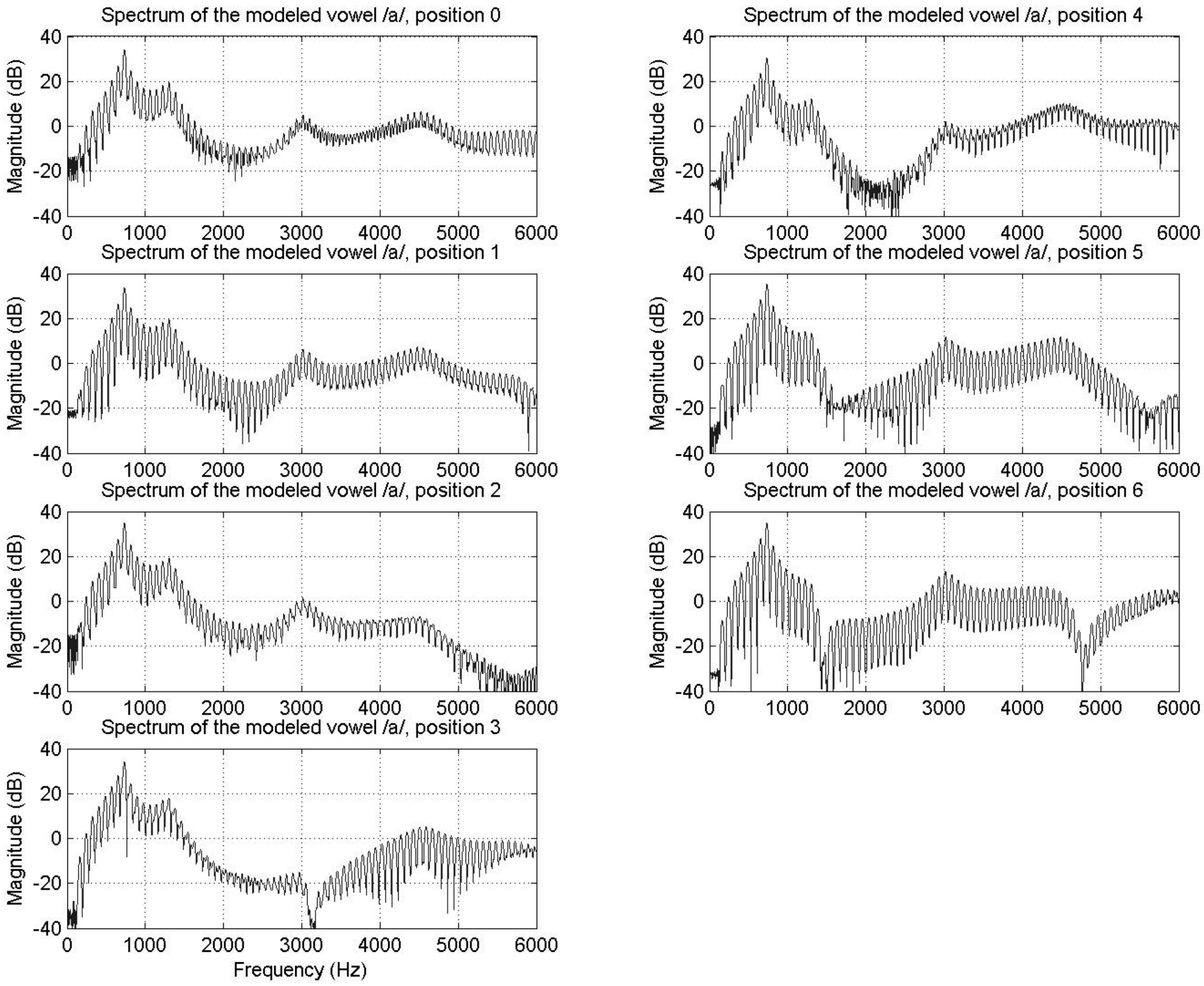
Spectrum of the modeled vowel /ε/, position 14

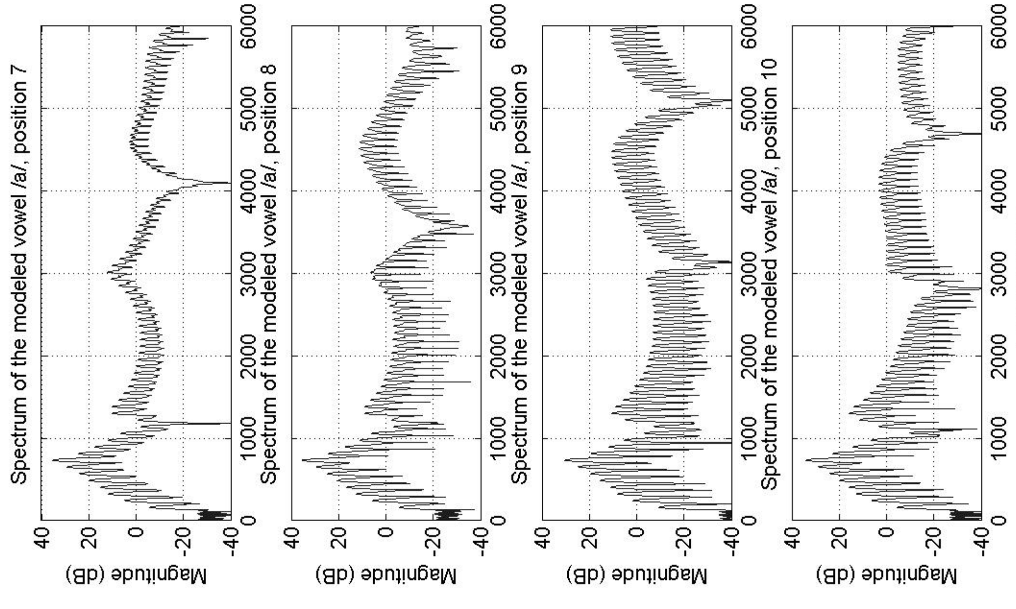
Spectrum of the modeled vowel /ε/, position 15

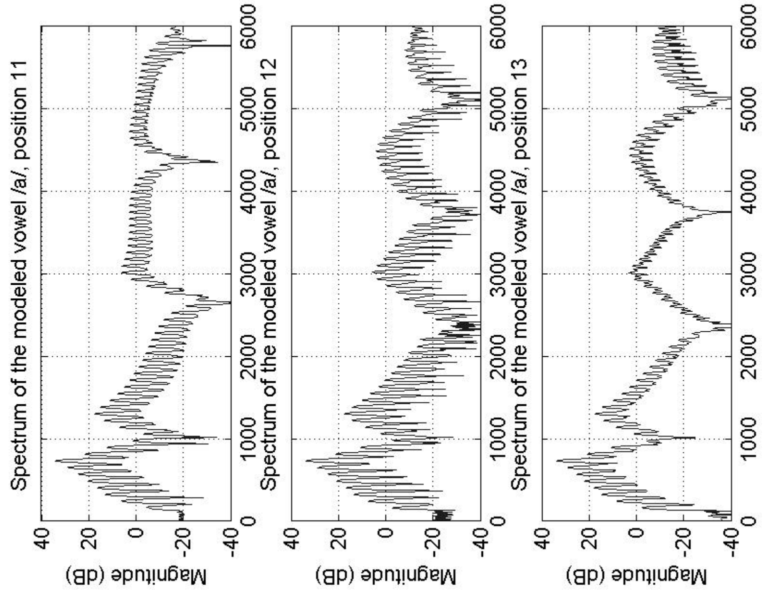
Spectrum of the modeled vowel /ε/, position 16

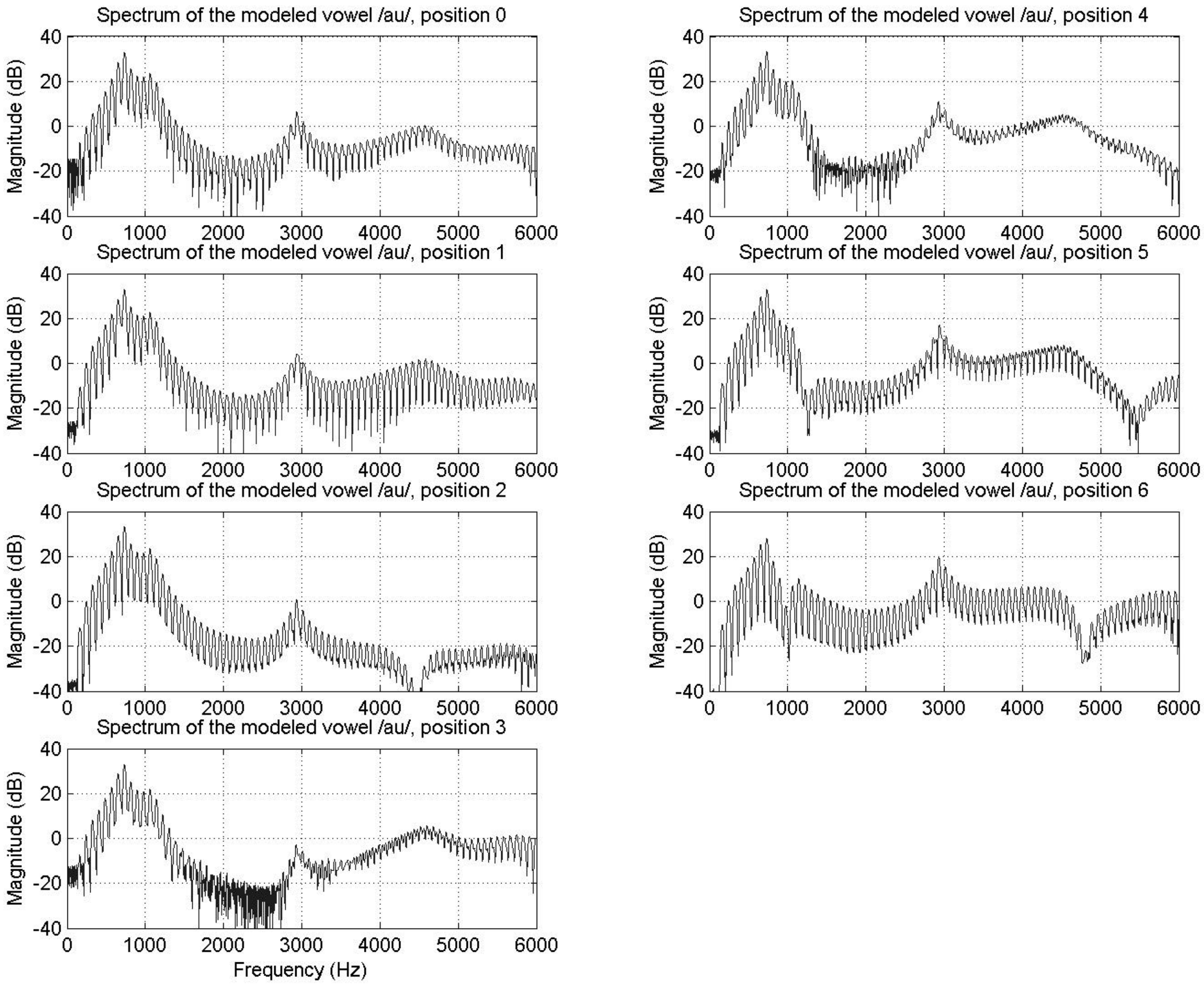

Spectrum of the modeled vowel /ae/, position 0
Spectrum of the modeled vowel /ae/, position 1
Spectrum of the modeled vowel /ae/, position 2
Spectrum of the modeled vowel /ae/, position 3
Spectrum of the modeled vowel /ae/, position 4
Spectrum of the modeled vowel /ae/, position 5
Spectrum of the modeled vowel /ae/, position 6
Magnitude (dB)
40
20
0
-20
-40
0
1000
2000
3000
4000
5000
6000
Frequency (Hz)

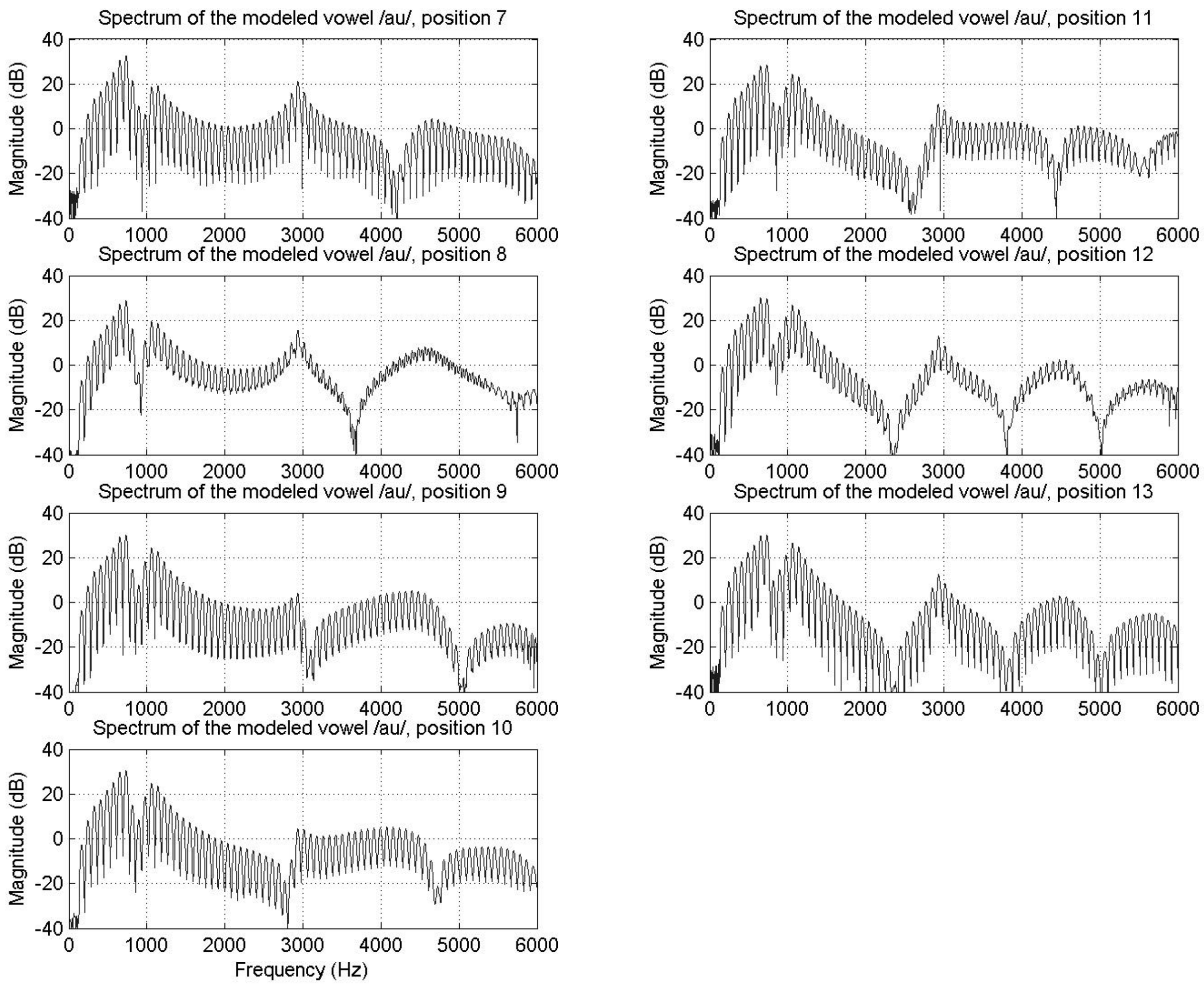

Spectrum of the modeled vowel /ae/, position 7
Spectrum of the modeled vowel /ae/, position 8
Spectrum of the modeled vowel /ae/, position 9
Spectrum of the modeled vowel /ae/, position 10
Spectrum of the modeled vowel /ae/, position 11
Spectrum of the modeled vowel /ae/, position 12
Spectrum of the modeled vowel /ae/, position 13
Magnitude (dB)
40
20
0
-20
-40
0
1000
2000
3000
4000
5000
6000
Frequency (Hz)

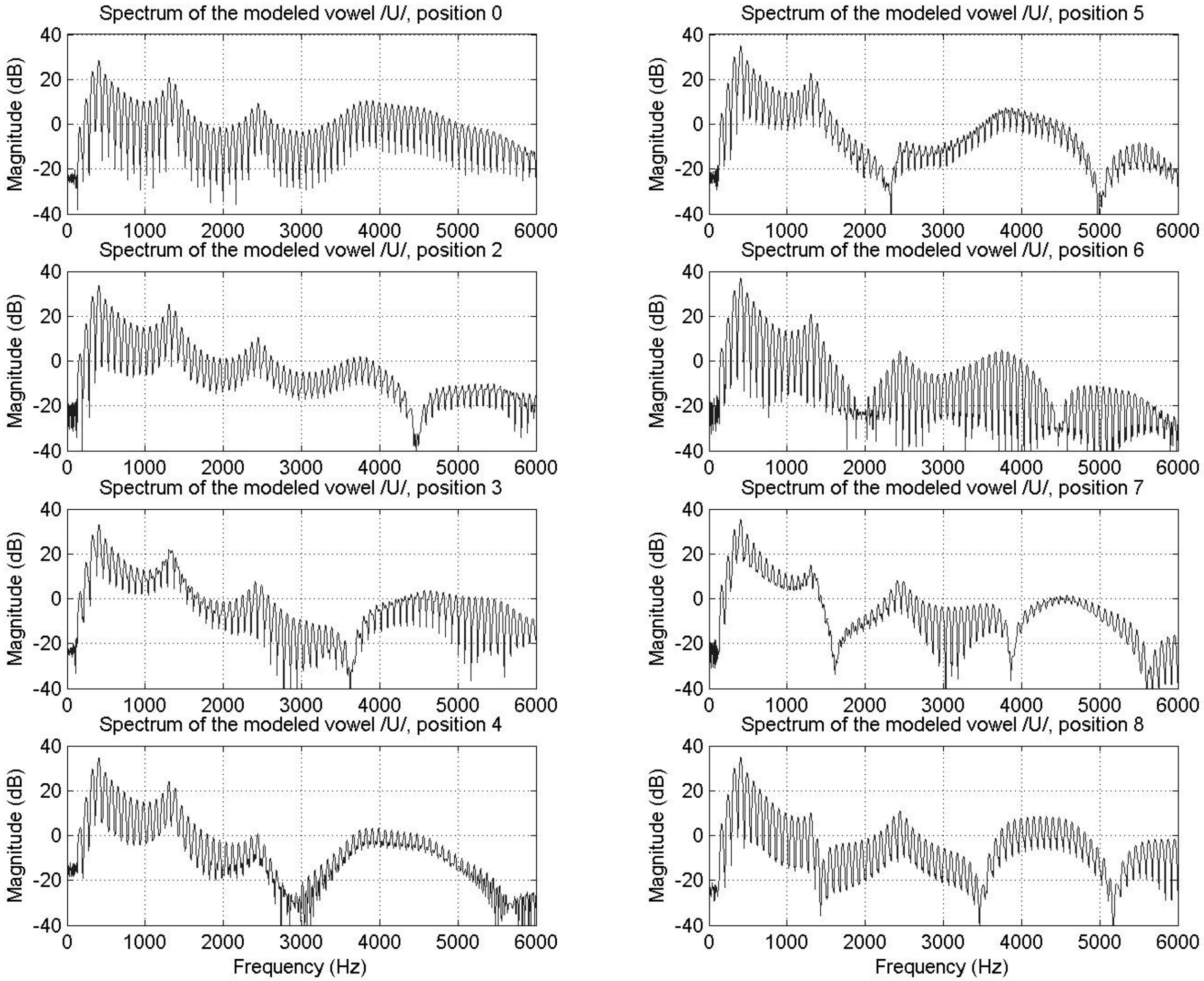

Spectrum of the modeled vowel /a/, position 0
Spectrum of the modeled vowel /a/, position 1
Spectrum of the modeled vowel /a/, position 2
Spectrum of the modeled vowel /a/, position 3
Spectrum of the modeled vowel /a/, position 4
Spectrum of the modeled vowel /a/, position 5
Spectrum of the modeled vowel /a/, position 6
Magnitude (dB)
40
20
0
-20
-40
0
1000
2000
3000
4000
5000
6000
Frequency (Hz)

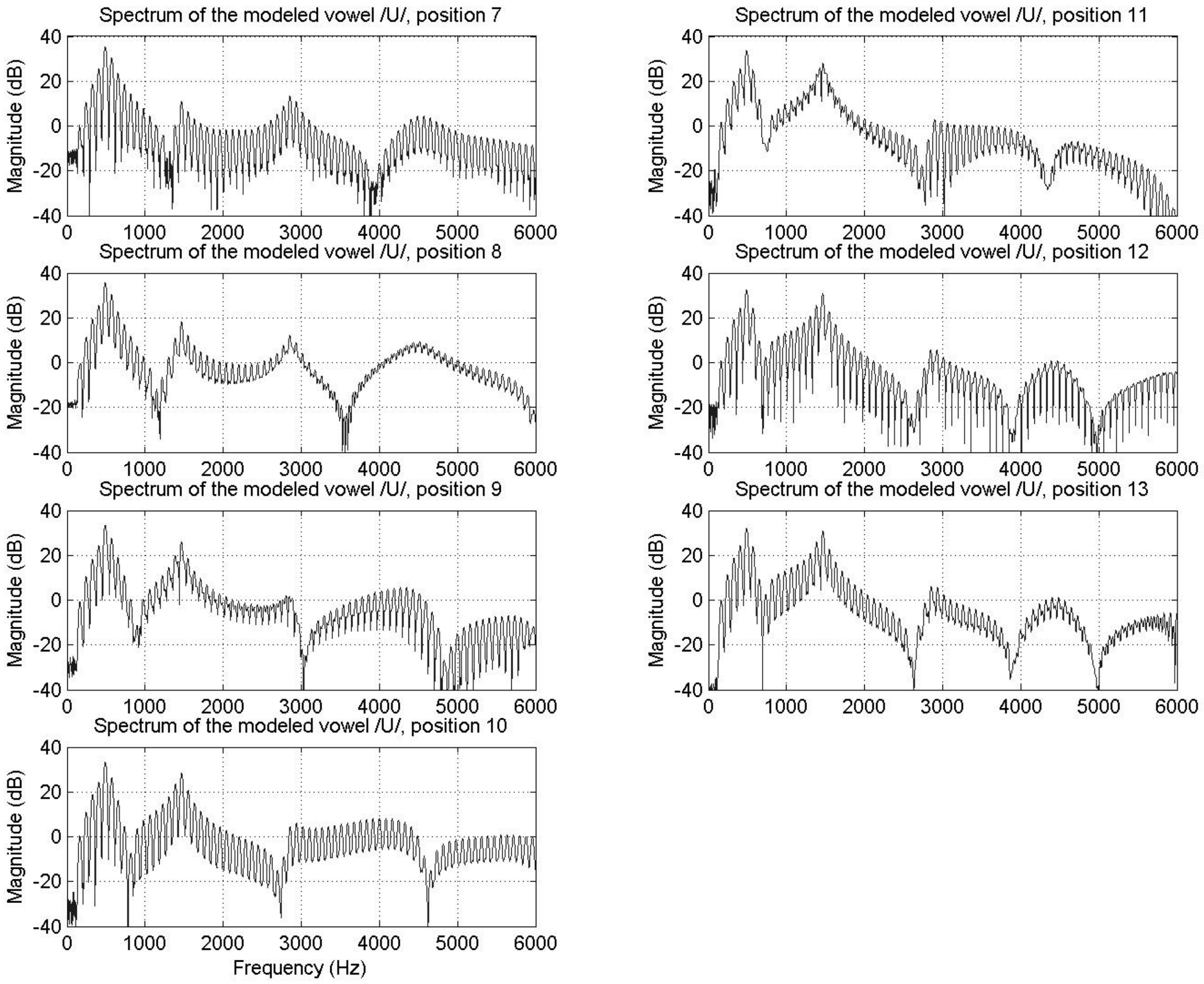

Spectrum of the modeled vowel /a/, position 7
Spectrum of the modeled vowel /a/, position 8
Spectrum of the modeled vowel /a/, position 9
Spectrum of the modeled vowel /a/, position 10
Magnitude (dB)
40
20
0
-20
-40
0
1000
2000
3000
4000
5000
6000
Frequency (Hz)


Spectrum of the modeled vowel /a/, position 11
Spectrum of the modeled vowel /a/, position 12
Spectrum of the modeled vowel /a/, position 13
Magnitude (dB)
40
20
0
-20
-40
0
1000
2000
3000
4000
5000
6000


Spectrum of the modeled vowel /au/, position 0
Spectrum of the modeled vowel /au/, position 1
Spectrum of the modeled vowel /au/, position 2
Spectrum of the modeled vowel /au/, position 3
Spectrum of the modeled vowel /au/, position 4
Spectrum of the modeled vowel /au/, position 5
Spectrum of the modeled vowel /au/, position 6
Magnitude (dB)
40
20
0
-20
-40
0
1000
2000
3000
4000
5000
6000
Frequency (Hz)


Spectrum of the modeled vowel /au/, position 7
Spectrum of the modeled vowel /au/, position 8
Spectrum of the modeled vowel /au/, position 9
Spectrum of the modeled vowel /au/, position 10
Spectrum of the modeled vowel /au/, position 11
Spectrum of the modeled vowel /au/, position 12
Spectrum of the modeled vowel /au/, position 13
Magnitude (dB)
40
20
0
-20
-40
0
1000
2000
3000
4000
5000
6000
Frequency (Hz)


Spectrum of the modeled vowel /U/, position 0
Spectrum of the modeled vowel /U/, position 2
Spectrum of the modeled vowel /U/, position 3
Spectrum of the modeled vowel /U/, position 4
Spectrum of the modeled vowel /U/, position 5
Spectrum of the modeled vowel /U/, position 6
Spectrum of the modeled vowel /U/, position 7
Spectrum of the modeled vowel /U/, position 8
Magnitude (dB)
40
20
0
-20
-40
0
1000
2000
3000
4000
5000
6000
Frequency (Hz)


Spectrum of the modeled vowel /U/, position 7

Spectrum of the modeled vowel /U/, position 8

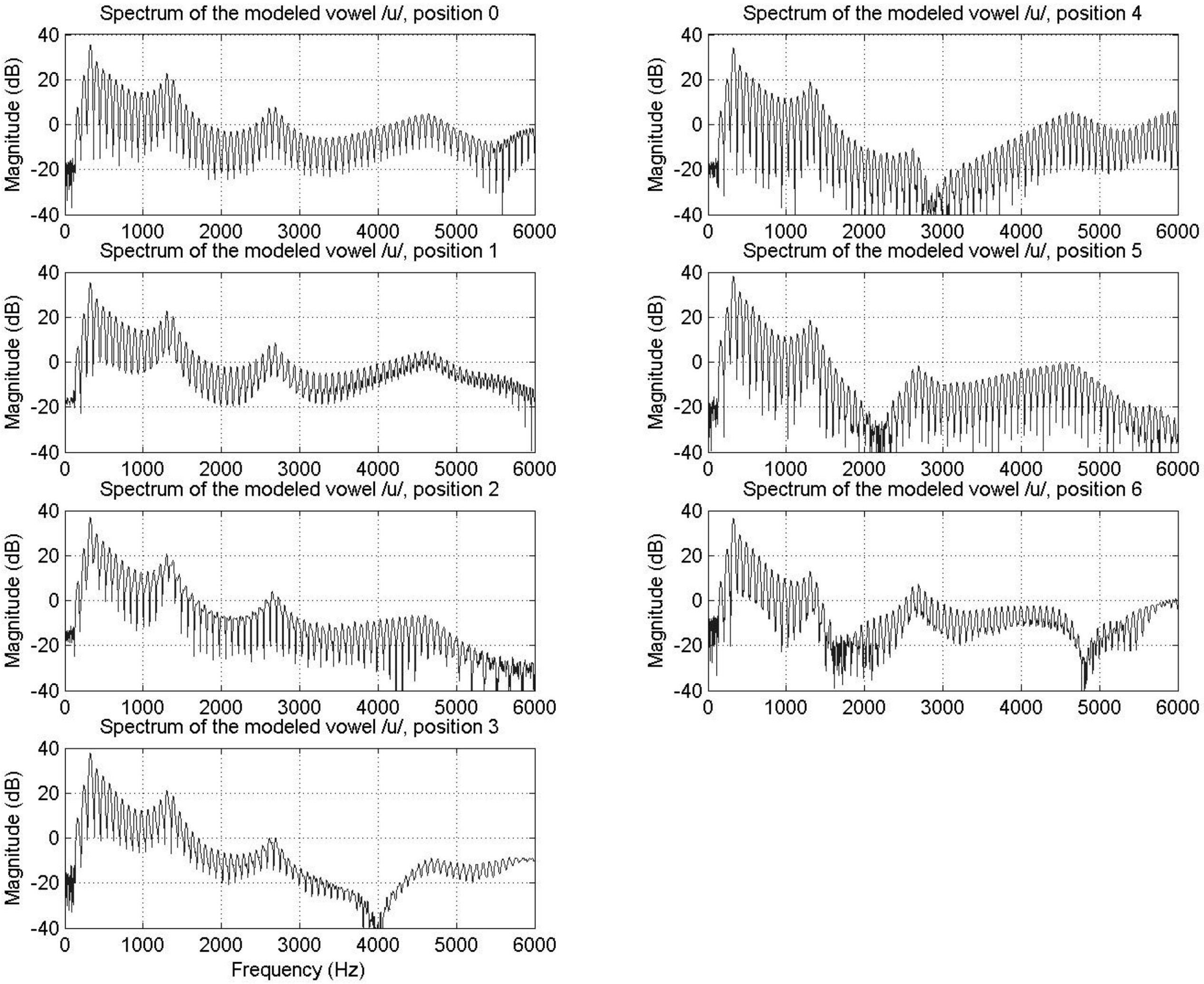
Spectrum of the modeled vowel /U/, position 9

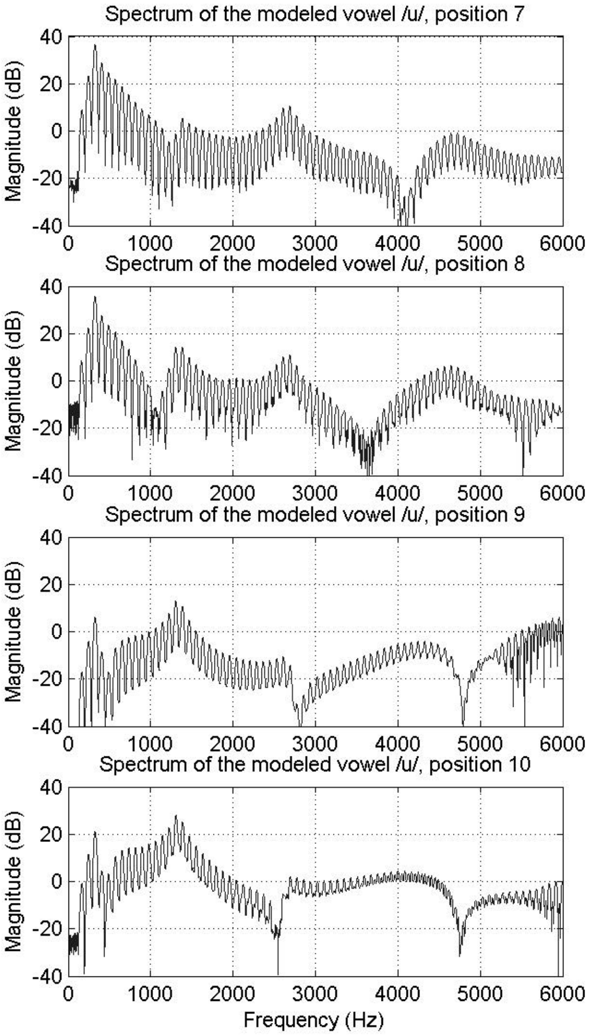
Spectrum of the modeled vowel /U/, position 10

Spectrum of the modeled vowel /U/, position 11

Spectrum of the modeled vowel /U/, position 12

Spectrum of the modeled vowel /U/, position 13


Spectrum of the modeled vowel /u/, position 0
Spectrum of the modeled vowel /u/, position 1
Spectrum of the modeled vowel /u/, position 2
Spectrum of the modeled vowel /u/, position 3
Spectrum of the modeled vowel /u/, position 4
Spectrum of the modeled vowel /u/, position 5
Spectrum of the modeled vowel /u/, position 6
Magnitude (dB)
40
20
0
-20
-40
0
1000
2000
3000
4000
5000
6000
Frequency (Hz)

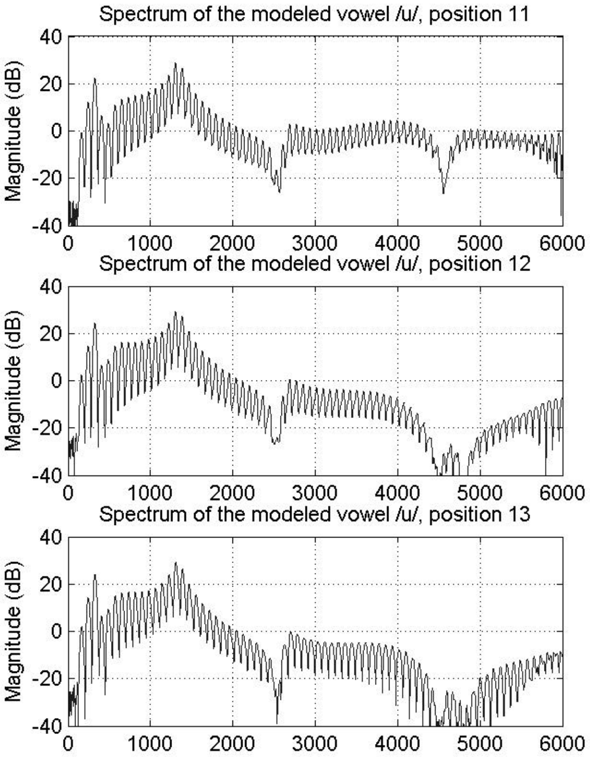


Spectrum of the modeled vowel /u/, position 7

Spectrum of the modeled vowel /u/, position 8

Spectrum of the modeled vowel /u/, position 9

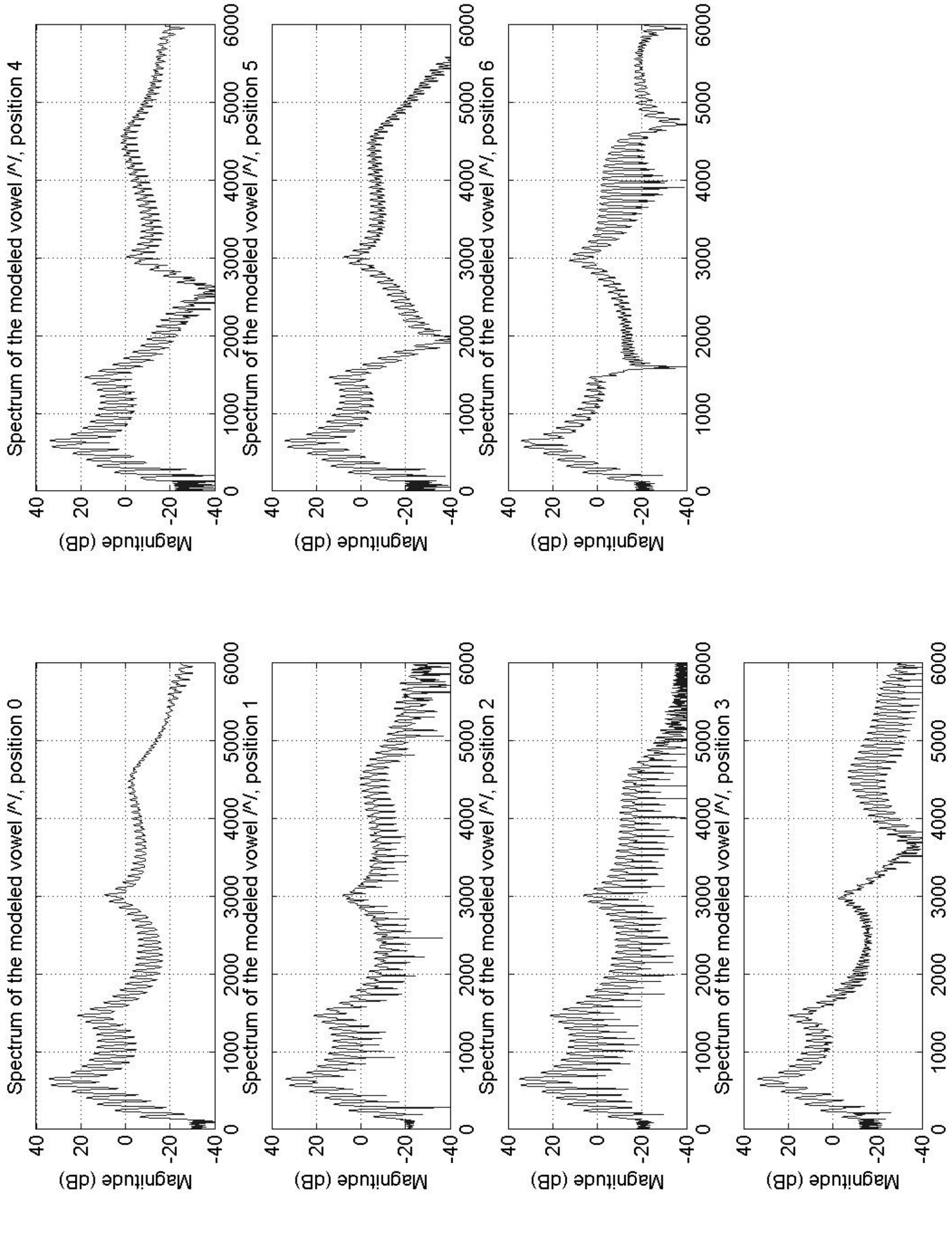
Spectrum of the modeled vowel /u/, position 10

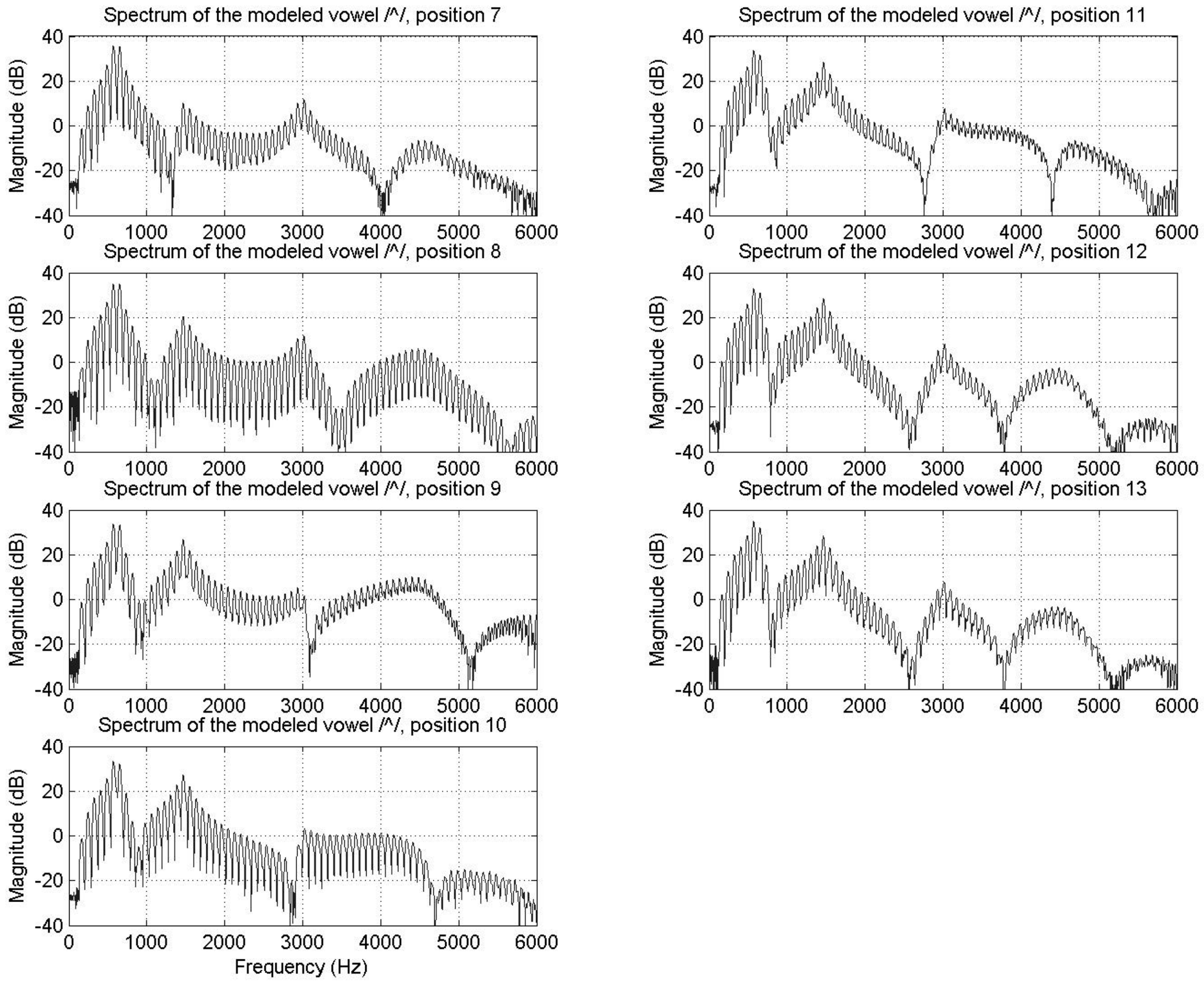


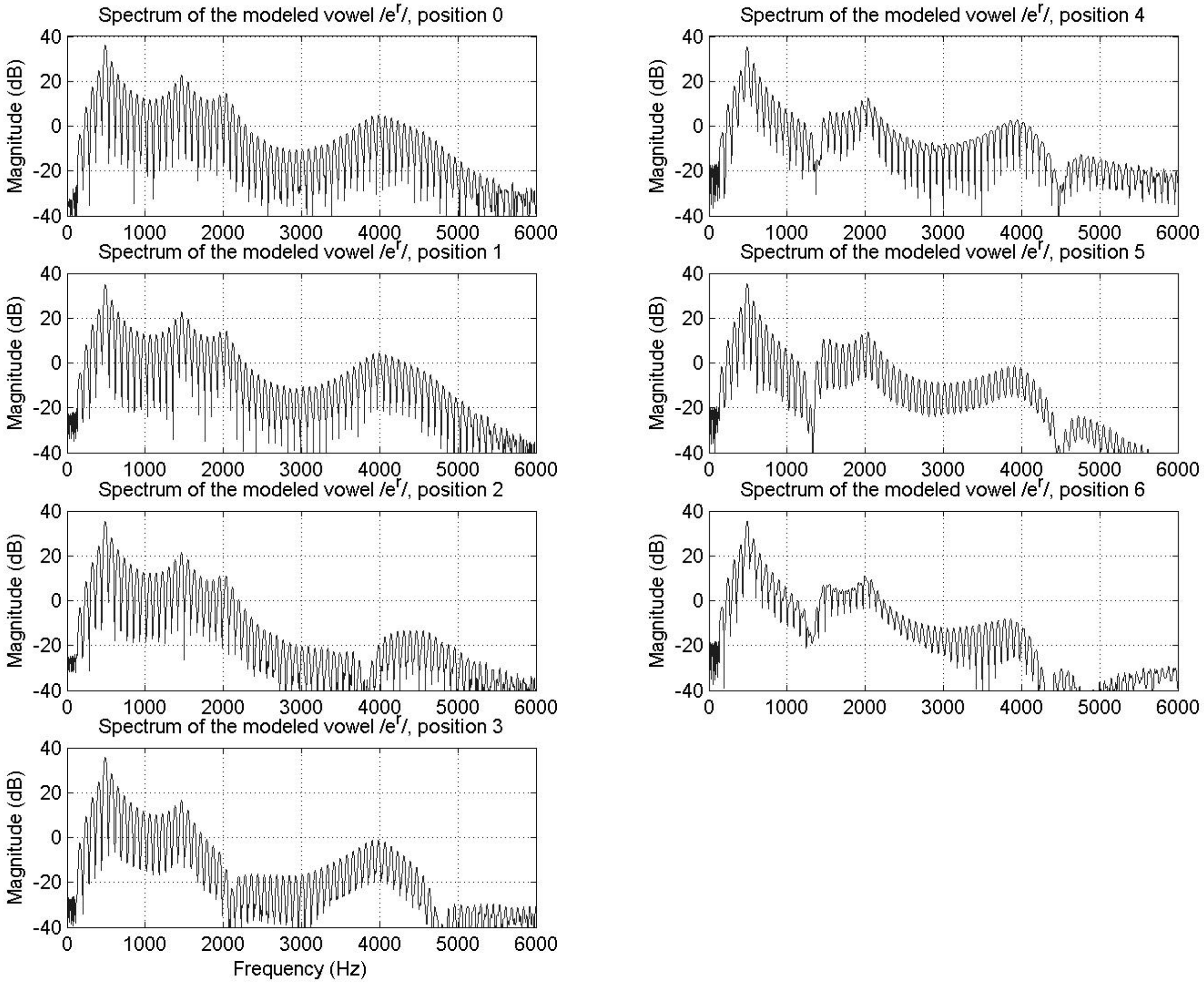
Spectrum of the modeled vowel /u/, position 11

Spectrum of the modeled vowel /u/, position 12

Spectrum of the modeled vowel /u/, position 13


Spectrum of the modeled vowel /^/, position 0
Spectrum of the modeled vowel /^/, position 1
Spectrum of the modeled vowel /^/, position 2
Spectrum of the modeled vowel /^/, position 3
Spectrum of the modeled vowel /^/, position 4
Spectrum of the modeled vowel /^/, position 5
Spectrum of the modeled vowel /^/, position 6
Magnitude (dB)
40
20
0
-20
-40
0
1000
2000
3000
4000
5000
6000
Frequency (Hz)


Spectrum of the modeled vowel /^/, position 7
Spectrum of the modeled vowel /^/, position 8
Spectrum of the modeled vowel /^/, position 9
Spectrum of the modeled vowel /^/, position 10
Spectrum of the modeled vowel /^/, position 11
Spectrum of the modeled vowel /^/, position 12
Spectrum of the modeled vowel /^/, position 13
Magnitude (dB)
40
20
0
-20
-40
0
1000
2000
3000
4000
5000
6000
Frequency (Hz)



Spectrum of the modeled vowel /e^r/, position 0

Spectrum of the modeled vowel /e^r/, position 1

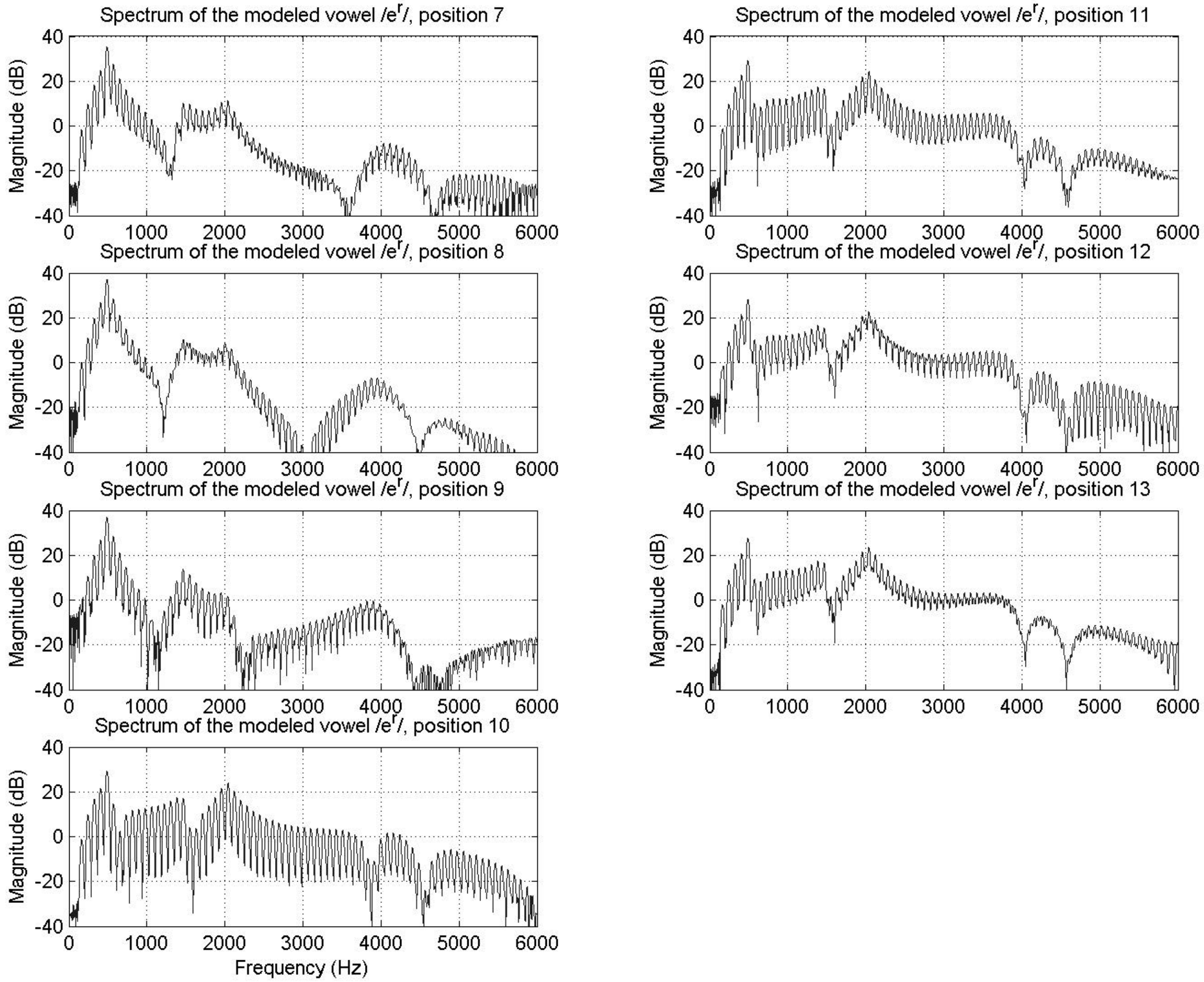
Spectrum of the modeled vowel /e^r/, position 2

Spectrum of the modeled vowel /e^r/, position 3

Spectrum of the modeled vowel /e^r/, position 4

Spectrum of the modeled vowel /e^r/, position 5

Spectrum of the modeled vowel /e^r/, position 6


Spectrum of the modeled vowel /e^r/, position 7

Spectrum of the modeled vowel /e^r/, position 8

Spectrum of the modeled vowel /e^r/, position 9

Spectrum of the modeled vowel /e^r/, position 10

Spectrum of the modeled vowel /e^r/, position 11

Spectrum of the modeled vowel /e^r/, position 12

Spectrum of the modeled vowel /e^r/, position 13

15. Appendix F. The Listener Consent Form for the Perceptual Experiments

MASSACHUSETTS EYE & EAR INFIRMARY

INFORMED CONSENT

TITLE:_Auditory-perceptual evaluation of the quality of speech produced using a new linear electrolarynx transducer as compared to speech produced using a Servox electrolarynx

INVESTIGATOR(S): Robert E. Hillman, Ph.D., Asako Masaki, B.S., Geoff Meltzner, M.S.

==

DESCRIPTION AND EXPLANATION OF PROCEDURES:

The following information is provided to assist you in deciding if you wish to give your voluntary informed consent to participate in a research study being conducted at the Massachusetts Eye and Ear Infirmary (MEEI). The purpose of this study is to help improve the quality of speech produced by patients who must use a mechanical device to communicate (electrolarynx) because of a loss of laryngeal (voice box) function.

You will be seated at a computer workstation and fitted with a pair of headphones. You will be asked to listen to pairs of different speech samples. Your task will be to decide which of two samples in each pair you think sounds more like normal natural speech. An example of normal natural speech (target) will be available for you to listen to. Directions on how to start and proceed through the listening session will be displayed on the computer screen. The entire session should take no more than 1 hour. If you have any questions during the test please do not hesitate to ask the experimenter.

RISKS AND DISCOMFORTS:

There are minimal risks involved with participating in this study. In order to prevent any potential discomfort, please adjust the headphones so it will fit your head snuggly. The volume of the speech stimuli can also be adjusted to accommodate your hearing comfort. If it any time you feel fatigued, you will be allowed to take a break.

POTENTIAL BENEFITS:

Unfortunately there are no immediate potential benefits for you. However, by participating in this study you will help research intended to benefit those who are no longer able to speak with their normal voices.

CONFIDENTIALITY:

Information derived from this study may be used for research purposes that may include publication and teaching. Your identity will be kept confidential.

RIGHT TO WITHDRAW:

Your participation in this study is entirely voluntary, and you may withdraw from the study even after signing this consent. The quality of care you will receive at the Massachusetts Eye and Ear Infirmary will not be affected in any way if you decide not to participate or if you withdraw from the study.

COMPENSATION:

In the unlikely event that you should be injured as a direct result of this study, you will be provided with emergency medical treatment through the emergency room at the MEEI at 617-573-3420. This treatment does not imply any negligence on the part of the Massachusetts Eye and Ear Infirmary or any of the physicians involved. When applicable, the Massachusetts Eye and Ear Infirmary reserves the right to bill third party payers for any emergency services rendered. The Massachusetts Eye and Ear Infirmary does not have any program to provide compensation as a result of any injuries. You should understand that by agreeing to participate in this study, you are not waiving any of your legal rights.

RIGHT TO ASK QUESTIONS:

You are free to ask any questions you may have about the study or your treatment as a research subject. Further information about any aspect of this study is available now or at any time during the course of the study from the principal investigator, Dr. Robert E. Hillman at (617) 573-4050. Additionally, you may contact Elayn Byron, Director of Research Administration, at (617) 573-4080 if you have any questions or concerns about your treatment as a research subject.

COSTS:

There will be no costs incurred by participating in this study. You will receive $20 for participating in one listening session.

CONSENT:

The purpose and procedures of this research project with its possible risks and benefits have been fully and adequately explained to me, and I understand them. I voluntarily agree to participate as a subject in the research project, and understand that by signing this consent form I am indicating that agreement. I have been given a copy of this consent form.

__________ ____________________ ________________________________
Date Name of Subject Signature of Subject

__________ ____________________ ________________________________
Date Name of Witness Signature of Witness

_______ ________________________________
Date Signature of Investigator

Investigators:
Robert E. Hillman
Voice and Speech Lab
MEEI
243 Charles Street
Boston, MA 02114
617-573-4050

Asako Masaki
Voice and Speech Lab
MEEI
243 Charles Street
Boston, MA 02114
617-573-4050

Geoff Meltzner
Voice and Speech Lab
MEEI
243 Charles Street
Boston, MA 02114
617-573-4050

16. Bibliography

Barney, H. L., Haworth, F. E., & Dunn, H. K. (1959). “An experimental transistorized artificial larynx.” In B. Weinberg (Ed.), *Readings in Speech Following Total Laryngectomy* (pp. 1337-1356). Baltimore: University Park Press.

Bell Laboratories, I. (1959). “New artificial larynx.” *Trans. Am. Acad. Ophthalmol. Otolaryngol.*, 63, 548-550.

Bessette, B., Salami, R, Lefebvre, R., Jelinek, M., Rotola-Pukkila, J., Vainio, J., Mikkola, H., Jarvinen, K. (2001). “The adaptive multirate wideband speech code (AMR-WB).” *IEEE Trans. Speech and Audio Proc.,* 10(8), 620-626

Chiba, T. & Kajiyama, M. (1941). *The Vowel: Its Nature and Structure.* Tokyo: Tokyo-Kaisekan Publishing Co., Ltd., 115-131.

Clark, J. G. (1985). “Alaryngeal speech intelligibility and the older listener.” *J. Speech and Hearing Dis*, 50(1), 60-65.

Cole, D., Sridharan, S., Moody, M., & Geva, S. (1997). “Application of noise reduction techniques for alaryngeal speech enhancement”. *IEEE TENCON*, 491-494.

Crystal, T.H, & House, A.S. “Segmental durations in connected speech signals: Preliminary results.” *J Acoust Soc Am*, 72(3), 705-717

Diedrich, W., & Youngstrom, K. (1977). *Alaryngeal Speech*. Springfield, IL: C.C. Thomas.

Edwards, A.L. (1957). *Techniques of Attitude Scale Construction*. New York: Appleton-Century-Crofts, Inc.

Espy-Wilson, C. Y., Chari, V. R., MacAuslan, J. M., Huang, C. B., & Walsh, M. J. (1998). “Enhancement of electrolaryngeal speech by adaptive filtering.” *J Speech Lang Hear Res*, 41(6), 1253-64.

Federal Information Processing Standards (1998) "Analog to Digital Conversion of Voice by 2,400 Bit/second Mixed Excitation Linear Prediction (MELP)," *Federal Information Processing Standards Publication (FIPS PUB) Draft*, May 28,1998

Gandour, J., Weinberg, B., Petty, S. H., & Dardarananda, R. (1988). “Tone in Thai alaryngeal speech.” *J Speech Hear Disord*, 53, 23-29.

Gao, Y., Shlomot, E., Benyassine, A., Thyssen, J., Huan-yu, S., and Murgia, C. (2001). “The SMV algorithm selected by TIA and 3GPP2 for CDMA applications.” *ICASSP ’01*. (2), 709-712.

Gates, G. A., Ryan, W., Cantu, E., & Hearne, E. (1982a). “Current status of laryngectomee rehabilitation: II. Causes of failure.” *Am J Otolaryngol*, 3(1), 8-14.

Gates, G. A., Ryan, W., Cooper, J. C., Jr., Lawlis, G. F., Cantu, E., Hayashi, T., Lauder, E., Welch, R. W., & Hearne, E. (1982b). “Current status of laryngectomee rehabilitation: I. Results of therapy.” *Am J Otolaryngol*, 3(1), 1-7.

Goldstein, E.A. (2003) *Prosthetic Voice Controlled by Muscle Electromyographic Signals.* Ph.D. Thesis, Harvard University, Cambridge

Goode, R. L. (1969). “The development of an improved artificial larynx.” *Trans Am Acad Ophthalmol Otolaryngol*, 73(2), 279-87.

Goodglass, H. & Kaplan E. (1983). *The Assessment of Aphasia and Related Disorders.* Philadelphia: Lea & Febiger.

Greer, S.C. & DeJaco, A. (2001). “Standardization of the selectable mode vocoder.” *ICASSP ’02* (4), 953-956.

Hanson, H. (1997). “Glottal characteristics of female speakers: Acoustic correlates.” *J. Acoust. Soc. Am.* 101(1), 466-481.

Hanson, H. & Chuang, E. (1999). “Glottal characteristics of male speakers: Acoustic correlates and comparison with female data.” *J. Acoust. Soc. Am.* 106(2), 1064-1077.

Harris, R.J. (2001). *A Primer of Multivariate Statistics.* Mahwah: Lawrence Erlbaum Associates.

Hillman, R. E., Walsh, M. J., Wolf, G. T., Fisher, S. G., & Hong, W. K. (1998). “Functional outcomes following treatment for advanced laryngeal cancer. Part I--Voice preservation in advanced laryngeal cancer. Part II--Laryngectomy rehabilitation: the state of the art in the VA System. Research Speech-Language Pathologists. Department of Veterans Affairs Laryngeal Cancer Study Group.” *Ann Otol Rhinol Laryngol Suppl*, 172, 1-27.

Hillman, R. E. (1999). *Development of an improved artificial electro-larynx communication system* (Funded Project Grant C2243-2DC). Boston MA: Veterans Administration Funded Project Grant number C2243-2DC.

House, A. S., & Stevens, K. N. (1958). “Estimation of formant bandwidths from measurements of transient response of the vocal tract.” *J. Speech Hear Res.,* 1 (4), 309-315.

Houston, K.M., Hillman, R.E., Kobler, J.B., & Meltzner, G.S. (1999). “Development of sound source components for a new electrolarynx speech prosthesis.” *ICASSP ’99,* 4, 2347-2350.

Kaiser, H.F. and Serlin, R.H. (1978). “Contributions to the method of paired comparisons.” *App Psych. Meas.* 2(3), 421-430.

Kelly, J.L. & Lochbaum, C.C. (1962). “Speech synthesis.” *Proc. Fourth Intern Congr. Acous.* Reprinted in: Flanagan, J.L. & Rabiner L.R., editors: (1973) *Speech Synthesis.* Stroudsberg: Dowden Hutchinson & Ross, Inc.

King, P. S., Fowlks, E. W., & Peirson, G. A. (1968). “Rehabilitation and adaptation of laryngectomy patients.” *Am J Phys Med*, 47(4), 192-203.

Klatt, D. H., & Klatt, L. C. (1990). “Analysis, synthesis, and perception of voice quality variations among female and male talkers.” *J Acoust Soc Am*, 87(2), 820-57.

Kommers, M. S., & Sullivan, M. D. (1979). “Wives' evaluation of problems related to laryngectomy.” *J Commun Disord*, 12(5), 411-30.

Krus, D.J. and Krus P. (1977) “Normal scaling of dominance matrices: the domain-referenced model.” *Educational and Psychological Measurement*, 37, 189-193.

Liljencrants, J. (1985). *Speech synthesis with a reflection-type line analog.* D.Sc. thesis, Royal Institute of Technology, Stockholm.

Ma, K., Demirel, P., Espy-Wilson, C., & MacAuslan, J. (1999). “Improvement of electrolarynx speech by introducing normal excitation information.” *EUROSPEECH ‘99*

Maeda, S. (1982). “Digital simulation of the vocal tract system.” *Speech Commun.* 1, 199-299.

McCree, A.V. and Barnwell, T.P. (1995) “A mixed excitation LPC vocoders model for low bit rate speech coding.” *ICASSP’95* 3(4), 242-250.

Meltzner, G.S., Kobler, J.B., & Hillman, R.E. (2003) “Measuring the neck frequency response function of laryngectomy patients: Implications for the design of electrolarynx devices.” *J Acoust Soc Am,* 114(2), 1035-1047.

Mitchell, H.L., Hoit, J.D., and Watson, P.J. (1996). “Cognitive-Linguistic Demands and Speech Breathing.” *J Speech Hear Res.* 39(1) 93-104.

Morris, H. L., Smith, A. E., Van Demark, D. R., & Maves, M. D. (1992). “Communication status following laryngectomy: the Iowa experience 1984-1987.” *Ann Otol Rhinol Laryngol*, 101(6), 503-10.

Mosteller, F. (1951). “Remarks on the method of paired comparisons II: The effect of an aberrant standard deviation when equal standard deviations and equal correlations are assumed.” *Psychometrika,* 16(2), 207-218.

Moulines, E. and Charpentier, F. (1990). “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones.” *Speech Comm.* 9(5), 453-467.

Myrick, R. and Yantorno, R. (1993). “Vocal tract modeling as related to the use of an artificial larynx.” *Bioeng. Conf., Proc IEEE Nineteenth Annual Northeast*. 212-214

Ng, M. L., Lerman, J. W., & Gilbert, H. R. (1998). “Perceptions of tonal changes in normal laryngeal, esophageal, and artificial laryngeal male Cantonese speakers.” *Folia Phoniatr Logop*, 50(2), 64-70.

Peterson, G. E., & Barney, H. L. (1952). “Control Methods Used in a Study of Vowels.” *J. Acoust. Soc. Am*, 24 (1), 175-184.

Qi, Y. Y., & Weinberg, B. (1991). “Low-frequency energy deficit in electrolaryngeal speech.” *J Speech Hear Res*, 34(6), 1250-6.

Richardson, J., & Bourque, L. (1985). “Communication after laryngectomy.” *J. Psychosoc. Oncol.*, 3, 83-97.

Sisty, N. L., & Weinberg, B. (1972). “Formant Frequency Characteristics of Esophageal Speech.” *J. Speech and Hear Re,* 15, 439-448.

Stevens, K.N. (1998) Acoustic Phonetics. Cambridge: MIT Press.

Story, B.H., Titze, I.R., & Hoffman, E.A. (1996) “Vocal tract area functions from magnetic resonance imaging.” *J. Acoust. Soc. Am.,* 100(1), 537-554.

Story, B.H., and Titze, I.R. (1998). “Parameterization of vocal tract area functions by empirical orthogonal modes.” *J. Phonetics*, 26(3), 223-260.

Torgerson, W.S. (1957) *Theory and Methods of Scaling.* New York: John Wiley and Sons.

Thurstone, L.L (1927). “A law of comparative judgment.” *Psych. Rev.* 34, 273-286.

Uemi, N., Ifukube, T., Takahashi, M., & Matsushima, J. (1994). “Design of a new electrolarynx having a pitch control function.” IEEE Workshop on Robot and Human Communication, 198-202

Verdolini, K., Skinner, M. W., Patton, T., & Walker, P. A. (1985). “Effect of amplification on the intelligibility of speech produced with an electrolarynx.” *Laryngoscope*, 95(6), 720-6.

Webster, P. M., & Duguay, M. J. (1990). Surgeons' reported attitudes and practices regarding alaryngeal speech. *Ann Otol Rhinol Laryngol*, 99(3 Pt 1), 197-200.

Weiss, M. S., & Basili, A. G. (1985). “Electrolaryngeal speech produced by laryngectomized subjects: perceptual characteristics.” *J Speech Hear Res.*, 28(2), 294-300.

Weiss, M. S., Yeni-Komshian, G. H., & Heinz, J. M. (1979). “Acoustical and perceptual characteristics of speech produced with an electronic artificial larynx.” *J Acoust Soc Am*, 65(5), 1298-1308.