Transforming High-Effort Voices Into Breathy Voices Using Adaptive Pre-Emphasis Linear Prediction

by

Karl Ingram Nordstrom
B.Eng., University of Victoria, 1995
M.A.Sc., University of Victoria, 2000

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Electrical Engineering

Transforming High-Effort Voices Into Breathy Voices
Using Adaptive Pre-Emphasis Linear Prediction

By

Karl Ingram Nordstrom
B.Eng., University of Victoria, 1995
M.A.Sc., University of Victoria, 2000

Supervisory Committee

Dr. Peter F. Driessen, Supervisor
(Department of Electrical Engineering)

Dr. George Tzanetakis, Departmental Member
(Department of Electrical Engineering and Department of Computer Science)

Dr. Wu-Sheng Lu, Departmental Member
(Department of Electrical Engineering)

Dr. Dale J. Shpak, Departmental Member
(Department of Electrical Engineering)

Dr. John Esling, Outside Member
(Department of Linguistics)

Supervisory Committee

Dr. Peter F. Driessen, Supervisor
(Department of Electrical Engineering)

Dr. George Tzanetakis, Departmental Member
(Department of Electrical Engineering and Department of Computer Science)

Dr. Wu-Sheng Lu, Departmental Member
(Department of Electrical Engineering)

Dr. Dale J. Shpak, Departmental Member
(Department of Electrical Engineering)

Dr. John Esling, Outside Member
(Department of Linguistics)

Abstract

During musical performance and recording, there are a variety of techniques and electronic effects available to transform the singing voice. The particular effect examined in this dissertation is breathiness, where artificial noise is added to a voice to simulate aspiration noise. The typical problem with this effect is that artificial noise does not effectively blend into voices that exhibit high vocal effort. The existing breathy effect does not reduce the perceived effort; breathy voices exhibit low effort.

A typical approach to synthesizing breathiness is to separate the voice into a filter representing the vocal tract and a source representing the excitation of the

vocal folds. Artificial noise is added to the source to simulate aspiration noise. The modified source is then fed through the vocal tract filter to synthesize a new voice. The resulting voice sounds like the original voice plus noise.

Listening experiments were carried out. These listening experiments demonstrated that constant pre-emphasis linear prediction (LP) results in an estimated vocal tract filter that retains the perception of vocal effort. It was hypothesized that reducing the perception of vocal effort in the estimated vocal tract filter may improve the breathy effect.

This dissertation presents adaptive pre-emphasis LP (APLP) as a technique to more appropriately model the spectral envelope of the voice. The APLP algorithm results in a more consistent vocal tract filter and an estimated voice source that varies more appropriately with changes in vocal effort. This dissertation describes how APLP estimates a spectral emphasis filter that can transform the spectral envelope of the voice, thereby reducing the perception of vocal effort.

A listening experiment was carried out to determine whether APLP is able to transform high effort voices into breathy voices more effectively than constant pre-emphasis LP. The experiment demonstrates that APLP is able to reduce the perceived effort in the voice. In addition, the voices transformed using APLP sound less artificial than the same voices transformed using constant pre-emphasis LP. This indicates that APLP is able to more effectively transform high-effort voices into breathy voices.

Contents

List of Tables

List of Figures

Acknowledgments

I would like to acknowledge the help of a number of people in completing this dissertation. This work started as an NSERC scholarship in collaboration with IVL Technologies in Victoria. Thanks goes to Brian Gibson at IVL for financially supporting the start of this project. At IVL and at associated TC-Helicon, Glen Rutledge mentored me in digital signal processing for voice and helped to establish the research project. Throughout the PhD, Peter Driessen, my supervisor, provided financial and other valuable ongoing support. I was initiated into the complexities of voice physiology through John Esling through extended discussions and a number of listening experiments. Anne Bateman also provided musical and phonetic expertise, as well as a collection of useful sound files. Mathieu Lagrange translated some of the algorithms that I developed into Marsyas, an audio processing framework developed by George Tzanetakis. In the mid to later stages of the process, I encountered writing challenges and the insightful help of George Tzanetakis helped me to break free and complete my research. I also want to thank Kevin Alexander and others at TC-Helicon for lending equipment and for providing related technical employment. None of this would have been possible

without my parents and their moral support. They established my life in a way that made this PhD achievable. Lastly and most importantly, my wife, Rachelann, has come along with me on this rocky ride and has always supported me. I thank her for her love. My children – Amber, Sarina and Kaden – have also joyfully come along for the ride, their voices, at times, playfully phonating vowels with varying quantities of breathiness and vocal effort.

Dedicated to:

Rachelann,

Amber, Sarina and Kaden

Chapter 1

Introduction

In the musical world today, singers are getting used to the idea of their voice as an instrument that can be digitally enhanced. This evolution from a purely acoustic instrument to an electronically enhanced instrument has already occurred for other instruments. The piano has evolved into the electronic keyboard and the acoustic guitar has evolved into the electric guitar. Innumerable effects have been created to electronically modify the sonic textures of these instruments. Recently, vocal effects have become more accepted and common in the creation of music. This dissertation concerns the improvement of a particular effect that adds breathiness to singing voices. The techniques developed here can also be transferred to a broad range of voice modeling techniques based upon linear prediction (LP).

Over the years, a range of effects have been developed to enhance and modify the voice during musical recording and performance. Many of these effects are subtle, related to recording techniques. Relatively subtle effects that have a close

relationship to acoustic phenomena are reverb and vocal doubling, where the voice is re-recorded over top of itself singing the same vocal line. Dynamics processing, such as compression, is often used to maintain the voice at the forefront of the recorded mix and de-essing is often used in these situations to reduce the resulting prominence of sibilants. Chorus effects have also been applied to thicken the sound of the voice.

Radical effects have also been explored such as the vocoder, guitar talk box, and distortion. Due to the extreme nature of these effects, they are only used on a minority of songs.

The most influential effect, and likely the most controversial, is pitch correction. This is an effect that is a significant modification of the voice, enabling many singers to sound better than they ever could in real life. Pitch correction has become an accepted part of the recording process, affecting almost every vocal recording in popular music today. Pitch correction has also lead to other effects such as pitch shifting that can create harmonies by making copies of the original voice at different pitches. One artifact in pitch correction as become known as the "Cher effect", where instead of gradual glide from pitch to pitch, heavy pitch correction leads to a sudden change as the pitch "pops" from one pitch to the next.

Pitch correction has been around long enough that it is now starting to be publicly accepted. This, in turn, has made people curious about other vocal modifications that can be made to the voice. The musical space for vocal effects with various sonic textures has only started to be explored.

The particular effect investigated in this dissertation is that of a breathiness

effect. This effect adds breathiness to a singing voice, making the original voice sound like it has more aspiration noise. This effect works by decomposing the voice into a voice source representing the air rushing through the vocal folds and a filter representing the the influence of the vocal tract using linear prediction (LP) [1, 2]. Synthetic noise representing aspiration noise at the vocal folds is added to the voice source [3]. The new vocal source is then passed through the vocal tract filter to synthesize the modified voice. The breathiness effect works well for voices that already sound a little breathy. However, for voices that do not exhibit breathiness, especially high-effort voices, the added noise does not blend easily into the voice and instead sounds like a segregated stream of sound, separate from the voice [4]. This dissertation explores the issue of why the breathiness effect does not blend easily into high-effort voices.

The breathiness effect is closely related to voice conversion [5, 6, 7, 8, 9], where the goal is to transform one voice into another using segmented processing. This typically involves breaking the voice signal into phoneme units. These phoneme units are then mapped to phoneme units from the target voice. As such, the resynthesis is often a form of concatenative synthesis [10]. The breathiness effect differs from voice conversion in that the goal of the breathiness effect is to transform only dimensions of the voice associated with breathiness and to do so in real-time with low latency. This means that that the algorithm will not map the phonemes themselves.

Another related field is that of audio morphing [11]. In the audio morph, the goal is to transform one audio sound into another audio sound to create entirely

new forms of sound. For example, one might want to transform a singing voice into a trumpet. Audio morphing involves mapping the audio characteristics of one sound to the audio characteristics of a new sound. There is some skepticism whether it is possible to create entirely new sounds through audio morphing due to the categorical nature of auditory perception. It is far more likely to create a "funny sounding trumpet" than it is to create a sound that people perceive to be entirely new. Voice conversion is a more narrowly defined version of audio morphing.

The remainder of this chapter is devoted to a description of high-effort and breathy voice qualities and a discussion of the problem at hand.

1.1 High-Effort and Breathy Voice Qualities

To digitally manipulate voice qualities such as breathiness and vocal effort, it is helpful to understand how these voice qualities are produced and how they manifest themselves in the voice signal.

Breathiness is associated with relaxed vocal folds and open glottis. When a voice is relaxed, the vocal folds move freely, with a slow rate of glottal closure. Air often leaks between the vocal folds when the voice is relaxed and there may not even be complete glottal closure. When air leakage causes significant aspiration noise and the vocal folds are relaxed, the voice is known as a breathy voice. To create a breathy voice, the vocal folds must be relaxed, free to vibrate, and without undue constriction in the lower vocal tract [12]. This is opposite to a high-effort

voice where the vocal folds are tense.

There are many terminologies describing various kinds of high-effort voices. Vocal effort has been chosen in the context of this research because increased effort describes a broad range of voice qualities where the vocal folds remain closed for a large portion of the glottal cycle. These voices have more high frequency harmonic content due to the short length of the glottal pulses and the rapid closure of the vocal folds, i.e., the glottal waveform approaches an impulse train. The high-effort terminology was also chosen because it describes something that most people can understand more easily than the standardized phonetic terminology [12]. People do not need specialized phonetic training to achieve a relatively consistent perception of vocal effort. It is more difficult to teach people the meaning of phonetic terms such as a pressed, laryngealized, creaky, or harsh voice. Vocal effort is a concept that both specialists and non-specialists can grasp and come to agreement over more easily [13, 14]. Since many of the subjects in the listening experiments are not experts in phonetics, the vocal effort terminology is most appropriate.

Vocal effort is a subjective term that describes a strained or tense voice quality. Although the most obvious consequence of increased vocal effort is increased sound intensity [15], people can distinguish the quantity of effort in a voice independent of the volume of the sample playback [13]. Vocal effort also affects the relative difference in sound pressure levels between vowels and consonants [16] as well as affecting the relative durations between vowels and consonants [17]. Pitch can also be an indication of vocal effort [16, 17] with higher pitches associated with higher levels of vocal effort.

In the case of singing, the pitch has already been specified. Therefore, the dominant cue of vocal effort for the singing voice is the spectral envelope of the signal [14, 18]. When a voice involves effort, it has more high frequency content than the same voice in a relaxed state [19].

The spectral envelope of the voice source provides one of the most important cues for the perception of vocal effort. This envelope varies from voice to voice and can vary within the context of a single phrase [20]. Studies show that it is possible to model the spectral envelope of the voice source with a third-order, all-pole, low-pass filter [21, 22]. These studies modeling the spectral envelope of the voice source show that the rate at which the vocal folds close (i.e., the rate of the glottal return phase) affects the spectral slope. A slow glottal return phase, such as in a breathy voice, results in a steeper slope starting at a lower frequency, producing little high-frequency content in the voice source. A quick glottal return phase, such as for a high-effort voice, results in a less steep slope and more high-frequency content in the voice source, because the instant of glottal closure is more abrupt and impulsive resulting in a flatter spectrum.

The frequency response of the vocal tract also influences the spectral envelope of the voice. Perceptually, the main characteristic of the vocal tract is that it produces the perception of vowels with narrow spectral peaks known as formants. However, the vocal tract filter also influences the spectral emphasis of the voice. The singer's formant results in the clustering of the third, fourth and fifth formants [23]. Acoustic resonances within the vocal tract can interact with the glottal source, creating small changes in the glottal waveform [24]. For example,

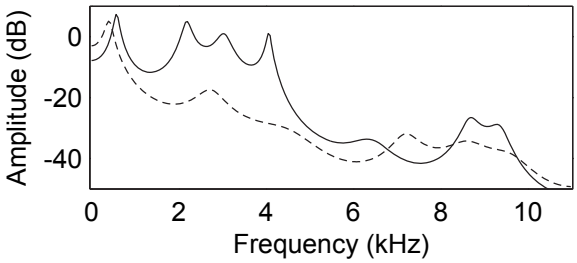


Figure 1.1: Spectral envelopes estimated by linear prediction without pre-emphasis: a breathy voice (dashed line) and a high-effort voice (solid line). In each plot the same voice is singing the same vowel on the same fundamental frequency. The breathy voice has less energy in the $1.5 - 4.5$ kHz range than the corresponding high-effort voice.

when the vocal tract is constricted, the load of the vocal tract upon the source can cause the glottal waveform to become skewed such that the opening of the glottis is more gradual and closure is more rapid. The lower vocal tract can change significantly in the production of different voice qualities [25, 26]. High-effort voices are often associated with constriction in the lower vocal tract and this leads to changes in the the vocal tract filter [27, 28].

Many attempts have been made to quantify the amount of breathiness in the voice and a number of quantitative measures have been developed in an attempt to measure breathiness. These measures have been derived from observations and intuitions about the nature of breathy voices:

- **H1:** amplitude of the first harmonic. Due to the more sinusoidal nature of glottal pulses in breathy voices relative to other voice qualities, the amplitude of the first harmonic should be higher.

- **H1-H2:** difference in amplitude between the first and second har-

monics. This measure converts H1 into a relative measure so that the measure is not dependent on gains applied during recording or processing.

- **H1-A1:** difference in amplitude of the first harmonic to the amplitude of the first formant, an indirect measure of first formant bandwidth [29]. It has been observed that breathy voices often have a wider first-formant bandwidth due to the larger glottal opening [30].

- **H1-A3:** difference in amplitude of the first harmonic to the amplitude of the third formant, a measure of spectral tilt. Since breathy voices have a slower rate of glottal closure, there is a larger negative slope to the spectrum of the signal.

- **Noise:** a variety of measures have been developed to quantify the amount of aspiration noise relative to the harmonic content in the voice.

The challenge with using these measures is that it can be difficult to achieve good correlation between the objective measures of breathiness and perceptual ratings of breathiness acquired in listening experiments [31]. It appears that it is possible, with carefully prepared samples and with carefully planned experiments to achieve a significant correlation between these measures [29]. However, in many cases, the results are inconsistent.

Objective measures of breathiness have been improved by taking into account

mechanisms of human perception. For example, one measure that has been developed assumes that breathiness primarily corresponds to the amount that the harmonic content of the voice is masked by aspiration noise, and the objective measure was calculated by passing these quantities through a perceptual model of the hearing process [32, 33]. In the perceptual evaluation of disordered breathy voices, this measure provided a high degree of correlation with perceptual ratings, whereas other measures such as H1-H2, H1-A1 and H1-A3 did not correlate well. Developing techniques to accurately quantify breathiness as perceived in listening experiments is an ongoing area of research [34, 35, 36].

1.1.1 Wider Bandwidth Signals

One of the things observed in the voice samples available in this research is that some high-effort voices exhibit a significant drop-off in frequency response between $4 - 5$ kHz as shown in Figure 1.1. Given that most phonetic analysis of the voice has taken place below approximately 5 kHz, there is little research on this topic. One relevant study uses a physical model of the vocal tract to analyze frequencies above 5 kHz. This study suggests that the cut-off frequency and the suddenness of the drop-off is due to throat constriction in the lower vocal tract [37].

The challenge with analysis beyond 5 kHz is that the acoustic waves in the vocal tract can no longer be assumed to be plane waves because the wavelengths are shorter than the width of the vocal tract. Since the spectral slope of the vocal tract can no longer be considered consistent throughout the frequency range, the drop-off observed in high-effort voice samples is a challenge to standard source-

filter methods. This is unfortunate because musical signals involve frequencies higher than 5 kHz and these frequencies significantly influence the aesthetics of the voice signal.

Most techniques for voice analysis and re-synthesis assume that the voice source is the predominant influence on voice qualities such as breathiness and that the filtering influence of the vocal tract remains relatively consistent. In addition, these techniques of voice analysis do not take into account the drop-off in frequency content that is observed in the samples at hand. This dissertation presents a way to deal with the drop-off when analyzing and resynthesizing the voice in musical applications. The following section provides an outline of the research and the organization of the dissertation.

1.2 Organization

Chapter 2 describes some preliminary thoughts about voice quality and a listening experiment that was carried out to choose between two particular voice terminologies.

Chapter 3 describes how the common implementations of LP result in estimated formant filters that vary with changes to the spectral emphasis of the voice. This chapter describes why the chosen pre-emphasis determines the spectral envelope of the voice source. Although this relationship between the pre-emphasis and the spectral envelope of the glottal source may be known to people with extensive use of LP for voice modeling, it has not been made clear in the literature. Since common

implementations of LP use constant pre-emphasis, the estimated voice source has a constant spectral envelope. This means that the filter estimated by LP captures the variation in the spectral emphasis and this could affect the perception of vocal effort.

The common technique of adding aspiration noise to the voice source implicitly assumes that the voice source is the primary influence on the perception of breathiness and vocal effort and that the estimated LP filter can be ignored. Chapter 4 describes two listening experiments that investigate the influence of the constant pre-emphasis LP filter upon the perception of breathiness and vocal effort. The purpose of these experiments was to verify whether the filters estimated by constant pre-emphasis LP would cause problems in implementing the breathy effect on voices with varying levels of vocal effort.

Chapter 5 presents adaptive pre-emphasis LP (APLP). APLP provides a way to separate changes in the spectral emphasis from the formant filter. Adaptive pre-emphasis has been used with LP, but its relationship to vocal effort and other voice qualities has not been elucidated. Adaptive pre-emphasis is often used to avoid ill-conditioning in fixed point algorithms due to the contrast in spectral slopes between voiced and unvoiced segments [2]. Some LP algorithms use adaptive pre-emphasis to improve speech recognition [38, 39] or accent detection [40].

APLP differs from other traditional techniques of voice source analysis. First, APLP focuses on signals that may not have been recorded in ideal conditions for phonetic analysis. Voice source analysis requires signals that retain phase information and no sound reflections, because the goal is to estimate the shapes of the

glottal pulses in the time domain. Any phase distortion or additional sound reflections will distort the shapes of these pulses. In musical signals, these conditions are not guaranteed. It may not be possible, even in theory, to extract reasonable estimates of the glottal pulses from musical signals, especially in live conditions. The APLP algorithm presented here does not depend upon the ideal retention of phase information.

The second reason why APLP differs from traditional techniques of source analysis is that it has a different goal. In phonetic analysis, the typical goal is to extract the shapes of the glottal pulses and the linguistic content of the voice. Frequencies above 5 kHz are not important for this analysis and are typically not considered. This produces a simpler vocal tract model because the vocal tract filter does not include the drop-off at $4-5$ kHz described above. The adaptive pre-emphasis algorithm presented here analyzes musical voice signals and manipulates them in a way that is musically relevant. In doing so, frequencies above 5 kHz are important; these frequencies influence the aesthetics of the voice signal.

In this dissertation, APLP is presented as a technique to track and manipulate the spectral emphasis of the voice, which influences perception of vocal effort. This spectral emphasis, once estimated, can be manipulated to change the perceived quantity of vocal effort in the voice. The goal is that, by reducing the perceived vocal effort, it will become easier to blend aspiration noise into the voice.

Chapter 6 describes how to use APLP to analyze and manipulate the perceived vocal effort in the voice. After describing the algorithm, a listening experiment is reported to demonstrate that APLP can transform the voice more effectively than

constant pre-emphasis LP.

The technique involved in APLP can be used during voice analysis as an indication of the perceived vocal effort in the voice [41]. Since vocal effort is influenced by a person's emotional state, this technique can be used to analyze the stress in a person's voice, which is a useful application in its own right. In a further application, the filters extracted with APLP can be manipulated to synthesize new voices with different levels of vocal effort and correspondingly different emotional states.

Aperiodic analysis and synthesis is capable of modifying the perceived vocal effort [42]. The type of vocal effort presented in aperiodic analysis and synthesis is different from the type of vocal effort manipulated by APLP in this dissertation. In the aperiodic synthesis, the perceived vocal effort is primarily modified by increasing variation in the aperiodic component. Increasing variation allows the production of voices with more roughness or harshness. This roughness is associated with vocal effort. However, APLP as presented here focuses on transforming voices that do not sound rough or harsh. In the absence of these vocal aperiodicities, vocal effort is, for the most part, influenced by changing the spectral emphasis.

This dissertation presents some discoveries about voice quality and about voice modeling using LP. The most significant contribution of this research is that LP, as commonly implemented with constant pre-emphasis, does not appropriately model the operation of the voice. When modeling ranges of voice qualities between high-effort and breathy voices, one needs to estimate a voice source with a spectral slope that follows the variations in the voice. However, constant pre-emphasis LP

estimates a voice source with an unchanging spectral envelope. This dissertation presents a solution to that problem using APLP to transform the voice effectively. The following chapter describes how to estimate a source-filter model of the voice using LP.

Chapter 2

Preliminary Exploration of Voice Quality

This chapter describes a preliminary investigation into the choice of terminology to describe non-breathy voices. The original intuition in this research was that the breathy effect does not work on constricted voices. This thought was inspired by some phonetic research that examines the mechanisms of phonation in a more complex way than the typical source-filter concept of voice modeling.

In source-filter modeling, it is typically thought that the vocal folds remain at a fixed location in the throat, with the mode of phonation (modal, breathy, harsh, creaky, etc. [12]) determined primarily by the tension in various directions in the vocal folds. However, the mechanism of phonation involves more than just the vocal folds. There are other folds above the vocal folds (aryepiglottic folds) that can constrict the flow of air, resulting in different voice qualities. Researchers in

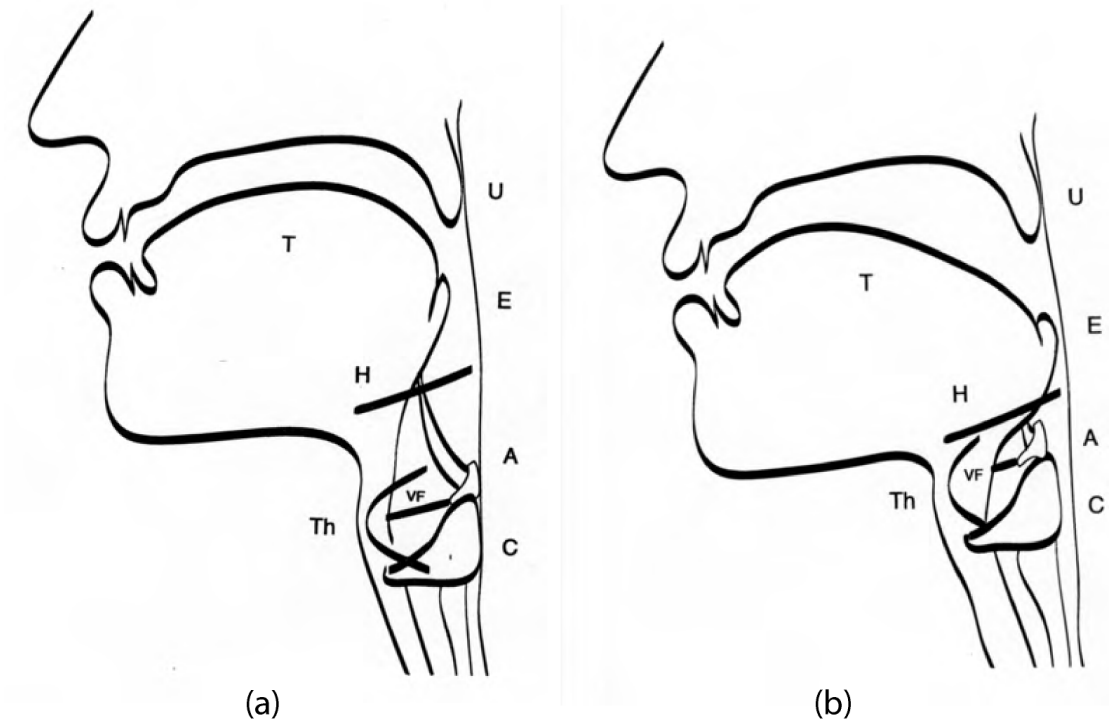


Figure 2.1: Two degrees of laryngeal constriction: (a) larynx in neutral position, (b) almost complete laryngeal constriction, with a narrowed aryepiglottic passage, shortened vocal folds, extreme larynx raising, and extreme tongue retraction. Labeling: T = tongue, U = uvula, E = epiglottis, H = hyoid bone, A = arytenoid cartilage, Th = thyroid cartilage, C = cricoid cartilage, AE = aryepiglottic folds, and VF = vocal folds. Used with permission [43].

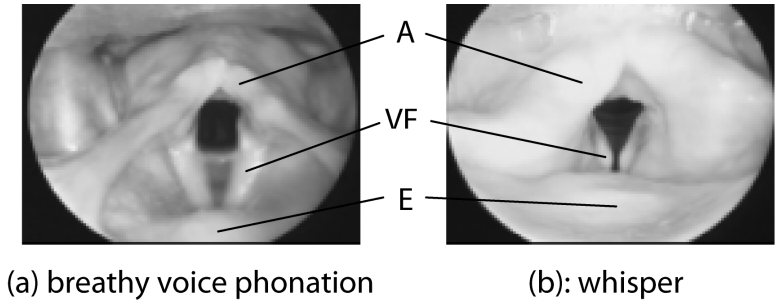


Figure 2.2: Two articulatory postures of the laryngeal articulator: A = arytenoid cartilages, VF = vocal folds, and E = epiglottis. Used with permission [43].

linguistics have been working to develop a map of these different voice qualities [25, 26], taking into account the influence of the aryepiglottic folds and other parts of the lower vocal tract. These constricted configurations come into play for some of the harsher voice qualities. Constriction in the lower vocal tract can change what would otherwise be a modal voice (i.e., a neutral voice) into a pressed voice or a harsh voice. During this constriction process, the larynx (the voice box) moves upwards and compresses the aryepiglottic folds as illustrated in Figure 2.1. The air pathway becomes constricted so that only a small gap remains for the air to escape. With large amounts of constriction, the vibrations in the lower vocal tract become aperiodic. This is known as a harsh voice and it can include vibration of aryepiglottic folds as well as the vocal folds. Some of these same mechanisms are involved in to a subtle degree during whispering as seen in Figure 2.2.

A whispery voice can result when applying the breath effect to a high-effort voice. To convert a high-effort voice into a breathy voice, it is not enough to add aspiration noise to the voice source. When aspiration noise is added to high-effort voices, the resulting voice does not sound like a typical breathy voice because it still exhibits effort. One obtains a voice that simultaneously exhibits effort and aspiration noise. If the artificial noise perceptually blends with this voice that exhibits some effort, the result is a whispery voice [25, 26]. An abstract representation of this transformation is presented in Figure 2.3. Alternately, transforming the spectral envelope of the high-effort voice into that of a breathy voice without adding noise yields a voice that sounds lax and unnatural. It gives the perception that the vocal folds are relaxed, but the aspiration noise that our ears expect to

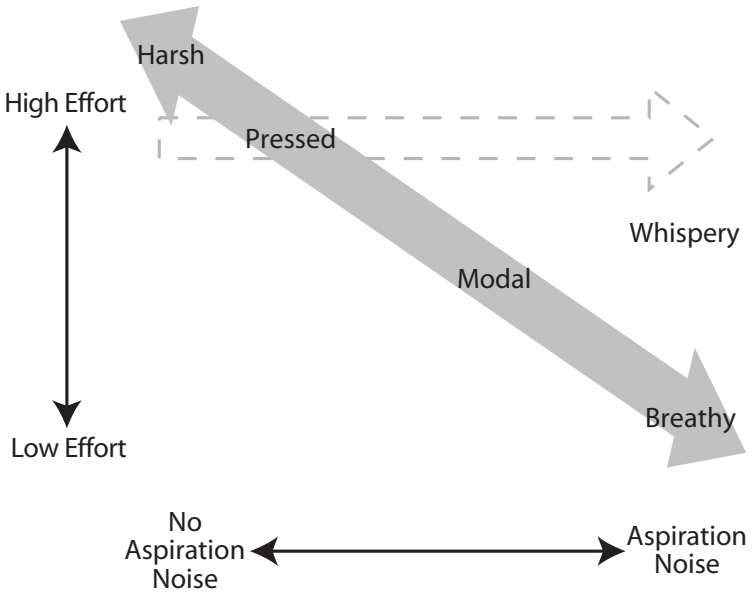


Figure 2.3: An abstract representation of various voice qualities on a continuum between pressed and breathy voices. The dashed arrow represents the result of adding aspiration noise without reducing the perceived vocal effort.

hear is missing.

Many of these terms are subjective and it can be difficult to find the appropriate terminology. In the early stages of the research, a voice conversion experiment was carried out that yielded twenty voice samples. This experiment was a preliminary version of the experiment described in detail in Section 4.1. Half of the samples were unmodified and the other half were modified through a voice conversion algorithm. In the experiment, a linguistics expert evaluated the voice samples relative to a benchmark according to perceived constriction, vocal effort and breathiness.

These evaluations were made on a scale from -5 meaning much less constriction to $+5$ meaning much more constriction.

This was just a preliminary experiment and some of the samples exhibited too many artifacts, but there was an interesting result. As expected, there was a negative correlation between breathiness and voice constriction: $-.39$. Also as expected, there was a positive correlation between constriction and vocal effort: 0.44. Surprisingly, there was an extremely strong negative correlation between breathiness and vocal effort: -0.98. This seems to indicate that vocal effort is better than constriction at describing voices opposite to breathiness. The results of this experiment indicated that it might be easier to work with the vocal effort terminology.

Regardless of the choice of terminology, the research into voice constriction raised a question. Does constriction in the lower vocal tract influence the performance of the breathy effect? In terms of voice modeling, the corresponding question might be: does the estimated vocal tract filter influence the performance of the breathy effect? Experiments presented later in this dissertation will examine this question. The following chapter introduces linear prediction (LP) as a technique for modeling the vocal tract.

Chapter 3

Linear Prediction and the Source-filter Voice Model

The approach taken in this study is to use a source-filter model of the voice (Figure 3.1) estimated by LP [44]. Linear prediction is the most common method of decomposing a voice into a source and a filter and is used extensively for both phonetic analysis and voice compression. In addition, IVL Technologies and TC-Helicon use LP in their commercial voice processing products. This chapter describes the operation of LP for voice analysis.

Linear prediction is well suited to the analysis of the voice, estimating a filter that behaves in a manner similar to the filtering influence of the vocal tract [45]. However, the linear model is not perfect [46]. Some interactions occur between the source and the filter [24]. Additionally, it is difficult to verify the appropriate separation between source and filter for a given voice, because the required mea-

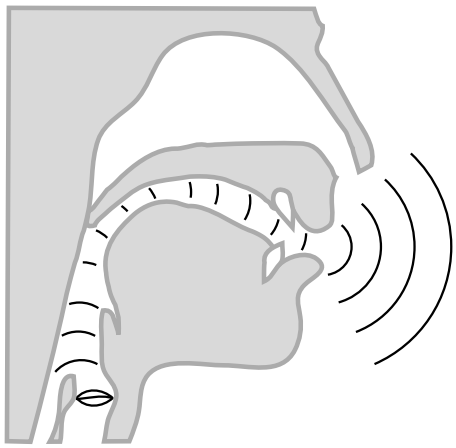


Figure 3.1: The voice can be viewed as a source and a filter. The pressure waves originating at the vocal folds provide the glottal source. The vocal tract filters these pulses resulting in resonances that correspond to the vowel sounds.

surements interfere with the operation of the voice. Despite these challenges, the source-filter model provides a good perceptual approximation to the vocal tract and is widely used for voice analysis and synthesis [47].

When a signal is fed into LP, LP estimates a filter that matches the spectral envelope of the signal. When the signal has been appropriately pre-emphasized, this estimate is a reasonable approximation of the filtering influence of the vocal tract. In phonetic research, a significant number of studies have used LP to extract glottal pulses from voice signals. Either these studies focus on working with carefully recorded voice signals or use artificially synthesized voice signals. In the case of artificially synthesized voices, the goal is often to use LP to extract the artificial source that was originally used to create the samples. If the artificial source can be recovered, this is an indication that LP could also work on real voice samples.

With careful preparation of the experiments using artificially synthesized voices,

LP is effective in separating the source and filter of the voice [48]. However, in the case of natural voices, it is not possible to verify whether the true source has been extracted. Neither is it possible using today's technology to accurately measure the true glottal source from the acoustic signal alone. Perhaps the most accurate measurement technique uses an electroglottograph, which measures the the electric potential across the vocal folds as they come into contact with each other, thereby providing detailed information about the nature of the contact. However, the glottal excitation of the voice is primarily caused by the dynamics of the airflow through the opening of the vocal folds, and the electroglottograph provides more information on the contact than the opening. This means that the electroglottograph provides only a secondary measurement of airflow. Using artificially synthesized vocal tract models, investigators using LP have extracted reasonable estimates of the glottal pulses, but it is not possible to verify whether this accuracy transfers to natural voices.

Investigators using LP can estimate a series of constant-diameter tubes corresponding to the cross-sectional areas of the vocal tract [49]. The number of tubes corresponds to the LP order. For a typical vocal tract, there are approximately twenty constant-diameter tubes concatenated together, so the spacial resolution is low. This series of tubes roughly corresponds to the cross-sectional areas of the vocal tract in that the tubes closer to the vocal folds are smaller while the tubes closer to the throat are larger. However, multiple configurations of tubes are capable of producing a similar vocal tract filter. Observing the estimated tube model in action, illustrates that the acoustic tube model does not result in a stable

estimate of tube sections. As the poles of the vocal tract filter estimated by LP move around, the diameters of the tubes suddenly change in a way that is not phonetically realistic. This happens when the poles of the filter suddenly swap. For example, two poles may be used to estimate a lower formant and one pole for a higher formant. Then, as the vocal waveform changes, suddenly one of the poles jumps from one formant to the other. Hence, a discontinuity forms in the model.

Another disadvantage of estimation of acoustic tubes is that it does not take into account the branching of the vocal tract into the nasal cavity. While the tube model corresponds to an all-pole filter, the branch corresponds to a zero in the transfer function of the vocal tract. The LP algorithm does not take this zero into account. It is possible to implement a method of analysis that includes zeros using Autoregressive Moving Average (ARMA) LP [50, 51]. However, this technique is not widely used because it is computationally more complex; because it is possible to take zeros into account by using a higher-order all-pole model; and because all-pole models have been found to work effectively in practical applications.

Considerable work has been carried out to interpret LP as a physical model of the voice. The results have been mixed since the LP filter does not represent precisely the physiology of the voice, that is, the estimated tube diameters are not accurate. However, LP can provide a reasonable approximation of the frequency response vocal tract filter. With careful preparation, LP can be used to obtain realistic estimates of glottal pulses. Accordingly, LP is thought of as a quasi-physical model of the voice. The model does not perfectly correspond to the voice, but it is sufficiently accurate to provide inspiration for further development.

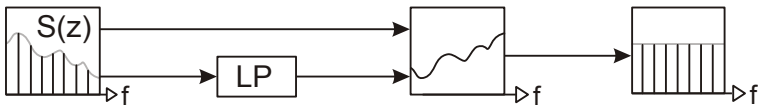


Figure 3.2: Linear prediction used to extract an excitation with a flat frequency response.

The physical interpretation of LP is part of the rationale for using adaptive pre-emphasis, which will be presented in Chapter 5.

Perhaps it is best to think of LP as a technique to model the spectral envelope of the voice. Linear prediction estimates an all-pole filter that fits the spectral envelope of the signal it receives. If one takes the original signal and inverse filters it to remove the spectral envelope, the result is an ideally flat excitation, as seen in Figure 3.2. The earliest voice models with LP used a formant filter, estimated by LP and a flat excitation, either an impulse train for voiced sounds or white noise for unvoiced sounds.

The true voice does not have a flat excitation. Instead, a linear model of the voice is illustrated in Figure 3.3(a) where:

- $G(z)$ = glottal excitation.
- $V(z)$ = influence of the vocal tract filter.
- $L(z)$ = influence of lip radiation.
- $S(z)$ = resulting spectrum of the voice.

To make LP correspond more closely to the physical voice, a pre-emphasis is

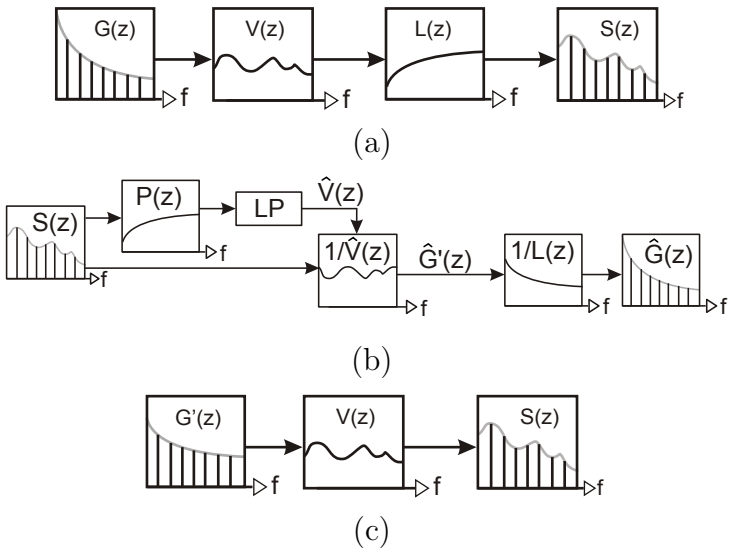


Figure 3.3: (a) Linear model of the voice. (b) Using LP to estimate the vocal tract filter, $\hat{V}(z)$, and the glottal source, $\hat{G}(z)$. (c) Simplified linear model of the voice where removing lip radiation is considered equivalent to taking the derivative.

typically applied as seen in Figure 3.3b. This pre-emphasis, when appropriately chosen, ensures that the estimated glottal spectrum, $\hat{G}(z)$, will have a spectral slope that, on average, represents what would be expected according to voice physiology.

The glottal signal is the flow of air beyond the glottis, which is the space between the vocal folds. This glottal signal is also known as the volume-velocity wave. The features of the glottal pulses can be seen more clearly when examining, $G'(z)$, also known as the derivative volume-velocity wave. For this reason, voice researchers, rather than working with $G(z)$, prefer to work with $G'(z)$. Using $G'(z)$ simplifies the model of the voice, as seen in Figure 3.3(c). This simplification is possible because $L(z)$ represents the equivalent of taking the derivative [52].

The LP technique fits an all-pole filter to the spectrum of the signal. The

all-pole filter is of the form:

$$\hat{V}(z) = \frac{1}{A(z)}, \tag{3.1}$$

where $A(z)$ is an all-zero filter and $\hat{V}(z)$ is an estimated vocal tract filter given by:

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k} \tag{3.2}$$

The order of the filter is defined by p. The operation of the LP algorithm [1] and its relation to the human voice have been thoroughly described in the literature [2].

3.1 Fixed-Rate and Closed-Phase LP

Several techniques allow computation of LP and the two most common techniques are fixed-rate autocorrelation LP and closed-phase covariance LP. The primary difference between these techniques is that fixed-rate LP analyzes a window of the voice signal over several glottal pulses, whereas closed-phase LP finds the spaces between the glottal closure instants and analyzes that portion of the signal using covariance LP.

For phonetic analysis, closed-phase LP is most often used. closed-phase LP provides the most realistic estimation of the glottal pulses, operating over the period where the assumptions underlying LP correspond most closely to the configuration of the vocal tract. This is because during the closed phase, the vocal tract can be modeled as a series of acoustic tubes with one end closed [49]. During the open

phase, the glottis is open and the trachea below the vocal folds acts as an additional resonator. In addition, the instant of glottal closure introduces an impulsive burst of energy into the voice signal that yields errors in the estimation of the LP coefficients.

In spite of the advantages of closed-phase LP, this technique is not appropriate for the current context. Closed-phase analysis requires that voices be recorded in a way that retains phase information. This is not always possible for an algorithm designed to manipulate singing voices in a musical context. In addition, in breathy voices the vocal folds are relaxed and may not have a significant closed phase. Lastly, closed-phase LP is less robust; the algorithm stops working when the glottal closure detection breaks down. For these reasons, autocorrelation LP is more appropriate in this context.

In summary, LP is the most widely used technique for source-filter analysis of the voice. It is not perfect but it can provide a reasonable estimation of the vocal tract filter and the corresponding glottal source. In the current application, autocorrelation LP is more appropriate than closed-phase LP, even if it deviates a little from the ideal methods used in phonetic analysis. Autocorrelation LP is more effective in analyzing practical musical signals and is more robust. The following chapter will discuss how various voice qualities appear in the source-filter model of the voice.

Chapter 4

Perceptual Investigation of Constant Pre-Emphasis Linear Prediction

The typical way to add breathiness to singing voices is to modify the estimated voice source by adding aspiration noise. However, high-effort voices are difficult to transform with the breathy effect because they retain the perception of high effort. Before setting out to improve the breathy effect, it is necessary to determine where the perception of effort originates. In the separation of source and filter, is the perception of effort primarily associated with the estimated source or the estimated filter? This chapter describes two experiments carried out to gain a better understanding of where the perception of breathiness and vocal effort arise in the source-filter model of the voice.

In the first experiment, two voices were decomposed into sources and filters using constant pre-emphasis LP. The sources were then exchanged and the voices were resynthesized as seen in Figure 4.1. The purpose of this experiment was to determine whether the source or the filter is more influential in the perception of breathiness and vocal effort.

In the second experiment, two voices were again decomposed into sources and filters. The filters were then excited with an artificial source. The purpose of this experiment was to determine how the filters influence the perception of breathiness and vocal effort. The benefit of this experiment is that it removes the confounding influence of the source, making the results more clearly explainable. Both of these experiments demonstrate that the vocal tract filter estimated by constant pre-emphasis LP does have a significant influence on the perception of breathiness and vocal effort.

4.1 Voice Conversion Experiment

A voice conversion [6, 7, 53] experiment was carried out to determine whether constant pre-emphasis LP estimates filters that capture some of what is perceived as vocal effort. The presented voice conversion technique was used to understand particular components of the voice quality without having to model all of the components in detail. The point of this evaluation is to determine whether the breathy effect is confined to the LP residual or whether some components of perceived breathiness are found within the estimated vocal tract filter.

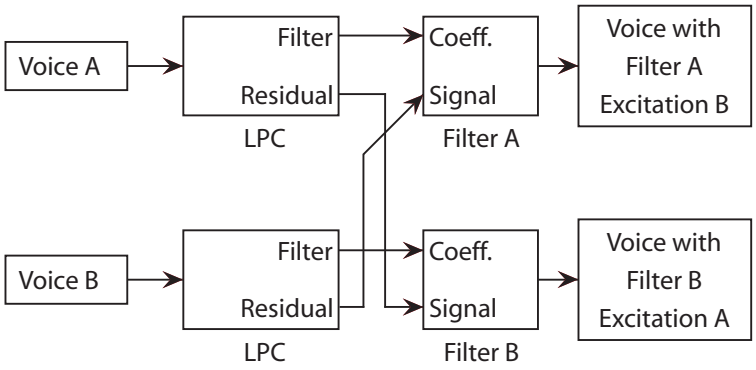


Figure 4.1: LP voice conversion concept.

The concept of the voice conversion algorithm is presented in Figure 4.1. A breathy and a non-breathy voice sing the same phrase with the same timing. The LP filter computed for each of these voices is depicted in Figure 4.2. The voices are then inverse filtered to extract the residual as seen in Figure 4.3. The LP residual from the breathy voice is then fed through the LP filter from the non-breathy voice. Likewise, the LP residual from the non-breathy voice is filtered by the LP filter from the breathy voice. Ideally, the synthesized voice should assume the glottal characteristics of the LP residual. The voice that was originally non-breathy should become breathy when given a breathy excitation. Likewise, the voice that was originally breathy should become non-breathy when it is given a non-breathy excitation.

4.1.1 Linear Prediction Modeling

Three pairs of voice samples were used in the experiment, collected from a variety of different sources. Some of them were available from previous experiments [54]

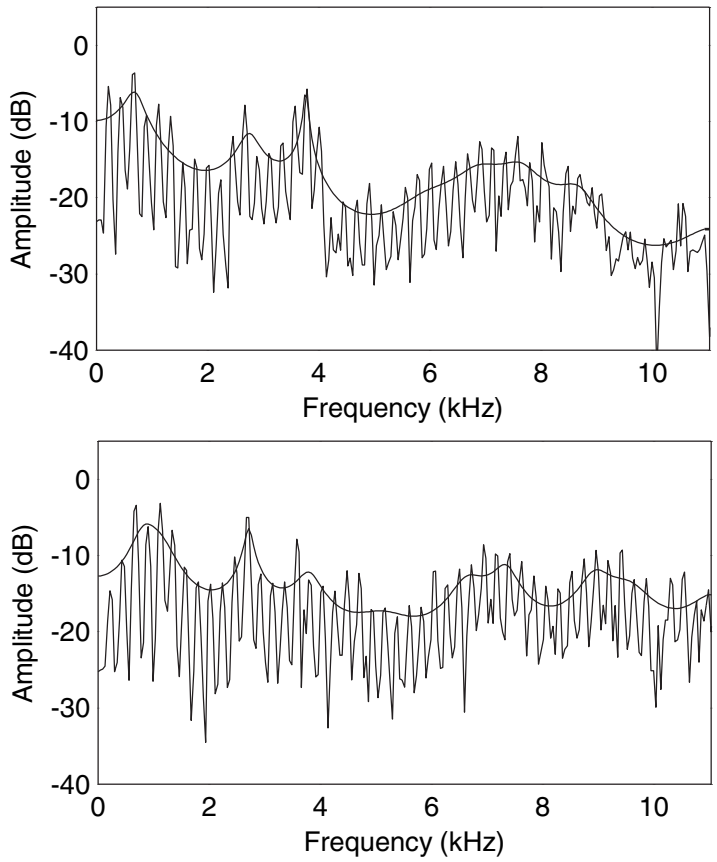


Figure 4.2: LP filters from a breathy voice (top) and a non-breathy voice (bottom). Both signals have been pre-emphasized.

while others were newly recorded. The ideal samples were those recorded by one person singing or speaking the same vowel with a breathy and non-breathy voice. The voices were recorded at a sample rate of $22050Hz$, which was chosen as a compromise between having enough bandwidth to capture the breathy quality and a low enough sample rate for LP to model the spectrum well.

During this early experiment, the LP algorithm was chosen to have an order of 20 because this corresponds to a typical vocal tract length of 15 cm long when LP

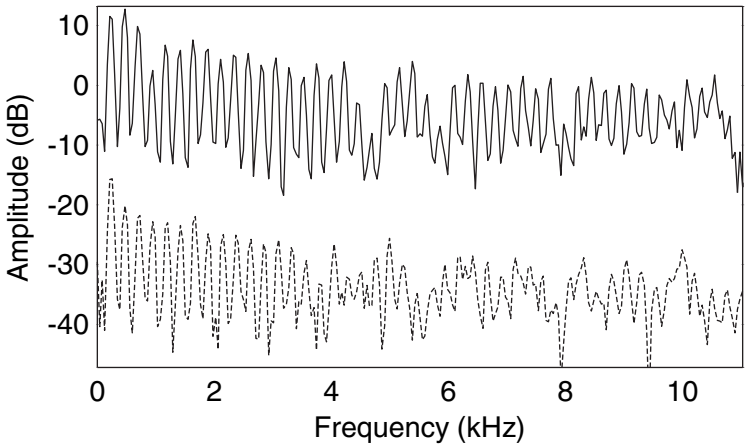


Figure 4.3: LP residuals from a breathy voice (dashed) and a non-breathy voice (solid). An arbitrary vertical offset has been applied for visualization. Pre-emphasis is included in this plot residual.

is interpreted as a series of concatenated, acoustic tubes [2]. The voice signal is pre-emphasized with a high-pass filter: $(1 - 0.98z^{-1})$. This pre-emphasis flattens the spectrum of the signal, making it easier for LP to fit the signal. Theoretically, the LP residual corresponds to the volume-velocity wave of the glottis if the pre-emphasis filter is appropriate.

4.1.2 Perceptual Testing

The results of the experiment were evaluated with the help of a linguistics expert. A preliminary test showed that it was difficult to achieve clear ratings with isolated samples. Therefore, the test was designed to measure the relative difference between a benchmark sample and the other samples. This approach has been used previously for evaluating breathy voices [29]. For each set of four samples, one of

the original samples was chosen as a benchmark by which the other corresponding samples were evaluated. The evaluator was not told how each sample was generated or whether the sample was natural or synthesized. The comparison samples were randomized.

The perceptual criteria for this test was drawn from other studies for evaluating breathy voices [48, 29]. The parameters from these tests were breathiness, naturalness, vocal effort and nasality. Some other parameters were also added in an attempt to gain a deeper understanding of the perceived configuration of the voice. The parameters included:

- Breathiness:

  (-5 = much less breathy, 0 = no change, 5 = much more breathy)

- Vocal effort:

  (-5 = much less vocal effort, 0 = no change, 5 = much more vocal effort)

- Nasality:

  (-5 = much less nasal, 0 = no change, 5 = much more nasal)

- Constriction above the glottis:

  (-5 = much less constriction, 0 = no change, 5 = much more constriction)

- Velarization:

  (-5 = much more velarized, 0 = no change, 5 = much less velarized)

- Creakiness:

  (-5 = much less creaky, 0 = no change, 5 = much more creaky)

Unnaturalness was evaluated separately, without a benchmark, to get a sense of whether the synthesized samples were close to the original samples in quality. Naturalness was defined as human-sounding.

The evaluation was carried out by Dr. John Esling, a professor in linguistics at the University of Victoria. Esling's research investigates different sound production mechanisms within the voice [25, 26]. He has a detailed understanding of the physiology of the voice mechanism and an experienced ear for detecting different voice qualities. The use of an expert listener reduces the risk inherent in the small sample size. However, the test should be repeated with a larger sample size to achieve more broadly applicable results.

4.1.3 Analysis of Perceptual Ratings

Factorial analysis [55] was carried out on the test data as shown in Figure 4.4. Differences in measures of constriction and velarization were not statistically significant. The most significant responses were for breathiness, vocal effort, unnaturalness, and creakiness. Creakiness and vocal effort were highly correlated but vocal effort had a larger range. Nasality was rated differently for different vocal tracts and did not change greatly with the excitation, as shown in Figure 4.4d.

The interaction plot for unnaturalness is found in Figure 4.4c. The most obvious observation from this plot is that the original samples sound more natural than the samples with swapped excitations. This is to be expected. However, it also raises the issue of whether unnatural sounds may have been a distraction in the evaluation.

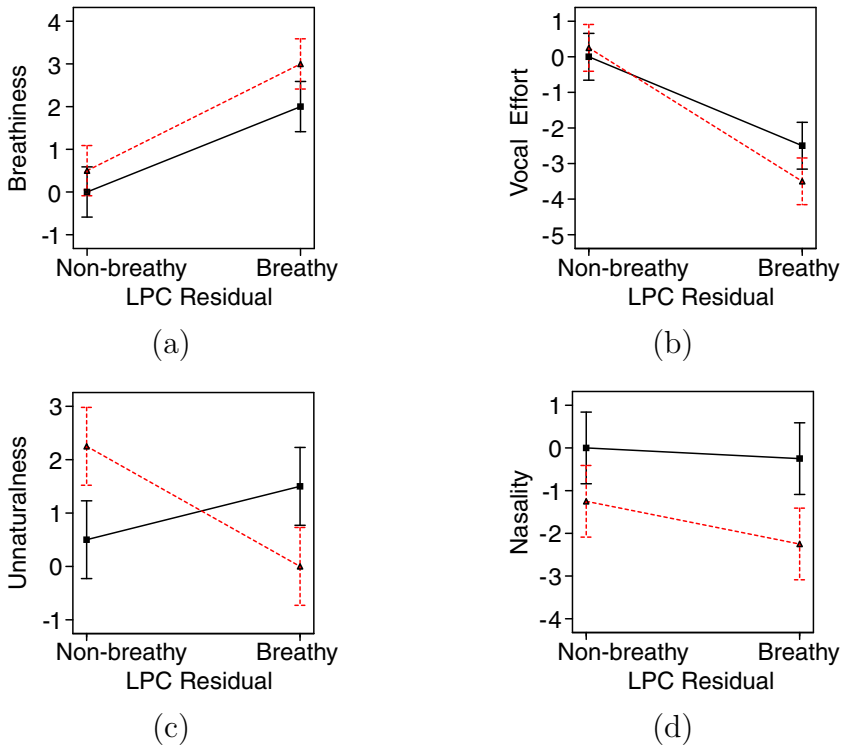


Figure 4.4: Interaction plots for (a) perceived breathiness, (b) perceived vocal effort, (c) perceived unnaturalness, and (d) perceived nasality. The horizontal axis represents the LP residual. The dotted lines represent data from the breathy LP filter. The solid lines represent data from the non-breathy LP filter. The 95% confidence intervals are also plotted.

The interaction plot for breathiness in Figure 4.4a shows a large increase in perceived breathiness when the LP residual from a breathy voice is fed through the LP filter for a non-breathy voice. As well, the newly synthesized voice does not achieve the same level of breathiness as the original breathy voice.

A similar phenomenon occurs in the interaction plot for vocal effort, but in reverse, in Figure 4.4b. Vocal effort negatively correlates with breathiness. When a breathy LP residual is fed into a non-breathy LP filter, the perceived vocal effort

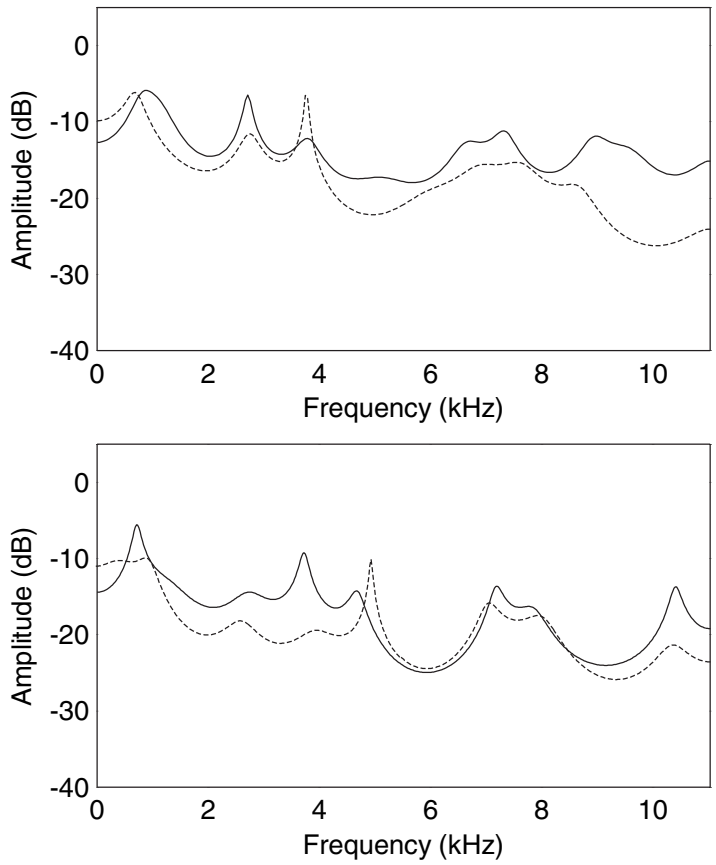


Figure 4.5: Constant pre-emphasis LP formant filters from the voice conversion experiment (male). Vocal tract filters for a man singing /ah/ at 210 Hz (top), and at 111 Hz (bottom). Solid line is non-breathy. Dotted line is breathy.

goes down. Again, the vocal effort does not go all the way to the level of the original breathy voice.

The breathy LP residual achieves most of the transformation but the transformation is not complete. The LP filter must account for some of the perceived breathy effect. Looking at the LP filters, one sees significant differences between breathy and non-breathy filters, even when the same voice is singing the same

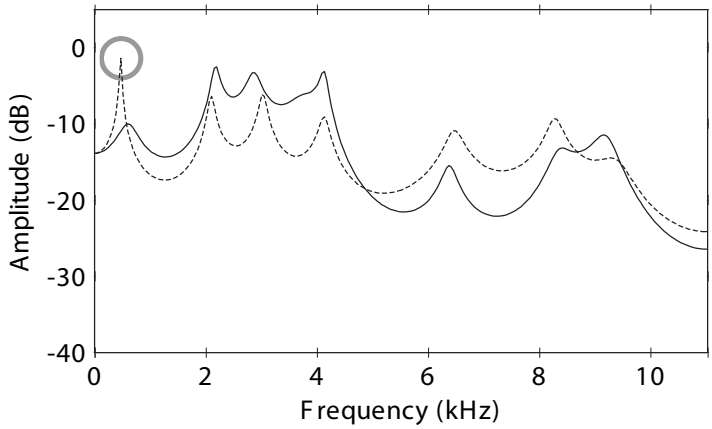


Figure 4.6: Constant pre-emphasis LP formant filters from the voice conversion experiment (female). Vocal tract filters for a woman singing /ay/. Solid line is non-breathy. Dotted line is breathy. The circled resonance lead to a distracting $500Hz$ artifact.

vowel at the same fundamental frequency (Figure 4.5 and 4.6).

Some artifacts were present in some of the synthesized data and they may have affected perceptions of breathiness in the cross-synthesized samples. The voice rated the most unnatural was a non-breathy LP residual fed into a breathy vocal tract. It sounded like a sine wave overlaid on the voice at approximately $500Hz$. The frequency of this artifact was confirmed by removing it with a narrow band filter. The artifact was generated by a large resonance in the breathy LP filter, as seen in Figure 4.6. The prevalence of the artifact might be reduced by using bandwidth expansion [56] to widen the peaks in the LP filters.

4.1.4 Discussion of the Voice Conversion Experiment

An attempt was made to convert a non-breathy voice into a breathy voice. The LP filter from a non-breathy voice was excited by the LP residual from a breathy voice. The consequent voice quality was not as breathy as the original breathy voice. This indicates that the perception of breathiness involves more than the LP residual. This phenomenon was analyzed with a factorial analysis experiment and the result was consistent. A breathy LP residual is not capable of fully transforming a non-breathy voice to a breathy voice. As expected, the perception of vocal effort was found to be inversely correlated with breathiness. However, the experiment should be repeated with more evaluators to gain greater confidence in the results.

Artifacts were present in some of the synthesized voices. This was partially due to peaky resonances in the LP filters due to poor modeling. For clearer results, these artifacts should be avoided before repeating the test.

The above algorithm is useful for examining the perceptual influence of different source-filter models. The source-filter models can be investigated without having to explicitly model the glottal pulses and aspiration noise. The greatest opportunity with this technique is to understand better how the vocal tract filter may affect the perception of different voice qualities. In this way, the LP modeling of breathy voices can be better understood.

The conclusions obtained from this experiment looked promising. However, the test was limited by having only one listener. Through doing this experiment, I thought of a way to evaluate more clearly the influence of the filter estimated by LP. By using the same excitation for two LP filters, one can more clearly see the

influence of the LP filter upon breathiness and vocal effort. The filters representing breathy and non-breathy vocal tracts were examined and found to be significantly different.

4.2 Artificial Excitation Experiment

The previous experiment demonstrated that the LP filter influences the perception of breathiness and vocal effort. However, one of the challenges of that experiment was that the the data were complex to interpret in that they involved both the source and filter. In addition, the previous experiment involved only one expert listener. The next experiment involves more listeners. The purpose of this experiment was to demonstrate more clearly that the LP filter influences the perception of breathiness and vocal effort. This was accomplished by using an artificial excitation to excite LP filters extracted from high-effort and breathy voices.

The LP filter captures changes to the spectral envelope that affect the perception of breathiness and vocal effort. This means that the estimated formant filter captures characteristics of the source. This can lead to problems when attempting to model the voice because the variation in the tilt of the source has instead been modeled by the estimated formant filter. For example, one can attempt to make a voice sound breathy by adding aspiration noise, but it becomes difficult to know how to change the spectral envelope of the source without having an estimate of the source envelope.

In the existing source-filter concept of the voice, it is generally assumed that the

filter does not influence the perception of breathiness. We evaluate this assumption through a listening experiment. Synthesized samples were created to have identical glottal sources and different vocal tracts. The fundamental frequency and the vowel remained constant for these samples. This ensured that any difference in the perceived voice quality would be due to the vocal tract filters. We used linear prediction to extract the voice filter.

According to the source-filter paradigm, the perception of breathiness and vocal effort should be primarily controlled by the glottal source and be little affected by the formant filter. This experiment investigates whether the formant filter estimated by LP can influence the perception of breathiness and vocal effort. The experiment starts with a pair of voice samples. One sample exhibits high-effort and the other sample exhibits breathiness. Linear prediction estimates a filter and residual for each sample. The influence of the residual is eliminated by providing both filters with the same artificial source during resynthesis. The synthesized samples differ only according to the difference between the two filters. Seven people evaluated three pairs of samples in listening tests. The results demonstrate that LP filters influence the perception of breathiness and vocal effort. When a voice changes from breathiness to vocal effort, the spectral envelope changes. This change is captured by the LP filter rather than by the residual. A closer look at the LP algorithm provides an explanation.

4.2.1 The Liljencrant-Fant model

The differences in the shapes of the glottal pulses can be seen by looking at some standard settings for the Liljencrant-Fant (LF) model [57]. The LF model provides time-domain pulses that represent the derivative of glottal flow. This model is the most popular for analyzing and synthesizing the glottal source. The LF model of the derivative glottal wave is described by the following formulas and is plotted in Figure 4.7:

$$\begin{aligned} g'(t) &= E_o e^{\alpha t} \sin(\omega_g t), \quad 0 \le t \le T_e \\ &= \frac{-E_e}{\varepsilon T_a}(e^{-\varepsilon(t-T_e)} - e^{-\varepsilon(T_c-T_e)}), \quad T_e \le t \le T_c \le T_o \end{aligned} \tag{4.1}$$

where:

T_o = period of the fundamental frequency

T_e = time of the glottal closure instant (GCI)

T_c = time of complete closure

T_a = "time constant" of the exponential decay

E_o = amplitude scaling of the sine wave

α = controls how much the envelope of the sine wave is skewed to the right

ω_g = π/T_p = wavelength of the sine wave

E_e = amplitude of negative peak in the glottal derivative wave

ε = rate of exponential decay during glottal closure

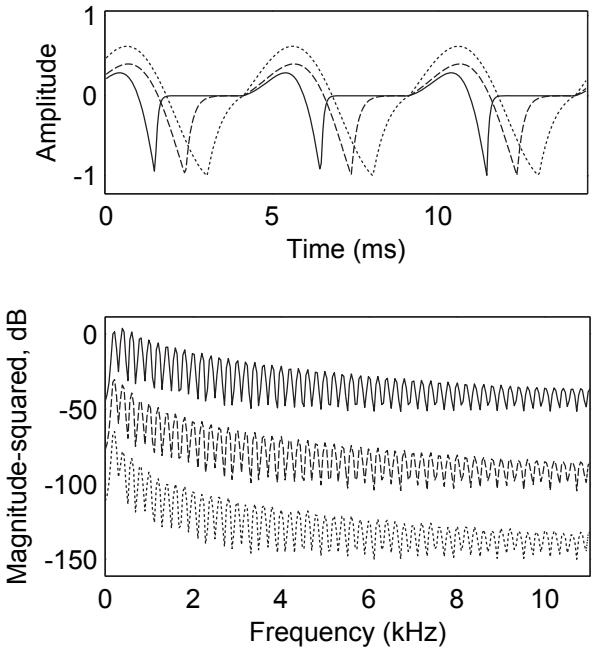


Figure 4.7: The LF model creates a pulse train representing the derivative of the glottal flow (top). Three voice types are represented: breathy (dotted line), modal (dashed line), and high-effort (solid line). Different pulse shapes result in different spectral slopes (bottom). The frequency spectra have been vertically offset for clarity.

Not all of these parameters are independent of one another. For example, there has to be continuity between the sine and exponential portion of the wave. In addition, the area above and below zero have to be equal to avoid drift if the derivative is integrated. Applying these constraints involves an optimization algorithm [58]. After the various constraints have been applied, there are five independent control parameters. However, controlling the shape of the derivative glottal wave with these five parameters isn't necessarily intuitive. For this reason, three dimensionless parameters have been developed to provide more intuitive

control [59]:

- $R_a = T_a/T_o$ = relative length of the return phase. Influences spectral tilt.
- $R_g = T_o/2T_p$ = a measure relating to the rise time. R_g increases with a shortening of the rise time.
- $R_k = (T_e - T_p)/T_p$ relative duration of the falling branch from the peak in the glottal wave at T_p to the discontinuity at T_e.

Another dimensionless parameter describes the wave shape:

$$R_d = (\frac{U_o}{E_e})\frac{1}{(110T_o)} \tag{4.2}$$

R_d describes a range of voice qualities between breathy voices with a high open-quotient ($R_d = 0.5$) to a neutral, modal voice ($R_d = 1$) to voices with a small open-quotient ($R_d = 2$). Fant also developed a mapping by which R_d can control R_a, R_g, and R_k [59]. Figure 4.7a illustrates the differences in the glottal pulses between a breathy and a high-effort voice. The corresponding differences in the frequency spectra have also been plotted in Figure 4.7b. This clearly demonstrates that a breathy source and a high-effort source each have a different frequency spectrum.

The primary concern was for the LF model to sound natural. In this experiment, R_d values between 0.5 and 0.8 worked well. A reasonable comparison between the LP filters can be obtained as long as the R_d parameter is kept identical for the sample pairs in the comparison (Figure 4.8).

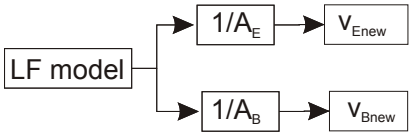


Figure 4.8: Artificial excitation for the experiment. The synthesized pair of voices were generated with the same artificial source using an LF model. The LP filters were extracted from the high-effort voice ($1/A_E$) and the breathy voice ($1/A_B$) using the same pre-emphasis filter. Any difference between the synthesized voices is due to differences in the LP filters.

4.2.2 Experiment setup

One way to evaluate the influence of the formant filter is to take two different formant filters and supply them with the same glottal source. In this situation, the only difference between the resulting synthesized voices is the filter. If the vocal tract filter does not influence the perception of breathiness, then both voices should be perceived to have the same amount of breathiness. If the formant filter does influence breathiness, then a difference will be observed. The process for creating and evaluating the samples has a number of steps:

1. Start with two samples in which the same person sings the same vowel at the same fundamental frequency but with differing voice qualities: high effort voice (V_E) and breathy voice(V_B).

2. Use LP (Figure 3.3(c)) on each voice to estimate filters ($1/A_E$ and $1/A_B$).

3. Excite the filters ($1/A_E$ and $1/A_B$) with an LF model (Figure 4.7) plus noise to generate synthesized voices: V_{Enew} and V_{Bnew}. Since the source is the

same for both voices, any difference between the voices will be due to the filters (see Figure 4.8).

4. Carry out a listening test evaluating the difference between the two filters.

   (a) Rate the relative difference in breathiness between the the original voices: V_E w.r.t. V_B.

   (b) Rate the relative difference in breathiness between the the synthesized voices: V_{Enew} w.r.t. V_{Bnew}.

   (c) A rating of zero indicates that there is no difference between V_{Enew} and V_{Bnew}, indicating that the filters ($1/A_E$ and $1/A_B$) do not influence the perception of breathiness. A non-zero rating indicates that the filters do influence the perception of breathiness. See Figure 4.9 for the results.

5. Repeat steps 4(a-c) for vocal effort.

4.2.3 Algorithm details

The voices were recorded at a sample rate of $22050Hz$, which was chosen as a compromise between having enough bandwidth to capture the breathy quality and a low enough sample rate for LP to model the spectrum well. Three pairs of voice samples were used in the experiment, collected from a variety of different sources. Some of them were available from previous experiments [54] while others were newly recorded. The characteristics of the extracted vowels are summarized in Table 4.1.

Table 4.1: Original voice samples for constant pre-emphasis LP experiment

Singer	Female A		Male A		Male B	
Tessitura	mezzo-soprano		baritone		baritone	
Vowel	[e]		[a]		[a]	
Note	A#3		G#3		A2	
Phonation	breathy	h.e.*	breathy	h.e.	breathy	h.e.
Fundamental freq. (Hz)	237	237	210	208	111	112
F1 (Hz)	475	475	630	830	610	670
F2 (Hz)	1900	1900	1270	1240	1230	1230

*h.e. = high effort

The rate at which the LF model provided glottal pulses was determined by the fundamental frequency from the original voices. The fundamental frequency was extracted using Praat phonetics software [60]. The profiles of the fundamental frequency contours were similar between the breathy and high-effort voices.

An LP order of 22 was chosen as it approximately corresponds to a typical vocal tract length [2]. A higher LP order, such as 50, with artificial excitation, yields a more natural-sounding voice. However, the additional spectral information from the higher order filter might artificially include detail about the breath quality that would otherwise remain in the LP residual. For this reason, the LP order was chosen to represent a physical vocal tract. Bandwidth expansion was carried out using the pole-scaling method [56] to reduce peakiness in the LP spectrum. The window size for the autocorrelation LP algorithm was 20 milliseconds. The LP coefficients were computed every 32 samples and linearly interpolated to reduce the influence of discontinuities between filters.

A pre-emphasis filter ($1 - 0.99z^{-1}$) was applied to the voice before it entered the LP analysis algorithm (see Figure 3.3). This pre-emphasis filter was chosen such that the LP residual approximately matches the spectral envelope of the LF model. This meant that no tilt adjustments were required when replacing the residual with the LF model.

The aspiration noise consisted of white noise with a square wave envelope that was synchronized with the pulses from the LF model. Providing noise pulses to the model helped the noise to blend into the voice more easily [48]. Three pairs of original samples resulted in three more synthesized samples. In total, six pairs of samples were evaluated. After synthesis, the samples were normalized to have the same energy level.

4.2.4 Listening Experiment

Listening experiments were carried out to evaluate the samples. A total of seven listeners participated. The perceptual criteria for this test were drawn from other studies for evaluating breathy voices [48, 29] and a prior test that we conducted [61].

The test was designed to measure relative differences between the high-effort sample and the breathy sample. This approach has been used for evaluating breathy voices [29]. The instructions were:

- Listen to the two samples and rate which one sounds more breathy.
- Listen to the two samples. Rate which voice sounds like it requires more effort to sing. Vocal effort would be associated with a tense voice rather

than a relaxed voice.

The difference in breathiness or vocal effort between the two samples was evaluated on a seven-point scale. For example, the possible ratings for breathiness ranged from much less breathy to no difference to much more breathy. Half of the rating scale can be seen on the vertical axis of Figure 4.9. Breathiness and vocal effort were evaluated in separate runs.

The listeners did not know which sample pairs were being provided or the order in which they were presented. Within each sample pair, the breathy or high-effort sample was randomly chosen to be first. This order was randomized for each run. In addition, the order of the six sample pairs was randomized for each run of the test for each listener. The evaluation process was automated for the listener and did not involve intervention on the part of the experimenter.

4.2.5 Results

The results of the subjective evaluation are displayed in Figure 4.9 and indicate the LP filters influence on the perception of breathiness and vocal effort. The ratings are presented with the high-effort voice relative to the breathy voice.

As expected, a large difference is apparent in the perceived breathiness between the original sample pairs (see Figure 4.9a). The rating of -2.3 indicates that the high-effort sample sounded less breathy than the breathy sample. When the LP filters from both of these samples were excited by the LF model, the perceived difference in breathiness was reduced to a rating of -1.1. The high-effort LP filter

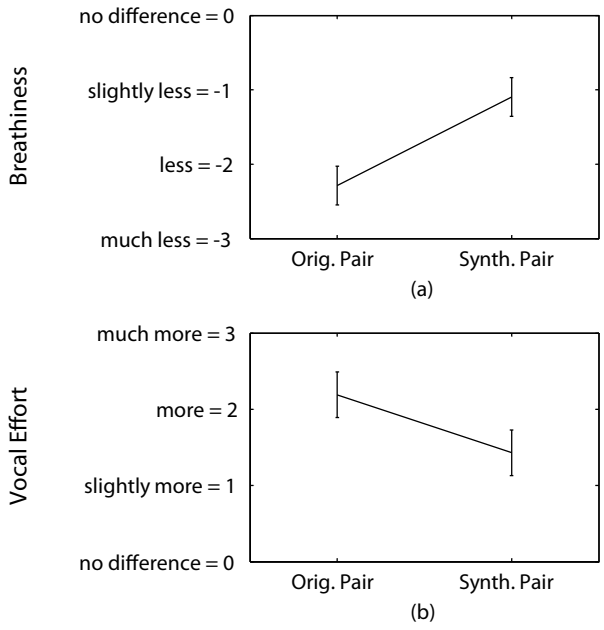


Figure 4.9: Statistical results from the artificial excitation experiment. Plot of the relative difference in (a) perceived breathiness and (b) perceived vocal effort within each sample pair. 95% confidence intervals have been plotted. "Orig. Pair" represents the rating of the original high-effort voice relative to the original breathy voice. "Synth. Pair" represents the rating of the synthesized high-effort voice relative to the synthesized breathy voice. The negative rating for breathiness indicates that the high-effort sample sounds less breathy than the corresponding breathy sample.

sounded slightly less breathy than the breathy LP filter. The 95% confidence interval indicates that the two filters did not sound the same. The LP filter from the breathy voice was clearly more breathy than the LP filter from the high-effort voice.

A large difference occurred in the perceived vocal effort between the original sample pairs (see Figure 4.9(b)). The rating of 2.2 indicates that the high-effort sample sounded like it had more effort than the breathy sample. When the LP

filters from both of these samples were excited by the LF model, the perceived difference in vocal effort was reduced to a rating of 1.4. The high-effort LP filter still sounded like it had more effort than the breathy LP filter. The 95% confidence interval indicates that the two filters did not sound the same.

A test for statistical significance was carried out. The F-test on the breathiness ratings yielded an F-value of 22.0, indicating less than 0.01% chance that the differences could occur due to noise in observations. The F-test on the vocal effort ratings resulted in an F-value of 6.8, indicating only a 1.4% chance that the differences could occur due to noise [62].

4.2.6 Discussion

The LP filters from the high-effort samples sound different than the LP filters from the breathy samples because a consistent difference occurs between their spectra. The spectra for the three pairs of filters have been plotted in Figure 4.10. The first formant for the breathy filters is at a lower frequency and is generally stronger than that for the high-effort filters. The breathy filters typically have less energy than the high-effort filters between 1000 Hz and 4500 Hz. The breathy filters accentuate the first formant while the high-effort filters emphasize higher frequencies.

The difference in spectral emphasis between high and low frequencies is likely due to physiological changes in the glottal source. Figure 4.7(b) shows how the glottal source changes between breathy and high-effort voices. This change in the spectra is captured by the LP filters, rather than the LP residual.

The LP algorithm does not take spectral changes to the glottal source into account. Whether the voice has much or a little vocal effort, whatever the shape of the glottal spectrum, the spectral envelope of the residual does not change. The spectral envelope of the LP residual is fully determined by the pre-emphasis filter, as seen in Figure 3.3c. No identification of an appropriate spectral envelope is present for the source in the standard LP algorithm.

4.2.7 Summary

This listening experiment showed that LP filters influence the perception of breathiness and vocal effort. The LP filters were estimated from pairs of voice samples where one was breathy and the other had high-effort. The pairs of LP filters were excited with the same LF excitation to ensure that the LP filters created the only differences between the samples. Listening tests were carried out to evaluate the differences between LP filters from the breathy and the high-effort voices. The data indicate that the LP filters retain some of the breathiness or vocal effort from the original voices. However, from a perceptual perspective, the formant filter should not contain information about breathiness or vocal effort. This information should be in the glottal source.

The LP formant filters capture some of the perception of breathiness and vocal effort since LP, as commonly implemented, assumes that the glottal source has a fixed spectral envelope. In contrast, the true glottal source varies according to different voice qualities, often within a single phrase. Consequently, the LP filter captures the variation in the spectral envelope. One way to vary the spectral

envelope of the residual is to implement an APLP algorithm.

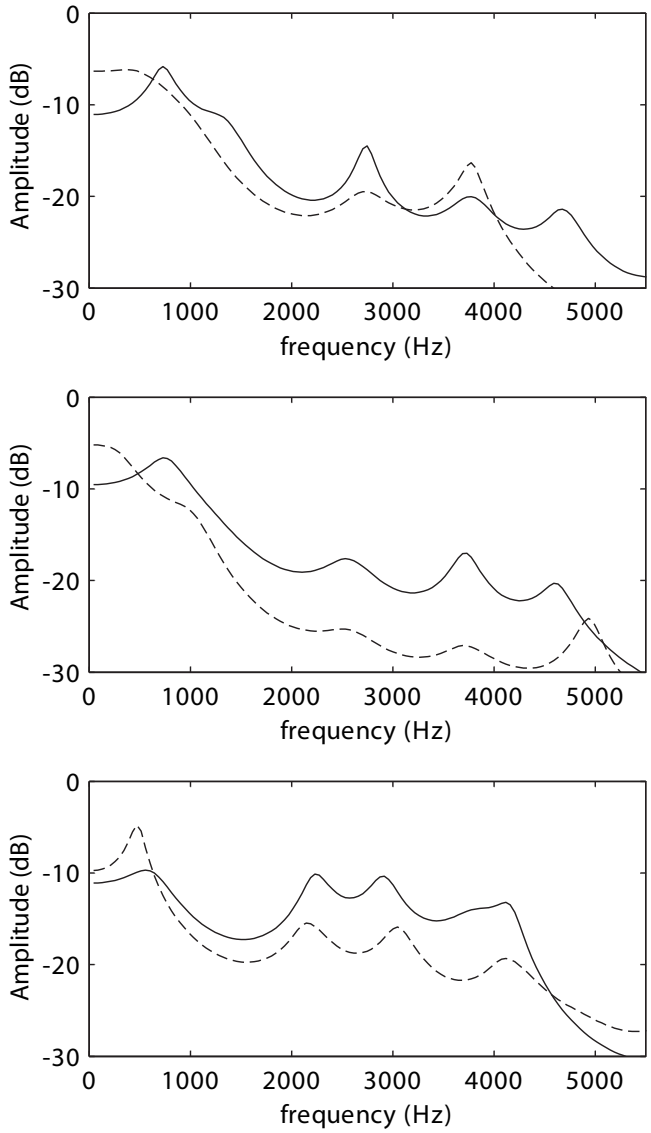


Figure 4.10: Frequency spectra from a number of LP filters for breathy voices (dashed lines) and high-effort voices (solid lines). In each plot the same voice is singing the same vowel on the same fundamental frequency. The filters estimated from the breathy voices apparently have a stronger first formant and less high-frequency content than the corresponding filters from the high-effort voices.

Chapter 5

Adaptive Pre-emphasis Linear Prediction (APLP)

Changes in vocal effort affect the spectral envelope of the voice source. This results in a voice source that varies. However, LP essentially specifies a fixed spectral envelope for the estimated source. This results in a mismatch. Because the spectral envelope of the estimated source does not vary, the variation in vocal effort is instead captured by the formant filter. Ideally, this information should be captured by the LP residual. Here, APLP is proposed as a technique to separate the influence of vocal effort from the formant filter. APLP estimates a formant filter that is more consistent across varying voice qualities.

The source-filter model of the voice separates the glottal source from the filtering influence of the vocal tract. However, when there are dramatic changes in the spectral envelope of the voice, constant pre-emphasis implementations of LP

do not appropriately separate the source and the filter. This happens because the pre-emphasis essentially specifies the spectral tilt of the estimated glottal source.

Earlier algorithms have implemented adaptive pre-emphasis for phonetic analysis [63, 64, 65]. The presentation of these algorithms did not make explicit the relationship between the pre-emphasis and the spectral envelope of the glottal source. This presentation of APLP does make that relationship more explicit. The following discussion will describe how the pre-emphasis before LP determines the spectral envelope of the glottal source. Another issue with these algorithms is that they depend upon ideally recorded voices where the glottal closure instants can be clearly determined. The algorithm to be presented here does not depend upon that requirement. In addition, there is another significant difference between how these algorithms have been implemented and the APLP algorithm presented here.

Later on in the chapter, the APLP algorithm will be applied to wider bandwidth signals for the purpose of modeling the vocal effort. This will entail a modification to the algorithm. Instead of strictly modeling the spectral envelope of the glottal source, the wide-bandwidth APLP algorithm will model the spectral emphasis of the voice signal. This spectral emphasis corresponds to the influence of vocal effort upon the glottal source and the frequency response of the vocal tract. With APLP modeling the spectral emphasis, the perception of vocal effort can be directly modified.

5.1 Influence of Pre-emphasis on the Estimated Glottal Source

The following derivation demonstrates that the spectrum of the estimated glottal source has a slope that is inverse to the pre-emphasis before LP.

The operation of the voice can be expressed as a linear model in the z-domain [2] (Figure 3.3(b)):

$$G(z)V(z)L(z) = S(z). \tag{5.1}$$

$G(z)$ represents the flow signal from the glottal source (the volume-velocity wave). $V(z)$ is a filter that represents the influence of the vocal tract. $L(z)$ represents the influence of lip radiation with a filter with z-transform $(1 - z^{-1})$ [52]. $S(z)$ is the acoustic pressure signal received at the ear.

Working with the derivative of the glottal airflow, $G'(z)$, makes it easier to see the features of the glottal pulses. Since the filter for lip radiation, $L(z)$, approximates taking the derivative [52], Equation 5.1 can be simplified:

$$G'(z)V(z) = S(z). \tag{5.2}$$

LP analyzes $S(z)P(z)$ to estimate $\hat{V}(z)$, where $P(z)$ is the chosen pre-emphasis filter:

$$S(z)P(z) = \hat{V}(z)E(z). \tag{5.3}$$

The excitation signal, $E(z)$ represents the spectrally flat excitation signal that LP

would estimate through inverse filtering. LP estimates an all-pole vocal tract filter, $\hat{V}(z)$, that approximates the spectral envelope of $S(z)P(z)$.

However, in the application of LP to voice analysis, the original signal, $S(z)$, is inverse filtered instead of $S(z)P(z)$. This filtering process can be seen in Figure 3.3c. By manipulating Equation 5.3, the result of inverse filtering can be shown to be:

$$S(z)/\hat{V}(z) = E(z)/P(z). \tag{5.4}$$

Since the excitation signal, $E(z)$, is spectrally flattened by LP, the extracted residual, $S(z)/\hat{V}(z)$, has a slope that is inverse to the slope of the pre-emphasis, $P(z)$. Note that for the excitation signal, $E(z)$, to be spectrally flat, the estimated vocal tract filter, $\hat{V}(z)$, and pre-emphasis, $P(z)$, must appropriately fit the voice signal, $S(z)$. Typically, LP is of sufficient order in estimating the formant filter, $\hat{V}(z)$, that the resulting excitation, $E(z)$, appears flat.

From Equation 5.2, we can also see how to estimate the glottal source, $\hat{G}'(z)$:

$$S(z)/\hat{V}(z) = \hat{G}'(z). \tag{5.5}$$

Combining Equation 5.4 and 5.5 shows that the estimated glottal source, $\hat{G}'(z)$, has a slope that is inverse to the pre-emphasis, $P(z)$:

$$\hat{G}'(z) = E(z)/P(z). \tag{5.6}$$

Constant pre-emphasis is commonly used in voice analysis. This enforces a

constant spectral envelope for the estimated glottal source, as seen in Equation 5.6. In contrast, the spectral envelope of the actual glottal source varies through time. This means that the estimated source does not capture variation in the spectral envelope of the actual glottal source. Instead, this variation is captured by the LP filter, $\hat{V}(z)$. In constant pre-emphasis LP, the influence of varying vocal effort is entangled in the estimated vocal tract filter, $\hat{V}(z)$, making it difficult to separate out and control the influence of effort.

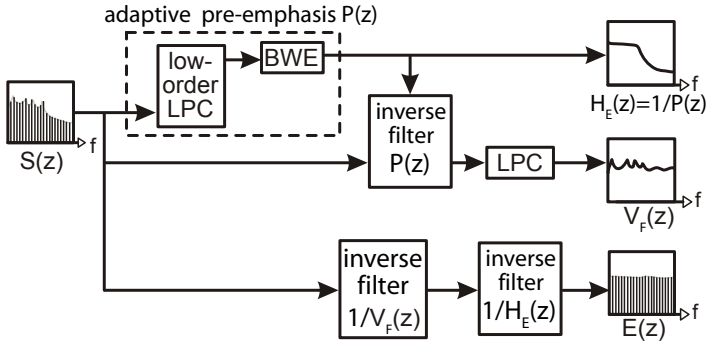


Figure 5.1: Adaptive pre-emphasis linear prediction for voice analysis. (BWE refers to bandwidth expansion.)

5.1.1 APLP analysis

The algorithm for APLP analysis is presented in Figure 5.1. In APLP, the pre-emphasis filter, $P(z)$, is estimated using low-order LP. This enables the pre-emphasis filter to track variations in the spectral envelope of the voice signal. The voice signal, $S(z)$, is inverse filtered with the pre-emphasis, $P(z)$, to spectrally flatten the signal before the second stage of LP. This second stage of LP captures the formant information in an estimated filter, $\hat{V}(z)$, using a higher or-

Table 5.1: Spectral slopes that result from constant and adaptive pre-emphasis in a linear model of voice production

$S(z)$ $dB/oct.$	$P(z)$ $dB/oct.$	$\hat{V}(z)$ $dB/oct.$	$\hat{G}(z)$ $dB/oct.$
-9 to -3	constant: 6	-3 to 3	-12
-9 to -3	adaptive: 3 to 9	0	-15 to -9

der for LP. Because the pre-emphasis flattens the signal before the second stage of LP, the estimated formant filter, $\hat{V}(z)$, becomes more consistent than the corresponding formant filter from constant pre-emphasis LP. In contrast, constant pre-emphasis LP causes variation in the spectral emphasis to be included in the estimated vocal tract filter, $\hat{V}(z)$. The differences in estimated slopes between constant pre-emphasis LP and APLP can be seen Figure 5.2 and Table 5.1.

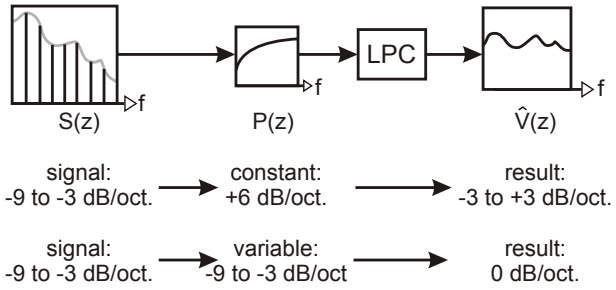


Figure 5.2: Spectral slopes from constant pre-emphasis LP and APLP. The estimated vocal tract filter, $\hat{V}(z)$, either has a constant or a varying tilt depending upon whether the pre-emphasis is constant or adaptive.

Constant pre-emphasis LP was compared to APLP by analyzing voices with contrasting levels of vocal effort. The same singer sang the same vowel at the same fundamental frequency but with different voice qualities. One of the voice qualities

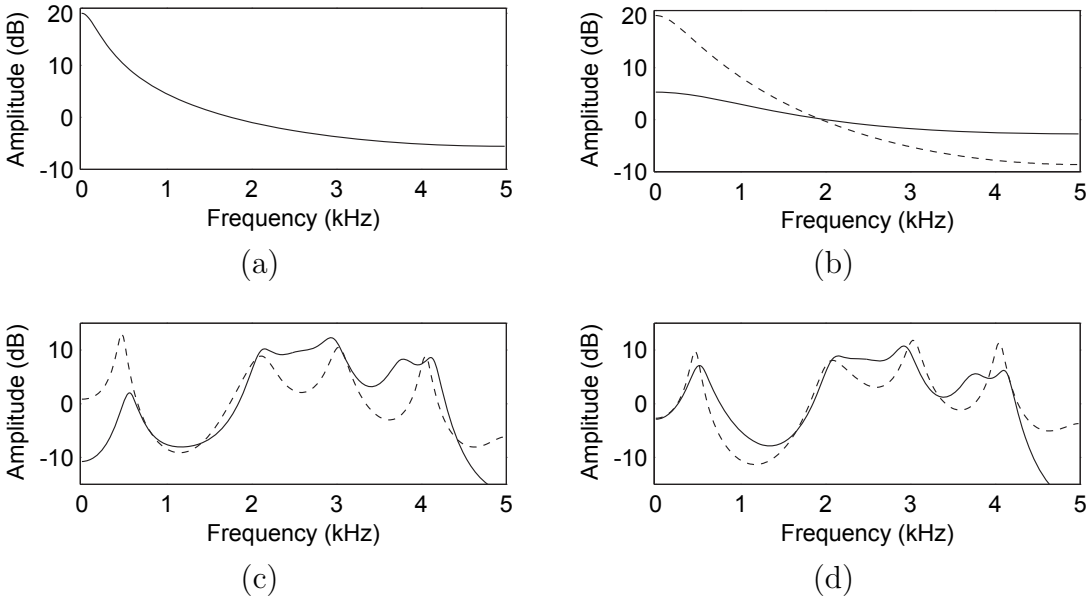


Figure 5.3: Pre-emphasis and vocal tract filters estimated using constant pre-emphasis LP (left) and adaptive pre-emphasis LP (right) for breathy (dashed lines) and high effort voices (solid lines). Inverse of the pre-emphasis filters for (a) constant pre-emphasis LP and (b) APLP. Estimated formant filters for (c) constant pre-emphasis LP and (d) APLP.

was breathy and the other exhibited high effort. There was a large difference in spectral tilt between the samples. Constant pre-emphasis LP and APLP were carried out on the samples to observe the resulting separation between source and filter.

The voice samples had a sampling rate of 10 kHz. The order of the LP formant filter was 16. The constant pre-emphasis filter was $(1 - 0.9z^{-1})$. The adaptive pre-emphasis filter was determined as the convolution of a constant filter $(1 - 0.5z^{-1})$ and a single-pole adaptive pre-emphasis filter. Using two filters in this way ensured that the adaptive pre-emphasis filter could adjust within an appropriate range.

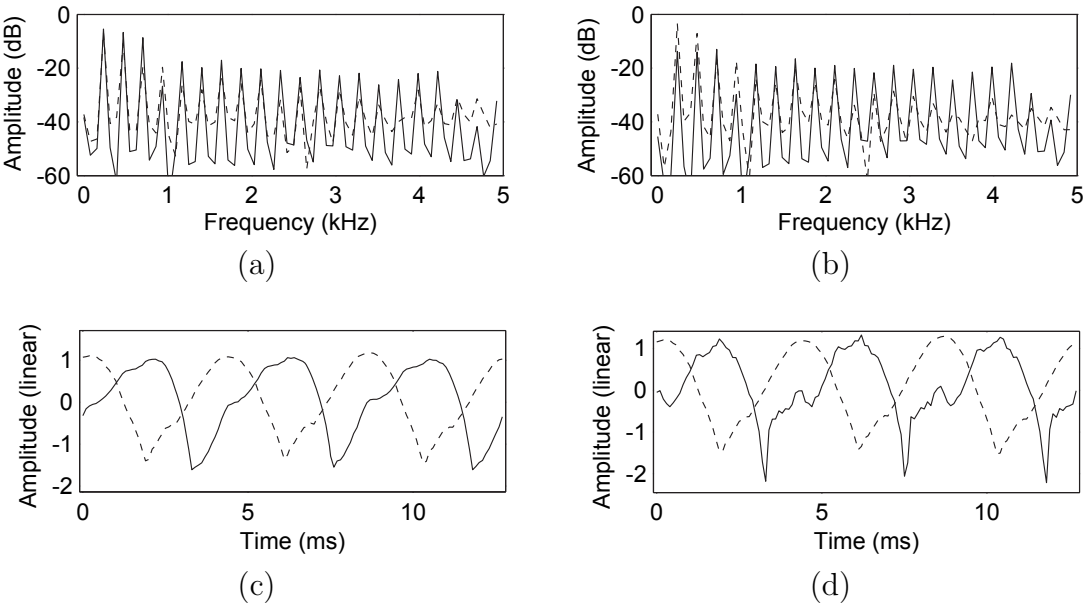


Figure 5.4: Voice source estimated using constant pre-emphasis LP (left) and APLP (right) for breathy (dashed lines) and high effort voices (solid lines). Spectra of the residual of (a) constant pre-emphasis LP and (b) APLP. If estimated correctly, this corresponds to the spectrum of the derivative volume-velocity wave, $\hat{G}'(z)$. The glottal volume-velocity wave has been estimated using (c) constant pre-emphasis LP and (d) APLP. Note that APLP extracts pulses with a much sharper instant of glottal closure. (The vertical scale has been scaled for visual comparison.)

The results of source-filter separation have been plotted in Figure 5.3 and 5.4. Each of the figures is the result of doing either constant pre-emphasis LP or APLP on a breathy and a high-effort voice.

The inverse of the pre-emphasis filters have been plotted in Figure 5.3(a,b). As mentioned above, the inverse of the pre-emphasis filters determine the tilt of the estimated voice source. It is clear from these plots that APLP fits a significant range of slopes for the estimated voice source. In contrast, constant pre-emphasis LP works only for voices that match the prior chosen value for the pre-emphasis.

The formant filters resulting from constant pre-emphasis LP and APLP have been plotted in Figure 5.3(c,d). Constant pre-emphasis LP does not allow the slope of the estimated voice source to adapt to the slope of the voice source. Instead, the formant filter adapts to this change in slope. Figure 5.3(c) shows how the formant filters from constant pre-emphasis LP vary depending upon whether the analyzed voice exhibits breathiness or effort. The variation appears in the contrast between the energy levels of the first formant and the upper formants. APLP estimates formant filters that are more consistent across varying voice qualities as seen in Figure 5.3(d).

Inverse filtering the original signal with the formant filter provides an estimate of the derivative volume-velocity wave at the glottis. The spectra of the residual have been plotted in Figure 5.4(a,b). If estimated correctly, this corresponds to $G'(z)$. Note that the breathy and high-effort residual from constant pre-emphasis LP look more similar to one another than the corresponding spectra from APLP. This is because LP with constant pre-emphasis results in a spectrally constant

residual.

To get an estimate of the volume-velocity wave, all that is required is an additional integration as seen in Figure 3.3(g,h). If we compare the volume-velocity pulses extracted using constant pre-emphasis LP to the pulses extracted using APLP, we can see a clear difference between the algorithms. In Figure 5.4(c), we can see that constant pre-emphasis LP estimates glottal pulses that look similar to one another regardless of whether the voice exhibits breathiness or vocal effort. Although there is a ripple in the pulses for the high-effort voice, the overall shape of the pulses are similar to the pulses from the breathy voice. In contrast, the pulses extracted using APLP look different depending upon whether the voice exhibits breathiness or high effort (Figure 5.4(d)). This is the type of difference in glottal pulses that we would expect to see between two voices that exhibit different voice qualities.

The estimation of the glottal pulses do not exactly match the shape that we would expect to see. This is because small phase errors or slightly different locations for the peaks of the glottal pulses can result in large changes in the shapes of the pulses. In this particular data set with this technique, we would not expect the results to be better. The samples at hand were not recorded under perfect recording conditions and we are using correlation LP rather than closed-phase covariance analysis. In addition, to achieve desirable results, careful manual control of the algorithm is often required, whereas in this situation, the algorithm was automated. The purpose of the algorithm to be developed is to transform the perceived vocal effort and breathiness in the voice for musical applications. Therefore, it is not

critical that the pulses be extracted perfectly. Musically recorded signals may not retain sufficient phase information to extract accurate glottal pulses. Therefore, time was not spent ensuring that the extraction of glottal pulses was perfect.

The key observation to take away in observing the differences between the pulse shapes in Figure 3.3(g,h) is that constant pre-emphasis LP estimates pulse shapes that appear very similar whereas APLP estimates pulse shapes that look very different from one another.

Carrying out constant pre-emphasis LP and APLP on high-effort and breathy voices revealed that the two algorithms operate differently. In constant pre-emphasis LP, variations in the spectral slope of the voice source show up in the estimated formant filter. In APLP, variations in the spectral slope of the voice source show up in the estimated source. This difference in spectral modeling also appears in the resulting estimates of the glottal volume-velocity wave. Constant pre-emphasis LP estimates volume-velocity waves that look similar to one another regardless of the nature of the original voice while APLP estimates volume-velocity waves that change along with changes to the voice source.

This comparison demonstrates how many typical implementations of LP using constant pre-emphasis are not able to follow variations in the slope of the glottal source while APLP is able to follow these variations.

5.1.2 Fixed-rate Versus Closed-phase Analysis

In fixed-rate LP analysis, the pre-emphasis determines the spectral envelope of the estimated glottal source. However, in closed-phase analysis, some of the variation

in the spectral envelope affects the estimated glottal source. This is because closed-phase analysis excludes large spikes in the time domain that occur at the instances of glottal closure. The amplitude of these spikes vary with the rate of glottal closure and this is closely related to the perception of vocal effort. With higher levels of vocal effort, the spikes are larger and this tends to flatten the voice spectrum. Since these spikes are excluded from analysis, their influence is passed on to the estimated glottal source, $\hat{G(z)}$, during inverse filtering. This results in an estimated glottal source that somewhat follows variation in the spectral slope of the voice signal. With careful preparation, closed-phase analysis can be used to make good estimates of the spectral slope of the glottal source [48]. That said, in at least one case, adaptive pre-emphasis has been used to improve closed-phase analysis [65].

We are using fixed-rate analysis because we cannot depend upon the source data being accurate enough to estimate reasonable glottal pulses. In the present application, it is safe to say that the pre-emphasis fully determines the slope of the estimated glottal source.

5.1.3 Wider Bandwidth Speech Signals

When working with voices sampled at or below 10 kHz, the adaptive pre-emphasis filter has a slope that varies with the level of vocal effort. However, voice signals sampled at higher frequencies often exhibit a trend that does not look like a simple slope. Instead, for high effort voices, the spectrum can look relatively flat up to 4.5 kHz and then drop off suddenly beyond that range. If we estimate a pre-emphasis filter using LP with an order above one, there is no longer a guarantee

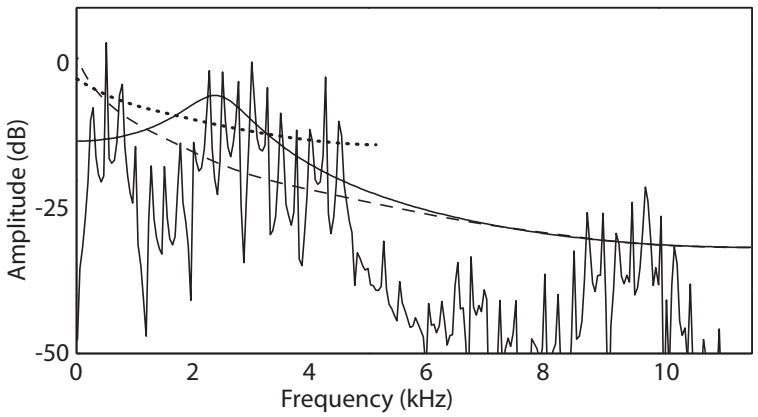


Figure 5.5: APLP fits the emphasis filter, $H_E(z)$, differently depending on the bandwidth of the signal and the order of the pre-emphasis. The dotted line represents the emphasis filter for a low bandwidth signal. The solid line represents the emphasis filter estimated using 3rd order LP. The dashed line represents how the emphasis filter might fit the signal if it were constrained to be a slope.

that the pole of the filter will be at 0 Hz. For breathy voices, the regime of the voice appears relatively consistent throughout the frequency range and the spectral emphasis filter, $H_E(z)$, still looks like a tilt. However, for high-effort voices, a pole of the filter ends up being at an intermediate frequency as seen in Figure 5.5. In this situation the pre-emphasis no longer represents the envelope of the glottal source.

Figure 5.5 illustrates how different pre-emphasis filters can result depending upon how the pre-emphasis algorithm is configured. If the signal goes up to 5 kHz, the pre-emphasis appears like the dotted line. However, for higher bandwidth signals, there can be different results. The solid line with a resonance between $2-3$ kHz was estimated using a pre-emphasis filter with an order of three. However, by constraining the estimated pre-emphasis to be a slope, the curve with the dashed

line could result. However, this curve does not correspond to the results achieved when the signal only goes up to 5 kHz. For wider bandwidth signals, the pre-emphasis is significantly influenced by the frequency content above 5 kHz.

This drop-off in the frequency content of the signal that happens above 5 kHz is likely due to a drop-off in the frequency response of the vocal tract filter as describe in Chapter 3. With appropriate adjustment, the APLP algorithm could be tuned to appropriately separate the slope of the glottal source from the drop-off in vocal tract frequency response. This could be accomplished by assuming that the vocal tract frequency response is flat up until 4 kHz. The pre-emphasis filter, $1/P(z)$, would be fit to this frequency range with one to three pole(s) at zero hertz. Frequencies above 4 kHz would be ignored during the fitting stage but would nonetheless be influenced by the filter. This would result in an estimated glottal source, $\hat{G}'(z)$, similar to traditional phonetic techniques while estimating a formant filter, $\hat{V}(z)$, that includes the drop-off in the spectrum between $4-5$ kHz.

If the drop-off between $4-5$ kHz is due to constriction in the lower vocal tract, then the pre-emphasis filter should be estimated using a bandlimited signal. To create a physiologically realistic model of the glottal excitation, it's necessary to estimate a pre-emphasis that is a slope rather than containing a resonance. In addition, this slope needs to be calculated based on the signal frequency content up to $4-5$ kHz. In developing this kind of algorithm, one could continue to draw on the wealth of knowledge built up for the phonetic analysis of the voice, continuing to focus on the spectral region below $4-5$ kHz. Two possible techniques for analyzing only the lower frequencies in wide-band signals are are split-band

analysis and warped LP [66].

Some techniques have been described for calculating a physiologically realistic glottal signal. However, in the present application, our goal is to manipulate the perceived vocal effort. Working towards this goal, a new technique has been developed.

5.2 APLP For Estimating Spectral Emphasis

The previous section described how APLP could be used to extract glottal pulses. This section describes how analyzing wider bandwidth speech signals results in more complications to the algorithm. For signals sampled above 10 kHz, the perception of vocal effort is caused by both the glottal source and the vocal tract filter. In the APLP algorithm presented here, we lump both of these influences into one spectral emphasis filter. While it is possible to use adaptive pre-emphasis LP to extract glottal pulses, our intent is to separate the spectral emphasis from the formant information and to manipulate the perceived type of vocal phonation.

When voices exhibit vocal effort, both the vocal tract and the glottal source change in a way that affects the spectral envelope of the voice. The slope of the glottal source steepens in a negative direction and the vocal tract exhibits a drop-off in the spectrum above $4-5$ kHz. The spectral emphasis filter, $H_E(z)$, with a resonance (Figure 5.6(b)) models this drop-off in the spectrum. This means that the spectral emphasis filter is now modeling more than just the slope of the glottal source. For this reason, a new model of APLP is introduced as seen in Figure 5.7.

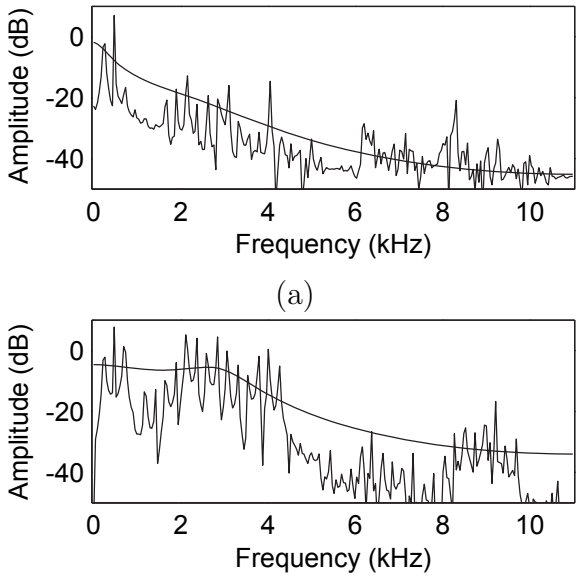


Figure 5.6: Resonance in spectral emphasis filter estimated by APLP. Voice spectra from (a) a breathy voice and (b) a high-effort voice. In each plot the same voice is singing the same vowel on the same fundamental frequency. The spectral emphasis filter ($H_E(z)$) estimated by low-order LP has also been plotted.

The signal entering the formant filter is no longer referred to as the glottal source but will now be described as a flat excitation, $E(z)$, shaped by a spectral emphasis filter, $H_E(z)$. In addition, $\hat{V}_F(z)$, has been given the subscript 'F' to indicate that this filter models the perceptual influence of the formants and not all of the frequency response of the vocal tract.

The algorithm for APLP analysis, presented in Figure 5.1, produces the following model of the voice:

$$E(z)H_E(z)\hat{V}_F(z) = S(z), \tag{5.7}$$

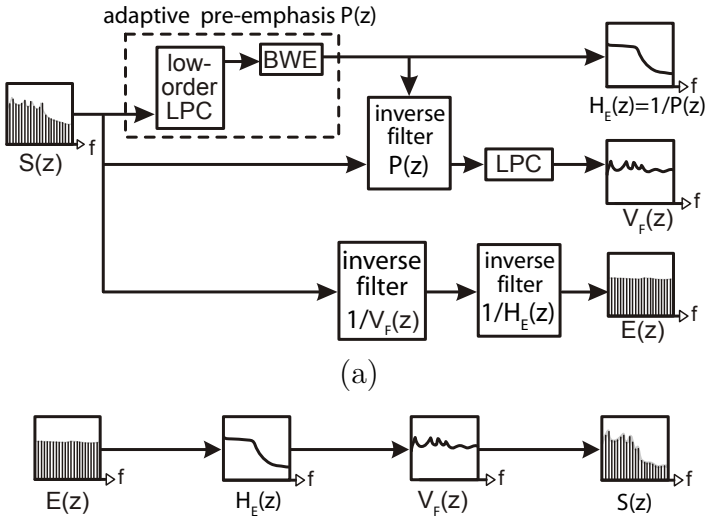
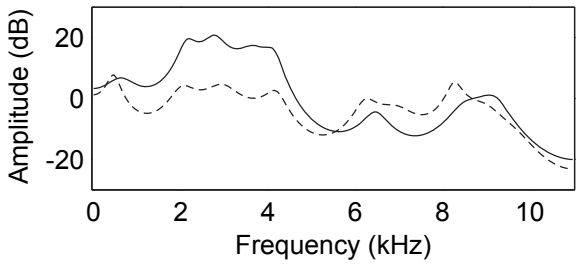
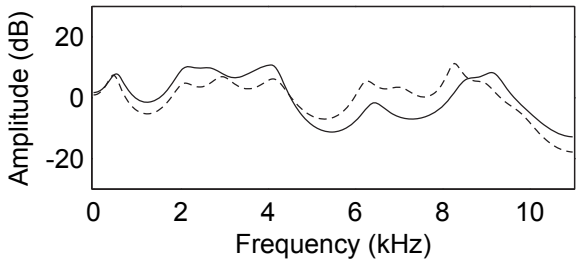


Figure 5.7: (a) APLP for estimating spectral emphasis, $H_E(z)$, and the formant filter, $V_F(z)$. The resulting excitation, $E(z)$, has a flat spectrum. (b) Linear model of the voice for controlling spectral emphasis. (BWE refers to bandwidth expansion.)

which appears quite different from the typical linear model of the voice as presented in Equation 5.1 and 5.2. A flat excitation, $E(z)$, is shaped by the spectral emphasis filter, $H_E(z)$, which is further shaped by the formant filter, $\hat{V}_F(z)$, to produce the sound received by the ear, $S(z)$. Since the spectral emphasis filter, $H_E(z)$, models the overall spectral emphasis, the excitation, $E(z)$, is spectrally flat and the formant filter, $\hat{V}_F(z)$, are more consistent than they would be without the spectral emphasis filter. The formant filter, $\hat{V}_F(z)$, captures the narrow peaks in the spectrum while the spectral emphasis filter, $H_E(z)$, captures the overall shape of the spectrum. The perception of high vocal effort is caused by changes to both the glottal source and the vocal tract filter. In the APLP algorithm presented



(a)



(b)

Figure 5.8: Formant filters estimated using (a) constant pre-emphasis LP and (b) APLP. The same voice sang the same vowel at the same fundamental frequency while varying the quantity of vocal effort. The dashed lines are estimated formant filters from the breathy voice while the solid lines are estimated formant filters from the high-effort voice. Note that the formant filters for breathy and high-effort voices are more similar for APLP than for constant pre-emphasis LP.

here, we lump both of these changes into one spectral emphasis filter, $H_E(z)$.

For signals sampled at 10 kHz and below, the spectral emphasis filter, $H_E(z)$, is a tilt that approximates the spectral envelope of the glottal source. In this situation, $E(z)H_E(z)$ corresponds to $\hat{G}'(z)$, and $\hat{V}_F(z)$ corresponds to $\hat{V}(z)$.

In order to manipulate vocal effort in singing voices, we want to separate the linguistic information (the formant filter) from the information about the perceived breathiness and vocal effort. APLP is useful because it is able to capture

the spectral emphasis of the voice, $H_E(z)$, independent of the formant filter, $\hat{V}_F$. As a result, APLP provides a more consistent estimate of the formant filter than constant pre-emphasis LP. This can be observed in Figure 5.8, which compares the formant filters from constant pre-emphasis LP and APLP. Because APLP better separates the linguistic information from the spectral emphasis, it is easier to manipulate the spectral emphasis independently of the formant filter. In contrast, constant pre-emphasis LP entangles the spectral emphasis information in the formant filter.

Practically speaking, when analyzing a particular voice signal, it is usually impossible to determine how much of the change in the spectral emphasis is due to a change in the glottal source versus the vocal tract. What is known is that, with increases in vocal effort, both the glottal source [22] and the vocal tract filter [37] emphasize higher frequencies. For this reason, we created the spectral emphasis filter to capture the combined influence of both the glottal source and the vocal tract filter on the overall spectral envelope of the voice. This deviates from the typical goal of separating the glottal source and the vocal tract. However, the spectral emphasis filter provides an easier way to manipulate the perceived vocal effort independently of the formant filter.

There are two key differences between the APLP algorithm presented here and other implementations of adaptive pre-emphasis for voice source analysis [63, 64, 65]. Firstly, voice source analysis extracts estimates of the glottal pulses whereas APLP extracts a spectrally flat excitation and a spectral emphasis filter. This spectral emphasis filter is distinct from the narrow spectral peaks associated with

the perception of formants. Secondly, current methods of voice source analysis operate on frequencies no higher than 5 kHz while APLP works with broader bandwidth speech signals.

The APLP algorithm requires a more complex model of the spectral emphasis. In standard methods of voice source analysis, the spectrum of the pre-emphasis has a simple slope but in APLP the pre-emphasis can have a more complex shape. This is because APLP aims to analyze musical voice signals and manipulate them in a way that is musically relevant. In doing this, frequencies above 5 kHz are important because they affect the aesthetics of the voice signal. However, when wider bandwidth signals are involved, the spectral emphasis no longer looks like a simple slope. This is likely because the frequency response of the vocal tract often drops off sharply beyond $4 - 5$ kHz [37].

In addition, APLP does not require the ideal retention of phase information. Voice source analysis aims to extract realistic estimates of the glottal pulses [45]. This means that the original signal must retain phase information to prevent the extracted glottal pulses from looking distorted in the time domain. In contrast, the APLP algorithm presented here is not intended to extract the shapes of the glottal pulses. APLP presented here focuses on musical signals where the absolute phase information is less important.

5.2.1 Bandwidth Expansion

LP typically results in formant filters have more pronounced peaks than a typical vocal tract. As such, it is useful to do bandwidth expansion to broaden the formant

resonances in the estimated vocal tract filter [56, 67, 68]. Perceptually, this results in a slight improvement in the sound of the voice model. When estimating the spectral emphasis filter, bandwidth expansion becomes even more critical.

APLP uses two stages of LP to estimate a spectral emphasis filter and a formant filter. The intent of this technique is that the formant filter captures the phonetic perception of vowels and the spectral emphasis filter captures the overall distribution of frequencies in the voice. If the spectral emphasis filter, $H_E(z)$, contains a sharply peaked resonance, then the emphasis filter could be perceived as a formant. Afterwards, if the spectral emphasis filter is manipulated with the intent to change the perceived vocal effort, then strange artifacts could be created due to a formant being inadvertently manipulated. In addition, accidentally modeling a formant in the spectral emphasis filter, $H_E(z)$, means that the formant filter, $\hat{V}_F(z)$, is not properly modeling that formant.

Bandwidth expansion widens the resonances in an all-pole filter, making the peaks more wide and less peaky. There are two key techniques that can be used for bandwidth expansion for LP [56]. One of these techniques modifies the autocorrelation coefficients before they enter LP. The other technique modifies the filter using pole scaling after LP has already estimated the coefficients. In this implementation, pole scaling was chosen because the other technique can modify the frequency of the resonance with the use of larger scaling values.

With radial pole scaling, the direct form filter is modified by replacing z with z/α. In the all-pole filter,

$$H_E(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{k=1}^{N} a_k z^{-k}} \tag{5.8}$$

This corresponds to substituting

$$a'_k = \alpha^k a_k \tag{5.9}$$

When $\alpha < 1$ this corresponds to moving the poles of the all-pole filter towards the origin. If there are any sharp peaks in the filter, these peaks are broadened. The value chosen for α for the spectral emphasis filter was 0.9 while the value for the formant filter was 0.975.

5.2.2 Chapter Summary

APLP can be configured to model the spectral emphasis of the voice. Perceptually, this spectral emphasis has a significant influence on the perceived vocal effort. For this reason, APLP can be configured to transform the spectral emphasis of the voice. By converting the spectral emphasis of a high-effort voice to the spectral emphasis of a breathy voice, it is thought that the perceived effort in the voice will be significantly reduced, thereby improving the blending of aspiration. The following chapter describes how APLP can be used to transform high effort voices into breathy voices.

Chapter 6

APLP for Voice Transformation

The first half of this chapter describes how the spectral emphasis filter estimated by APLP can be manipulated to reduce the high frequency content associated with vocal effort, thereby improving the blending of aspiration noise. The second half of the chapter contains a listening experiment demonstrating that APLP transforms high-effort voices into breathy voices more effectively than constant pre-emphasis linear prediction.

6.1 Voice Transformation Algorithm

APLP can be used to estimate a spectral emphasis filter. This spectral emphasis is a significant cue to the perception of vocal effort. To transform high-effort singing voices into breathy voices, it is necessary to reduce the spectral emphasis of the upper frequencies; this can be achieved with the estimated spectral emphasis filter.

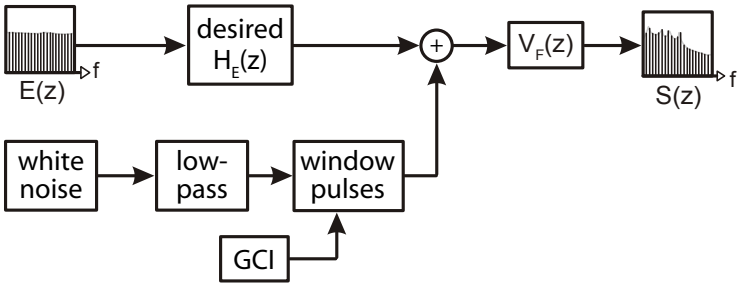


Figure 6.1: APLP synthesis configured to modify the perception of vocal effort. The flat excitation, $E(z)$, is shaped by the desired spectral emphasis filter, $H_E(z)$. White noise is low-passed with a first-order, all-zero filter to remove the uppermost frequencies. The coloured noise is then pulsed according to the glottal closure instants (GCI). The noise is added to the spectrally shaped excitation. Then the formant filter, $V_F(z)$, is applied to generate the newly synthesized voice signal, $S(z)$.

First, APLP analysis removes the formants by inverse filtering with the estimated formant filter, $V_F(z)$. Then, APLP analysis removes the the spectral emphasis by inverse filtering with the spectral emphasis filter, $H_E(z)$. This process can be observed in Figure 5.7. During synthesis, the excitation, $E(z)$, is filtered by the desired spectral emphasis, $H_{Enew}(z)$; artificial noise is added to simulate aspiration noise; and the formant filter, $V_F(z)$, reapplies the formants. This resynthesis process can be observed in Figure 6.1. If the new spectral emphasis filter and the artificial noise appropriately match the target breathy voice, the high effort voice, in theory, should be transformed into a breathy voice.

At first glance, it may appear that this approach would not be any better than filtering the signal with a constant filter to reduce the high-frequency content in the signal. However, simple filtering is helpful only in situations where the

characteristics of the original signal are already known. For example, filtering should only occur when the the original signal exhibits high effort. Otherwise, filtering would result in a voice with too little high frequency content. This type of voice would sound unnatural.

The advantage of APLP is that it computes a running estimate of the spectral emphasis of the signal. As a result, the quantity of spectral emphasis to remove is known. If a high effort voice and a breathy voice enter the APLP algorithm, the voices exiting the algorithm will have a similar spectral emphasis.

While developing the algorithm, a sound file was used where the same person sequentially sang two voice samples on the same vowel and at the same fundamental frequency but with two different voice qualities: high-effort and breathy. The APLP algorithm was designed to match the spectral emphasis of the breathy voice. As a result, the spectral emphasis of the breathy voice was not modified. The only change made to the the breathy voice was the addition of noise, which lead to a slight increase in the perception of breathiness. In contrast, the spectral emphasis of the high-effort voice was significantly changed. The added noise was also more obvious due to the lack of aspiration noise in the original high-effort voice. The input to the APLP algorithm was a sound file with highly contrasting voice qualities. The output from the algorithm was a sound file with two voices that sounded more similar.

Controlling the Emphasis Filter

As an alternative to using the spectral emphasis filter from the breathy voice, a parametric, third-order filter can be used to control the emphasis. This filter has three controls: tilt (t), centre frequency (θ) and bandwidth (r) as described in the equation below.

$$H_{Enew} = \frac{1}{(1 - tz^{-1})(1 - 2r\cos\theta z^{-1} + r^2 z^{-2})} \tag{6.1}$$

The tilt parameter, t, controls a first-order tilt filter. The other parameters, θ and r control a second-order resonance filter.

To use the filter it was necessary to find appropriate starting values. These values can be derived from the spectral emphasis filter estimated from the target breathy voice. To that end, various voices were analyzed in order to extract reasonable starting values for the spectral emphasis filter, $H_{Enew}(z)$. The values for this filter are summarized in Table 6.1 and the spectra of these filters have been plotted in Figure 6.2.

Interpolated LP coefficients

The LP coefficients were computed every thirty-two samples. This resulted in some artifacts as the filter coefficients instantaneously changed from one value to the next. For this reason, the reflection coefficients from LP were interpolated to provide a smoother change from filter to filter. When interpolating the coefficients, it is important to use reflection coefficients rather than direct form coefficients.

Table 6.1: Filter values for spectral emphasis filter, H_{Enew} (Equation 6.1), estimated from breathy and high-effort voices using APLP.

Voice	Tilt (t)	Centre Frequency (θ)	Bandwidth (r)
Popeil high effort	0.7193	0.2774π (3060 Hz)	0.8027
Popeil breathy	-0.1735	0.0306π (335 Hz)	0.8217
male high effort	-0.0974	0.0896π (990 Hz)	0.8748
male breathy	-0.1839	0.0213π (235Hz)	0.8007
ab high effort	-0.2101	0.1174π (1295Hz)	0.8795
ab breathy	-0.2528	0.0843π (930Hz)	0.8642

Reflection coefficients are guaranteed to be stable as long as the two filters being interpolated are also stable.

Getting the Noise to Blend

The greatest challenge in synthesizing breathy voices is that the noise does not blend easily into the voice. Instead, listeners typically perceive the noise to be a separate sound distinct from the voice signal. In a breathy voice, the voice source, $E(z)H_E(z)$, has a noise floor that is approximately flat. This noise floor is caused due to the aspiration noise in the voice. When transforming high-effort voices, it is easiest to add aspiration noise in after the excitation has already been shaped by the spectral emphasis filter. By inserting noise at this stage, coloured noise can be used because it approximately matches the desired spectral shape for the aspiration noise. A first-order, low-pass, all-zero filter ($1 - 0.98z^{-1}$) was applied to shape the white noise. This reduces some of the uppermost noise frequencies that do not blend easily.

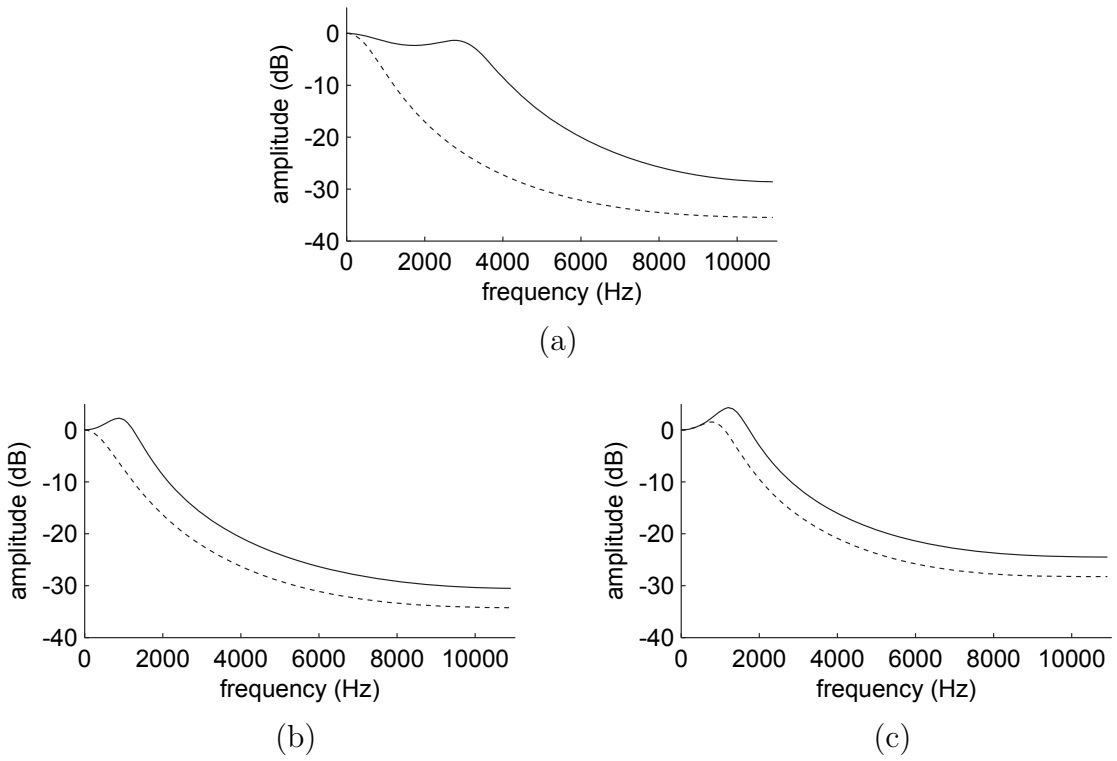


Figure 6.2: Spectral emphasis filters for (a) Popeil, (b) male and (c) ab voice samples (also described in Table 6.1). Note the contrast between emphasis filters from breathy voices (dashed lines) and high-effort voices (solid lines).

Another factor that improves breath blending is to use pulsed noise. During the oscillation of the vocal folds, the quantity of aspiration noise fluctuates with the opening and closing of the glottis. The maximum air turbulence contributing to the perception of aspiration noise as been calculated to occur just after the vocal folds open and just before they close [69]. This is because the likelihood of turbulence is proportional to the air flow and inversely proportional to the size of the aperture. At these instants, the size of the aperture is smallest.

Hermes carried out a psychoacoustic experiment to determine the influence of phase when inserting noise bursts into the glottal excitation [70]. He inserted the noise bursts various degrees out of phase with the glottal excitation instant. What he found was that the noise sounds most closely integrated with the glottal excitation when the noise is close to the glottal closure instant.

Glottal closure instant detection was carried out by threshold detection on the LP residual. When this technique was not effective, the Praat open-source software was used to pre-process the sound file and detect the glottal closure instants [60].

Unlike the artificial excitation experiment in Section 4.2, a von Hann window was used to shape the glottal pulses in time. It was found that this shaped window slightly improved the blending of artificial aspiration noise into the voice when compared to the previously used square window.

In Hermes experiments [70], he used short pulses relative to the length of time involved in a full glottal cycle. This leads to the timing of the breath pulses being more critical. The noise pulses can blend in and "hide" in the glottal pulses most easily when the two are synchronized. The choice of a wide or narrow glottal pulse affects how easily the noise will blend into the voice and be perceived as breathiness. Hermes used an artificial source for the experiment. As such, he had good control over the glottal pulses themselves. In the limiting case of a breath pulse so narrow that it appears to be an impulse, it's obvious that it would blend in and "hide" most easily when it is synchronized with the most impulsive part of the glottal signal. In this limiting case, the noise pulse is not really acting like a noise pulse any longer. It's behaving like an impulse.

When working with signals generated by natural voice, it is easy to make errors in the estimation of glottal closure instants. Using narrow breath pulses makes any errors become more evident. When tuning the breath algorithm, I found it easiest to maintain most of the benefits of breath pulses while using wider pulses to reduce the evidence of any glottal closure errors. By increasing the width of the breath pulses to about half of the glottal cycle, this timing becomes less critical while still maintaining the advantage of pulsing the noise rather than continuous noise. A 50% duty cycle has been used in other voice synthesis experiments [48].

In short, to simulate breath aspiration, pulsed white noise is added to the excitation, $E(z)$, that has been shaped by the spectral emphasis filter, $H_E(z)$. The resulting signal is a voice source with the desired spectral emphasis plus simulated noise. This voice source is then fed through the formant filter, $\hat{V}_F(z)$, from the high-effort voice to synthesize the modified voice. The spectral emphasis from the high-effort voice has been removed and replaced with the spectral emphasis from a breathy voice. Simulated aspiration noise has been added to simulate breathiness. The result is a high-effort voice transformed to be more similar to a breathy voice.

Glottal Closure Instant Detection

Synthesized noise blends more easily when it is pulsed in sync with the glottal closure instants [70]. Getting the phase correct is critical in getting the noise to blend in appropriately with the original voice signal. If the phase is not correct, this leads to an audible degradation of blending. A previous study shows that this takes place and this is confirmed in the present work. Getting the phase correct

makes a difference.

In the first attempt to create glottal closure instants, we used pitch detection to determine the fundamental frequency and then converted the fundamental frequency to glottal closure instants. However, this resulted in noise that did not blend much better than unpulsed noise. The noise pulses and the glottal closure instants were not synchronized and this resulted in two signals that did not easily blend into one signal.

These glottal closure instants were computed by spectrally flattening the signal using the spectral emphasis filter and the formant filter. This results in a signal that looks similar to an impulse train. To compute the glottal closure instants (GCI), a simple threshold was used on this signal. This method was quite crude but it worked effectively for the isolated vowel samples that we had recorded.

When the voices involved phrases or when the voices were recorded in less ideal conditions, the above technique did not work nearly as well. For these samples, we used Praat software [60] to identify the GCI ahead of time. Praat uses an accurate autocorrelation-based pitch detector to estimate GCI candidates [71]. This meant that the signal had to be pre-processed. During voice processing, the sound file and the GCI from Praat were simultaneously fed into the APLP algorithm.

Bursts of noise were added, synchronized with the GCI. Von Hann windows provided the shapes of the pulses and a duty cycle of approximately 50%.

Getting the Delays Correct

Ideally, using an LP algorithm on the voice will result in a formant filter and an excitation that are independent of one another. However, in practice, the estimated formant filter and the resulting residual are closely linked. It is not difficult to achieve perfect reconstruction by recombining the two components. However, when there are changes to one or the other, artifacts quickly start appearing in the re-synthesized signal. If these changes are small, these artifacts will be insignificant. The larger the changes, the more significant the artifacts become. For this reason, it is critical to maintain synchronization between the estimated formant filter and the extracted residual.

One of the things that required the most work in getting the algorithm to work was in getting the delays correct in the processing of the samples. There are a number of parts of the process that involve delays:

- windowing
- interpolation
- filtering
- GCI detection

For the algorithm to work correctly, it is necessary to get all of the various parts of the algorithm to be delayed appropriately. The residual has to line up with the filter from which it was analyzed. In addition, the same filter should be used to decompose and resynthesize the signal. In an ideal world, this separation

between source and filter would result in two independent signals. However, in reality, these two components are inextricably linked. If a signal is decomposed with one filter and recomposed with another filter, artifacts will be heard. This is likely why the artificial excitation resulted in artifacts in the previous section. The pulses were created with a pitch detector and the phases did not line up. Probably the most sensitive synchronization required is to line up the GCI with the pulses in the residual. The phase and synchronization of these signals audibly changes how well the two signals blend.

Now that the synthesis algorithm has been described, the following section will present a listening experiment that was carried out to demonstrate that APLP is an improvement over constant pre-emphasis LP in transforming high-effort voices into breathy voices.

6.2 Listening Experiments

We have described how the APLP algorithm can transform the spectrum of the voice. This section describes a perceptual experiment to verify whether these changes are are subjectively perceivable.

As source data, there were pairs of voice samples where the same person phonated the same vowel at the same fundamental frequency but with two different voice qualities: one breathy and one high-effort. The goal was to transform the high-effort sample into the breathy sample. There were three different sample pairs as source data as described in Table 6.2.

Table 6.2: Original voice samples for voice transformation experiment

Singer	Female A		Male A		Female B	
Tessitura	mezzo-soprano		baritone		soprano	
Vowel	[e]		[a]		[a]	
Note	A#3		G#3		G#4	
Phonation	breathy	h.e.*	breathy	h.e.	breathy	h.e.
Fundamental freq. (Hz)	237	237	210	208	415	421
F1 (Hz)	475	475	630	830	830	860
F2 (Hz)	1900	1900	1270	1240	1250	1270

*h.e. = high effort

Table 6.3: Comparison of voice samples in voice transformation listening experiment

The listener is presented with pairs of voice samples.

orig. breathy voice	vs.	orig. high effort voice (o.h.e.v.)
orig. breathy voice	vs.	o.h.e.v. through constant pre-emphasis LP
orig. breathy voice	vs.	o.h.e.v. through first order pre-emphasis APLP
orig. breathy voice	vs.	o.h.e.v. through third order APLP
orig. breathy voice	vs.	o.h.e.v. through commercial processor

The listening experiment compared the capabilities of three different algorithms to transform high-effort voices into breathy voices: APLP with first-order pre-emphasis, APLP with third-order pre-emphasis, and constant pre-emphasis LP. A high-effort voice without processing was also included in the comparison. The listeners rated the effectiveness of the four methods (including no processing) with respect to the target breathy sample for the voice being transformed. The various comparisons made for each sample pair are listed in Table 6.3.

We applied APLP synthesis as described in Figure 6.1 to transform high-effort voices into breathy voices. The spectral emphasis, $H_E(z)$, used during synthesis was extracted from the target breathy voice in the sample pair. Constant pre-emphasis LP does not estimate a spectral emphasis filter. This is equivalent to using the same spectral emphasis for analysis and synthesis, meaning that the spectral emphasis is not modified. The same quantity of aspiration noise was added to both the APLP algorithm and the constant pre-emphasis LP algorithm. A commercial voice processor was also available to process one of the voice samples. In a prior session, TC-Helicon's VoicePro had been used to add as much breathiness as possible to a high-effort voice sample. The VoicePro's algorithm is equivalent to the constant pre-emphasis LP algorithm with some additional fixed filtering to shape the source spectrum and the spectrum of the added noise. Since the algorithm is proprietary, the exact shapes of the filters cannot be described here. As much breathiness as possible was added to the voice using the VoicePro while trying to maintain a natural sound.

The onset of increasing amplitude for high-effort voices is typically much faster than the onset for breathy voices. A steady-state section was extracted from the center of each voice sample to eliminate the influence of voice onsets upon the perceptual rating.

There were sixteen listeners in total. The listeners for the experiment were audio engineers with experience in voice processing (ten listeners), trained linguists (three listeners), and experienced singers (three listeners). The processed voice samples were rated relative to benchmark breathy samples. The unprocessed high-

effort voice samples were also rated relative to benchmark breathy samples. The variously processed sample pairs were presented in a random order. In addition, the processed samples were randomly presented before or after the benchmark breathy samples without specifying the order to the listener.

The listeners went through the experiment three times to make three different ratings:

- BREATHINESS: please listen to the two samples and rate how much more BREATHY one sample sounds than the other sample. BREATHINESS corresponds to a soft, relaxed voice.

- VOCAL EFFORT: please listen to the two samples and rate how much more EFFORT is required on the part of the singer to generate one sample rather than the other sample. VOCAL EFFORT corresponds to a strained or tense voice.

- ARTIFICIALNESS: all of the samples have been digitally modified in some way. Please listen to the two samples and rate how ARTIFICIAL one sample sounds than the other.

In each iteration of the experiment, the order of the samples was re-randomized. The relative rating was on a seven-point scale as per ITU Standard 1284 [72]. This is a double-ended scale ranging from one sample sounding much more breathy, to both samples sounding the same, to the other sample sounding much more breathy. Since the synthesized sample was presented randomly before or after the target sample, the scale had to be double ended. Otherwise, the listener would

become biased by being told which sample should sound more breathy. During the the analysis, the ranking was re-ordered to become relative to the target breathy voice. This resulted in a mapping of the seven-point, double-ended scale to a four-point, single-ended scale.

Each listener evaluated 36 pairs of samples. Each pair of voice samples (breathy and high effort sung by the same person) results in four pairs: breathy vs. high-effort, breathy vs. constant pre-emphasis LP, breathy vs. first order APLP, and breathy vs. third order APLP. This results in twelve pairs of samples (3 sample pairs $\times$ 4 processing methods). In addition, one of the sample pairs had been processed through the commercial processor, bringing the total up to 13 sample pairs. The experiment is repeated for each "quality" being evaluated: breathiness, vocal effort, and artificialness. This means that each listener evaluates 39 sample pairs (13 sample pairs $\times$ 3 voice qualities evaluated).

The results of the experiment are presented in Figure 6.3. A test for statistical significance was carried out. The F-test on the breathiness, vocal effort and artificialness ratings resulting in an F-values of 16.7, 16.1, and 10.0 respectively. This indicates that there is less than a 0.01% chance that the observed differences could occur due to noise in each of the sets of ratings.

All of the processing techniques provided a increase in perceived breathiness (Figure 6.3a). The listeners rated the original high-effort voice as being between less breathy and much less breathy than the corresponding breathy voice. After the transformation, all of the voice samples sounded only slightly less breathy than the breathy voice. The third order APLP algorithm performed slightly better than the

other techniques. This may be due to the third order APLP algorithm modeling the drop-off in the high-effort voice spectrum while the other techniques do not.

The similar ratings for breathiness between the APLP and the constant pre-emphasis LP algorithms might give the impression that APLP is not much more effective than constant pre-emphasis LP. However, there are other factors to consider. A breathy voice should exhibit low effort and the transformation should, ideally, be free of unnatural artifacts.

The APLP algorithm was able to reduce the perceived effort of the high-effort voice more effectively than the commercial processor, which in turn performed slightly better than constant pre-emphasis LP (Figure 6.3b). The constant pre-emphasis LP algorithm exhibited nearly the same amount of vocal effort as the original high-effort voice.

The best performance of the APLP algorithm was in the perceptual rating of artificialness (Figure 6.3b). Constant pre-emphasis LP sounded more artificial than the original high-effort voice. The listeners rated the commercial processor as sounding even more artificial. Interestingly enough, the original high-effort voice (that is, with no processing) was rated as sounding slightly more artificial than the corresponding breathy sample. This is likely because the breathy sample sounded relaxed and "natural" while the high-effort voice sounded more strained in a way that was less comfortable to hear. Listeners rated the APLP algorithm as having the same amount of artificialness as the original high-effort voice, indicating that APLP does not significantly increase the perceived artificialness of the original voice.

In conclusion, all of the voice transformation algorithms increased the perceived breathiness to a similar level. This is likely because aspiration noise is one of the most significant cues of breathiness and all of the algorithms added similar quantities of noise. Constant pre-emphasis LP did not reduce the perceived effort. In contrast, APLP outperformed both constant pre-emphasis LP and the commercial processor in reducing the perceived effort. Constant pre-emphasis LP and the commercial processor increased the perceived artificialness. APLP did not add artificialness into the re-synthesized voices. This indicates that APLP is able to synthesize natural-sounding voices when transforming high effort voices into breathy voices.

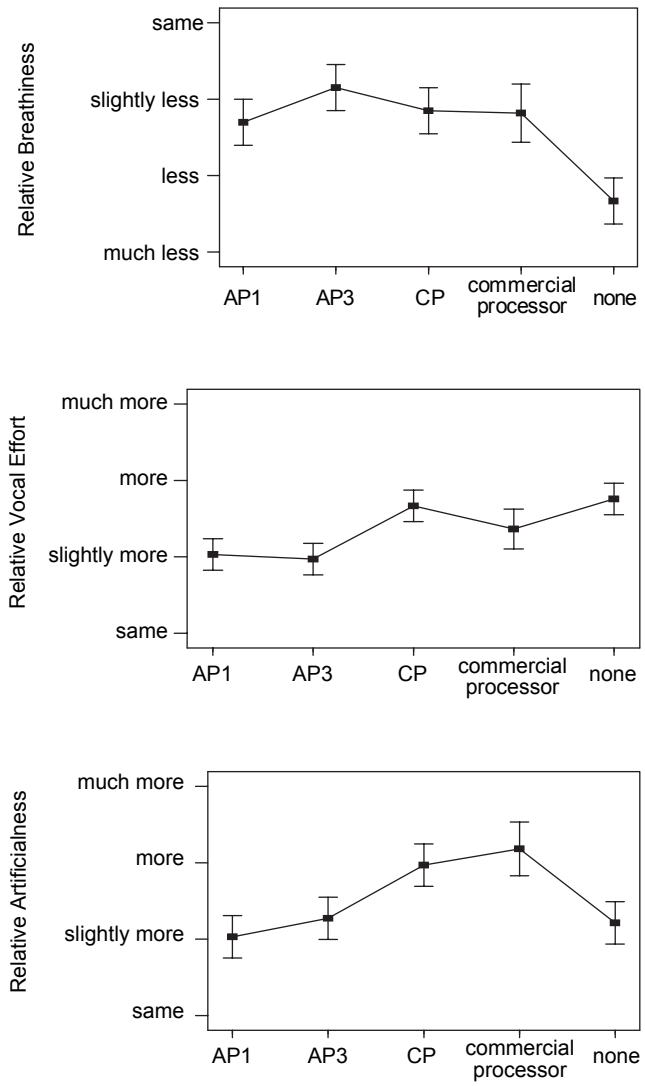


Figure 6.3: Statistical results from relative ratings of breathiness (top), vocal effort (middle), and artificialness (bottom). The horizontal axis represents the processing applied to the high-effort voice: AP1 = first order adaptive pre-emphasis, AP3 = third order adaptive pre-emphasis, CP = constant pre-emphasis, commercial processor and original = no processing. All samples were rated relative to a target breathy voice. 95% confidence intervals have been plotted.

Chapter 7

Conclusion

This dissertation presented an algorithm for modeling the perceived vocal effort in singing voices. It was found that constant pre-emphasis LP results in a vocal tract filter that varies with changes in vocal effort. APLP is able to track variations in the spectral emphasis of the voice signal. This results in an estimated formant filter that is more consistent across varying voice qualities. APLP also estimates a spectral emphasis filter that can be used to manipulate the perceived effort in the voice.

An algorithm was implemented using APLP to transform the spectral envelope of the voice, with the goal of transforming the perceived vocal effort during the synthesis of breathy voices. A listening experiment was carried out to compare the performance of APLP to constant pre-emphasis LP. It was found that APLP resulted in breathy voices that exhibited less perceived vocal effort and fewer artifacts than the corresponding transformation using constant pre-emphasis LP.

The APLP algorithm effectively transforms high-effort voices into breathy voices, thereby providing an interesting effect for musicians to use. One common vocal effect is creates harmony voices by pitch shifting the original voice to other notes. The breath effect could be applied to these harmony voices to create a more distinctive sound for these voices.

The APLP algorithm also estimates a spectral emphasis filter that can be used as an indicator for the quantity of vocal effort present in the voice. This indicator could be used to selectively apply effects depending upon the quantity of effort in the voice. For example, one might want to apply a distorted or other radical effect to the voice only when the singer exhibits a large degree of effort. Alternately, there might be effects that one would only want to apply when the performer is singing softly.

APLP could be applied along with other techniques for voice transformation. For example, one such technique analyzes and re-synthesizes the aperiodic component of the voice source while depending upon LP for initial analysis [73, 42]. This technique can be used to modify the perception of vocal effort by adding aperiodic variation to simulate roughness and harshness in the voice. Using APLP instead of standard LP in this technique would produce a formant filter that is more consistent across varying voice qualities. In addition, the spectral emphasis filter estimated by APLP would establish a valuable starting point for any further manipulations to the spectral tilt.

The techniques developed here could also be used in other applications where it is useful to analyze or manipulate the perceived vocal effort in the voice. For

example, the techniques developed here could be used to synthesize artificial voices that are more expressive.

7.1 Possible Improvements

There are a number of ways that the APLP algorithm could be improved. This dissertation focused on reducing the perceived effort in the voice. To increase the vocal effort, a different problem is encountered. It is easy enough to increase the spectral emphasis of the voice. However, if the original voice is breathy, the resulting voice will exhibit too much noise. To synthesize high-effort voices, it is necessary to reduce the noise from the original voice and to increase the harmonic content in the voice, especially at higher frequencies where breathy voices lack harmonics.

Low-order LP was used to estimate the spectral emphasis filter. However, for high-effort voices, low-order LP does not always model the full steepness of the drop-off around $4 - 5$ kHz (see Figure 5.6(b)). A more precise estimate of the shape of the the drop-off could result in a more effective transformation of vocal effort.

The perception of vocal effort is caused by both the glottal source and the vocal tract filter. In the APLP algorithm presented here, we lump both of these influences into one spectral emphasis filter. An alternate implementation of APLP could estimate the pre-emphasis based on the voice spectrum up to 5 kHz while ensuring that the pre-emphasis remains a simple slope. This would result in an

estimated source that is closer to the true glottal source.

APLP was applied to signals sampled up to 22 kHz. Implementing a warped [66] version of APLP could result in an algorithm that works for even wider bandwidth signals.

The algorithm presented here transforms the vowel content of the voice. However, some of the cues of vocal effort include the consonant information in the voice. High effort voices exhibit notes with a more abrupt attack while breathy voices exhibit notes that start and end more slowly. To appropriately transform high-effort voices into breathy voices, it may even be necessary to change the relative durations of the vowels and the consonants [16]. APLP effectively transforms the vowel content but the consonants also need to be modified.

The research here focuses on a relatively small number of vowels. The samples used in most of the listening experiments were open vowels where the perception of breathiness is more apparent. It is possible that the influence of the APLP algorithm may not be consistent across different vowels. When a voice changes from one vowel to another, some of the spectral changes are similar to changes that occur between high and low effort voices. For example, in changing between higher and lower vowel qualities, there is an increase in the first formant frequency that also increases the amplitude of the higher formants [74]. In implementing APLP for phrases, it may be necessary to adjust for changes between different vowels. It would be valuable to carry out further research to better understand how APLP behaves across large vowel changes.

APLP transforms the spectral envelope of the voice using a spectral emphasis

filter to reduce the perceived vocal effort. This dissertation presents an algorithm that uses pulsed noise to create the perception of aspiration noise during the synthesis of breathy voices. The pulsed noise is synchronized with the glottal closure instances that have been estimated from the voice signal. However, in live performance and with signals that do not retain phase information, the glottal closure detection will not be reliable. In these circumstances, artificial noise can still be added but it will not be pulsed. This non-pulsed noise will still simulate aspiration noise, but it will not be as effective as using pulsed noise.

Bibliography

[1] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, pp. 561–580, April 1975.

[2] J. D. Markel and A. H. Gray, *Linear Prediction of Speech.* New York: Springer-Verlag Berlin Heidelberg, 1976.

[3] M. W. Macon, L. Jensen-Link, J. Oliverio, M. A. Clements, and E. B. George, "A singing voice synthesis system based on sinusoidal modeling," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP97)*, vol. 1, Munich, Germany, Apr. 1997, pp. 435–438.

[4] F. Thibault and P. Depalle, "Adaptive processing of singing voice timbre," in *Canadian Conference on Electrical and Computer Engineering (CCECE 2004)*, Niagara Falls, Ontario, Canada, May 2004, pp. 871– 874.

[5] D. G. Childers, "Glottal source modeling for voice conversion," *Speech communication*, vol. 16, pp. 127–138, 1995.

[6] D. G. Childers, K. Wu, D. M. Hicks, and B. Yegnanarayana, "Voice conversion," *Speech Communication*, vol. 8, pp. 147–158, 1989.

[7] D. G. Childers, B. Yegnanarayana, and K. Wu, "Voice conversion: Factors responsible for quality," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 10, pp. 748–751, April 1985.

[8] J. M. Gutierrez-Arriola, Y. S. Hsiao, J. M. Montero, J. M. Pardo, and D. G. Childers, "Voice conversion based on parameter transformation," *Proceedings of 5th International Conference on Spoken Language Processing*, 1998.

[9] A. Kumar and A. Verma, "Using phone and diphone based acoustic models for voice conversion: A step towards creating voice fonts," *IEEE International*

Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, pp. 720–723, April 2003.

[10] M. Bulut, S. Narayanan, and A. Syrdal, "Expressive speech synthesis using a concatenative synthesizer," in *7th International Conference on Spoken Language Processing (ICSLP 2002)*, Denver, Colorado, Sept. 2002, pp. 1265–1268.

[11] J. Janer and A. Loscos, "Morphing techniques for enhanced scat singing," in *Proc. 8th International Conference on Digital Audio Effects (DAFX-05)*, Madrid, Spain, Sept. 2005, pp. 190–193.

[12] J. Laver, *The Phonetic Description of Voice Quality*. New York: Cambridge University Press, 1980.

[13] H. Traunmüller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women and children," *Journal of the Acoustical Society of America*, vol. 107, no. 6, pp. 3438–3451, June 2000.

[14] J.-S. Liénard and M.-G. Di Benedetto, "Effect of vocal effort on spectral properties of vowels," *Journal of the Acoustical Society of America*, vol. 106, no. 1, pp. 411–422, July 1999.

[15] G. D. Allen, "Acoustic level and vocal effort as cues for the loudness of speech," *Journal of the Acoustical Society of America*, vol. 49, no. 6, pp. 1831–1841, June 1971.

[16] G. Fairbanks and M. S. Miron, "Effects of vocal effort upon the consonant-vowel ratio within the syllable," *Journal of the Acoustical Society of America*, vol. 29, no. 5, pp. 621–626, May 1957.

[17] A. Andersson, A. Eriksson, and H. Traunmüller, "Cries and whispers: Acoustic effects of variations of vocal effort," in *Speech, Music and Hearing Quarterly Progress and Status Report (TMH-QPSR)*. Sweden: KTH, 1996, pp. 127–130.

[18] B. Granström and L. Nord, "Neglected dimensions in speech synthesis," *Speech Communication*, vol. 11, pp. 459–462, 1992.

[19] S. Ternström, M. Bohman, and M. Södersten, "Loud speech over noise: Some spectral attributes, with gender differences," *Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1648–1665, Mar. 2006.

[20] A. N. Chasaide and C. Gobl, "Voice source variation," in *Handbook of Phonetic Sciences*, W. J. Hardcastle and J. Laver, Eds. Blackwell Publishers, 1997, pp. 427–461.

[21] B. Doval, C. d'Alessandro, and N. Henrich, "The voice source as a causal/anticausal linear filter," in *In Proc. Voice Quality: Functions, Analysis and Synthesis, ISCA workshop (VOQUAL'03)*, Geneva, Switzerland, Aug. 2003, pp. 15–20.

[22] ——, "The spectrum of glottal flow models," *Acustica United with Acta Acustica*, vol. 92, pp. 1026–1046, 2006.

[23] J. Sundberg, *The Science of the Singing Voice*. Dekalb, IL: Northern Illinois University Press, 1987.

[24] D. G. Childers and C.-F. Wong, "Measuring and modeling vocal source-tract interaction," *IEEE Transactions on Biomedical Engineering*, vol. 41, no. 7, pp. 663–671, July 1994.

[25] J. H. Esling, "The laryngeal sphincter as an articulator: How register and phonation interact with vowel quality and tone," in *Western Conference on Linguistics*. UBC, November 2002.

[26] J. H. Esling and J. G. Harris, "Expanded taxonomy of states of the glottis," *15th International Congress of Phonetic Sciences*, vol. 1, pp. 1049–1052, 2003. [Online]. Available: http://web.uvic.ca/ling/research/phonetics/SOG/

[27] I. R. Titze and B. H. Story, "Acoustic interactions of the voice source with the lower vocal tract," *Journal of the Acoustical Society of America*, vol. 101, no. 4, pp. 2234–2243, April 1997.

[28] B. H. Story, I. R. Titze, and E. A. Hoffman, "The relationship of vocal tract shape to three voice qualities," *Journal of the Acoustical Society of America*, vol. 109, no. 4, pp. 1651–1667, April 2001.

[29] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, February 1990.

[30] G. Fant and A. Kruckenberg, "Voice source properties of the speech code," *Speech, Music and Hearing Quarterly Progress and Status Report (TMH-QPSR)*, vol. 4, pp. 45–56, 1996.

[31] J. Kreiman and B. R. Gerratt, "Sources of listener disagreement in voice quality assessment," *Journal of the Acoustical Society of America*, vol. 108, no. 4, pp. 1867–1876, October 2000.

[32] R. Shrivastav, "The use of an auditory model in predicting perceptual ratings of breathy voice quality," *Journal of Voice*, vol. 17, no. 4, pp. 502–512, June 2003.

[33] R. Shrivastav and C. M. Sapienza, "Objective measures of breathy voice quality obtained using an auditory model," *Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2217–2224, October 2003.

[34] J. Kreiman, B. Gerratt, and M. Ita, "When and why listeners disagree in voice quality assessment tasks," *Journal of the Acoustical Society of America*, vol. 122, no. 4, pp. 2354–2364, 2007.

[35] R. Shrivastav, "Multidimensional scaling of breathy voice quality: Individual differences in perception," *Journal of Voice*, vol. 20, no. 2, pp. 211–222, 2006.

[36] R. Shrivastav and C. Sapienza, "Some difference limens for the perception of breathiness," *Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 416–423, 2006.

[37] H. Imagawa, K.-I. Sakakibara, N. Tayama, and S. Niimi, "The effect of the hypopharyngeal and supra-glottic shapes on the singing voice," in *Proc. of the Stockholm Music Acoustics Conference (SMAC03)*, Stockholm, Sweden, Aug. 2003, pp. 471–474.

[38] J. Picone, "Signal modeling techniques in speech recognition," *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1215–1247, September 1993.

[39] S. E. Bou-Ghazale and J. H. L. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Transaction on Speech and Audio Processing*, vol. 8, no. 4, pp. 429–442, July 2000.

[40] M. Heldner, "Spectral emphasis as an additional source of information in accent detection," in *Prosody 2001: ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*, Red Bank, NJ, Oct. 2001, pp. 57–60.

[41] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Methods for stress classification: Nonlinear teo and linear speech based features," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP99)*, vol. 4, Phoenix, AZ, Mar. 1999, pp. 2091–2094.

[42] G. Richard and C. d'Alessandro, "Analysis/synthesis and modification of the speech aperiodic component," *Speech Communication*, vol. 19, pp. 221–224, 1996.

[43] J. H. Esling, "There are no back vowels: The laryngeal articulator model," *Canadian Journal of Linguistics*, vol. 50, pp. 13–44, 2005.

[44] G. Fant, *Acoustic Theory of Speech Production*. The Hague: Mouton De Gruyter, 1970.

[45] D. Y. Wong, J. D. Markel, and A. H. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 4, pp. 350–355, Aug. 1979.

[46] M. A. Little, P. E. McSharry, I. M. Moroz, and S. Roberts, "Testing the assumptions of linear prediction analysis in normal vowels," *Journal of the Acoustical Society of America*, vol. 119, no. 1.

[47] P. R. Cook, "Toward the perfect audio morph? singing voice synthesis and processing," in *Proc. of the COST-G6 Workshop on Digital Audio Effects (DAFx-98), Barcelona, Spain*, 1998.

[48] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394–2410, 1991.

[49] J. L. Kelly and C. C. Lochbaum, "Speech synthesis," *Proceedings of the 4th International Congress on Acoustics*, pp. 1–4, 1962, paper G42.

[50] A. P. Lobo and W. A. Ainsworth, "Evaluation of a glottal ARMA model of speech production," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP92)*, San Francisco, CA, Mar. 1992, pp. 13–16.

[51] L. B. Jackson, "Noncausal arma modeling of voiced speech," *IEEE Transactions on Acoustics, Speech and Signal Processing (ASSP)*, vol. 37, no. 10, pp. 1606–1608, Oct. 1989.

[52] J. L. Flanagan, *Speech Analysis Synthesis and Perception*. New York: Springer-Verlag, 1972.

[53] P. Lupini, D. J. Shpak, and B. Gibson, "Targeted vocal transformation," Submitted for international patents in May 1998, January 2002, u.S. patent 6,336,092.

[54] L. A. Bateman, "Soprano, style and voice quality: Acoustic and laryngographic correlates," Master's thesis, University of Victoria, 2004.

[55] R. L. Mason, F. G. Gunst, and J. L. Hess, *Statistical Design and Analysis of Experiments with Applications to Engineering and Science*. New York: John Wiley and Sons, 1989.

[56] P. Kabal, "Ill-conditioning and bandwidth expansion in linear prediction of speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP03)*, Hong Kong, Apr. 2003, pp. 824–827.

[57] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *Speech, Music and Hearing Quarterly Progress and Status Report (STL-QPSR)*, vol. 4, pp. 1–13, 1985.

[58] Q. Lin, "Speech production theory and articulatory speech synthesis," Ph.D. dissertation, Royal Institute of Technology (KTH), Stockholm, Sweden, Jan 1990.

[59] G. Fant, "The LF-model revisited: Transformations and frequency domain analysis," *Speech, Music and Hearing Quarterly Progress and Status Report (TMH-QPSR)*, vol. 2-3, pp. 119–156, 1995.

[60] P. Boersma and D. Weenink, "Praat: A system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[61] K. I. Nordstrom, G. A. Rutledge, and P. F. Driessen, "Using voice conversion as a paradigm for analyzing breathy singing voices," in *Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM05), Victoria, BC*, 2005, pp. 428 – 431.

[62] J. S. Milton and J. C. Arnold, *Introduction to Probablility and Statistics: Principles and Applications for Engineering and the Computing Sciences*. New York: McGraw Hill, 1990.

[63] P. Alku, "An automatic method to estimate the time-based parameters of the glottal pulseform," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP92)*, San Francisco, CA, Mar. 1992, pp. 29–32.

[64] ——, "Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering," *Speech Communication*, vol. 11, pp. 109–118, 1992.

[65] H. R. Pfitzinger, "Influence of differences between inverse filtering techniques on the residual signal of speech," in *Proc. of DAGA 2005*, vol. 1, Munich, Germany, Mar. 2005, pp. 223–224.

[66] A. Harma and U. K. Laine, "A comparison of warped and conventional linear predictive coding," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 579–588, July 2001.

[67] P. Kabal, "Ill-conditioning and bandwidth expansion in linear prediction of speech," TSP Lab, Department of Electrical Engineering, McGill University, Ontario, Canada, Tech. Rep., October 2003.

[68] Y. Tohkura, F. Itakura, and S. Hashimoto, "Spectral smoothing technique in PARCOR speech analysis-synthesis," *IEEE Transactions on Acoustics, Speech, and Signal Processing (TASSP)*, vol. 26, no. 6, pp. 587–596, December 1978.

[69] P. R. Cook, "Identification of control parameters in an articulatory vocal tract model, with applications to the synthesis of singing," Ph.D. dissertation, Department of Electrical Engineering, Stanford University, 1991.

[70] D. J. Hermes, "Synthesis of breathy vowels: Some research methods," *Speech Communication*, vol. 10, pp. 497–502, 1991.

[71] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Institute of Phonetic Sciences (IFA) Proceedings*, no. 17, University of Amsterdam, 1993, pp. 97–110.

[72] *General methods for the subjective assessment of sound quality*, International Telecommunication Union Std. ITU-R BS.1284-1.

[73] C. d'Alessandro, V. Darsinos, and B. Yegnanarayana, "Effectiveness of a periodic and aperiodic decomposition method for analysis of voice source," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 12–23, Jan. 1998.

[74] K. N. Stevens, *Acoustic Phonetics*. Cambridge, MA: MIT Press, 2000.