Abstract

The synthesis of singing is investigated using an articulatory model which simulates the human vocal tract as a network of digital filter simulations of acoustic tubes. Methods for identifying the vocal tract shape parameters are presented, and a new method of adaptively tracking vocal tract shape from features of an input speech signal is defined and demonstrated. Methods of modeling the glottal source are discussed, and identification of the glottal source waveform using non-invasive deconvolution techniques is discussed.

The identification and control of glottal source control parameters is investigated. Studies were conducted using four highly trained singers, with additional results from an additional eight trained singers. Standard methods of pitch detection are summarized, and a new method of pitch detection is presented. The behavior of low and high frequency components of the vocal pitch deviation control signal under various phonation conditions is investigated. Noise generation mechanisms in the vocal tract are discussed, and the generation of noise at or near the glottal source is investigated theoretically and experimentally. A fluid dynamic analysis of flow-induced noise generation in the glottal folds is conducted, and the results indicate that noise bursts could occur in the glottal source for frequencies below 200 Hz. Methods of extracting the periodic and non-periodic components of quasi-periodic signals are discussed, and new extraction, quantification, and visualization methods are presented. The pulsed-noise behavior which was hypothesized from the analytical results was verified by the experimental data, most clearly in low bass-singer tones. Other aperiodicities such as subharmonics are also investigated experimentally.

The synthesis model and user interface features of two computer programs which were written for singing synthesis are described. One program controls a Digital Signal Processor (DSP) chip in real time to synthesize the singing voice. The other program is a text-driven software synthesis system.

IDENTIFICATION OF CONTROL PARAMETERS IN AN ARTICULATORY VOCAL TRACT MODEL, WITH APPLICATIONS TO THE SYNTHESIS OF SINGING

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

By
Perry Raymond Cook
September 1991

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

John Chowning
(Department of Music)
(Principal Adviser)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Bernard Widrow

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Stephen Boyd

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Julius O. Smith
(Department of Music)

Approved for the University Committee on Graduate Studies:

Dean of Graduate Studies

IDENTIFICATION OF CONTROL PARAMETERS IN AN ARTICULATORY VOCAL TRACT MODEL, WITH APPLICATIONS TO THE SYNTHESIS OF SINGING

Perry Raymond Cook, Stanford University, 1991

The synthesis of singing is investigated using an articulatory model which simulates the human vocal tract as a network of digital filter simulations of acoustic tubes. Methods for identifying the vocal tract shape parameters are presented, and a new method of adaptively tracking vocal tract shape from features of an input speech signal is defined and demonstrated. Methods of modeling the glottal source are discussed, and identification of the glottal source waveform using non-invasive deconvolution techniques is discussed.

The identification and control of glottal source control parameters is investigated. Studies were conducted using four highly trained singers, with additional results from an additional eight trained singers. Standard methods of pitch detection are summarized, and a new method of pitch detection is presented. The behavior of low and high frequency components of the vocal pitch deviation control signal under various phonation conditions is investigated. Noise generation mechanisms in the vocal tract are discussed, and the generation of noise at or near the glottal source is investigated theoretically and experimentally. A fluid dynamic analysis of flow-induced noise generation in the glottal folds is conducted, and the results indicate that noise bursts could occur in the glottal source for frequencies below 200 Hz. Methods of extracting the periodic and non-periodic components of quasi-periodic signals are discussed, and new extraction, quantification, and visualization methods are presented. The pulsed-noise behavior which was hypothesized from the analytical results was verified by the experimental data, most clearly in low bass-singer tones. Other aperiodicities such as subharmonics are also investigated experimentally.

The synthesis model and user interface features of two computer programs which were written for singing synthesis are described. One program controls a Digital Signal Processor (DSP) chip in real time to synthesize the singing voice. The other program is a text-driven software synthesis system.

Dedication

This is dedicated to Liz and Mom, for their support throughout,
and to the memory of Robert C. Cook.

Acknowledgements

I thank Professor John Chowning, whose profound interest in singers and singing caused me to undertake this project, and who provided funding for me to work at CCRMA. Thanks to Dynacord Inc. for funding my research at CCRMA. Thank you to Professor Julius O. Smith, my principle reference and advisor on topics of digital signal processing and speech. My thanks to Professor Chris Chafe for providing ideas on linguistics, non-linear systems, and for first bringing up the idea of mechanisms of pulsed noise generation. I thank Professors Bernard Widrow and Stephen Boyd for their input on repeated rewritings of this dissertation. Thank you to Professor Norman Abelson, who was my first voice teacher to mention that singing has quite a lot to do with acoustics. I thank Professors Earl Schubert, Max Mathews, and John Pierce, who were encouraging mentors and living reference books. Thanks to Mary Douthitt and Liz Douthitt for their proofreading services. Finally, thanks to the other folks at CCRMA, Patte Wood, Heide Kugler, Glen Diener, Xavier Serra, Doug Kieslar, Dave Mellinger, David Jaffe, Carl Müller, and all the others who have made my research time at CCRMA interesting and productive.

Preface

The research documented in this dissertation was conducted to construct a singing synthesis system which provides physical parameterization over the articulatory features of the speech organ, and to provide rules for perturbation of the voice source which yield more natural synthesis of the singing voice.

The synthesis of singing has been investigated from within the frameworks of most audio synthesis techniques and methods. Primitive speech systems use time domain methods, such as playback of digitally recorded phonemes or words, but such systems are generally inappropriate for the flexible synthesis of music. Other systems are based on the final spectrum. Still others, specifically source/filter systems such as linear predictive coding, are more closely related to the physics of the vocal tract and include parameters for controlling formants or other spectral features directly. The mapping of such non-physical or pseudo-physical systems onto the actual physical components of the vocal mechanism is usually not sufficient to allow direct synthesis (rather than by analysis of recordings) based on notions from vocal pedagogy and speech physiology such as tongue position and glottal effort. Toward the goal of creating a system which is controlled from intuitive physical descriptions, a model of the vocal tract filter was developed which simulates the human vocal tract as a network of connected digital filters. The digital filters simulate the solutions of wave equations inside acoustic tubes, and thus provide control based on the shape of the simulated vocal tract.

Perturbations in the source signal are the primary focus of the experimental research presented in this dissertation. The research was motivated by a profound realization which comes quickly to new students of computer music. This realization is that computers most easily make irritating and boring sounds; sounds which play perfectly periodic waveforms

with perfect fidelity. Sounds of this type are offensive to the human ear, specifically they are the types of sounds which cause humans to decide that machine-produced music and speech is unacceptable. One of the principal reasons for the laughable quality of speech synthesis in airport trains and soft-drink machines is the lack of natural deviations in the signal. Such systems sound like machines imitating speech. A few simple rules for perturbing the instantaneous frequency of the glottal source, and for applying additive noise to the glottal source make a profound difference in the quality of the synthesis of the speaking and singing voice. By applying these rules for source perturbation to a voice synthesis model, more natural speech and singing synthesis is achieved.

Contents

List of Tables

List of Figures

Chapter 1

Synthesis of the Singing Voice

When embarking on research of the singing voice, the first question that the researcher must consider is, "Is singing significantly different from speech?" Since singers are humans who have trained their vocal mechanism to emit specialized sounds, it is common to view singing as a special case of speech. There are profound differences between speech and singing, and some of the differences motivate the notion that singing and speech are different areas of investigation. Certainly there are different priorities in the synthesis of singing and speech; that intelligibility is the principle goal in speech and speech synthesis, and quality is the principle goal in singing and singing synthesis (often compromising intelligibility). To motivate the notion of the research of singing not merely as a special case of speech research, but as a study of quite different phenomena, a brief list of differences between singing and speech follows:

- Voiced/Unvoiced Ratio - The time ratio of voiced/unvoiced/silent phonation is roughly 60%/25%/15% in speech, compared to the nearly continuous 95% voiced time of singing [128].

- Singer's Vibrato - Intentionally introduced deviation in the voice pitch. Section refVibratoSection deals extensively with this topic, as well as unintentional pitch deviations.

- Singer's Formant - Acoustical phenonenon brought about by grouping the third, fourth, and sometimes fifth formants together for increased resonance [20][5]. The

singer's formant would be less evident if the glottal source of the singer did not have such a rich spectral content in this range [16]. Solo singers use the singing formant to be heard, particularly above instruments.

- Singer's Vowel Modification - Intentional and unintentional practices of mutating the vowel sound as a function of pitch for comfort, projection, and/or intelligibility [1][24][12]. Some modification in the sound is an artifact of wider harmonic spacing under the vocal tract filter spectrum envelope, rather than a spectral envelope change [22].

- Nasal Airway Use - For western BelCanto singing, the pathway through the nose is not used as often as it is in speech[18] [23]. This is caused partly by arching the soft palate (velum) to aquire the 'singer's resonance'. It is also a defense mechanism allowing the singer to sound much the same with or without a cold.

- Average Pitch and Range of Pitch - The average speaking pitch is different from the average singing pitch. The comfortable speaking pitch is often different than the comfortable singing pitch. The range of speaking is determined by the speaker's comfort and emotional state. The singer's range is first determined by physiology and training. While performing a particular musical piece, the range is determined by the composer.

- Average Volume and Range - The average level of the speaking voice is softer than the average level of the singing voice. The dynamic range of singing is greater than that of typical speech. Greater flow rates and greater excursions of the vocal folds imply that the singing system is likely to operate in higher orders of non-linearity [128].

- Singer Vocal Training - The singer exercises his/her vocal folds regularly in different regimes, thus differences exist between the source signals of singers and non-singers. When asked to phonate loudly, untrained singers (and speakers of loud or angry speech) move toward a pressed (efficient but squeaky) mode of phonation. Trained singers show no such tendency, yielding a more consistent timbre across a wide dynamic range [70].

- Neurology - Some classic studies on head injury document cases of people who are

unable to speak, but still sing perfectly and are even able to learn and perform new songs. These and other studies point up the likelihood of completely different areas of the brain controlling speech and singing [169].

- Statutory - It is unclear at the present time whether singing is protected under the First Amendment to the Constitution of the United States of America. This part of the Bill of Rights protects free speech, but some recent interpretations have tended toward defining certain artistic performances as non-speech.

The synthesis of the singing voice has been investigated in the past using various techniques. The simplest methods of generating vocal sounds involve the playback of entire words or stored waveforms. Such systems can produce intelligible speech, but do not provide the flexibility of pitch and timbral control required for the synthesis of singing. The techniques used successfully for singing synthesis are divided into two broad categories; spectral models and source/filter models. A typical spectral model is a Fourier analysis/resynthesis system, involving the identification and synthesis of important features in the frequency spectrum of the vocal signal. A source/filter model considers the fact that the vocal tract is a resonant system driven by various sound sources, such as glottal pulses or noise, and provides control over the source and filter elements.

1.1 The Fourier Model of Speech and Singing

The Fourier model is capable of identity resynthesis if analysis is performed by frequency transformation and resynthesis is performed by inverse transformation of all frequency components without modification. However, the task of speech analysis/resynthesis is usually undertaken to either decrease the data required to represent the signal, or to gain some type of flexible control over the synthesis process. Systems for the synthesis of singing almost always fall into the latter category, using the flexible control features to shape the sound musically or provide some type of special effect. Using the Fourier model, a signal is analyzed using Short Time Fourier Analysis (STFA), the spectrum can be modified, and an inverse transform is performed on the modified spectrum yielding a modified time-domain signal [26][27]. The phase vocoder is a frequency-representation based analysis/modification/resynthesis system which has experienced popularity as a music synthesis

technique [31][30][35]. Typical speech and music synthesis applications taking advantage of parametric control over the model involve time compression/expansion, pitch shifting, and cross-synthesis [37][160].

In some spectrally based analysis/resynthesis systems, a decision is made as to whether the analyzed voice signal is periodic or noisy, i.e. voiced or unvoiced. If the signal is periodic, the harmonic sinusoidal peaks are located, and the signal is resynthesized using sinusoidal generators at the appropriate frequencies, amplitudes, and phases. Techniques for economically generating sets of harmonics can be exploited for synthesis [29]. If the signal is noisy, as in the case of a consonant, resynthesis is accomplished by passing white noise through a filter designed to match the spectral shape of the analyzed consonant. Alternatively, inverse transformation of a spectrum exhibiting the magnitude response of the noise signal being modeled, and random phase components in each resynthesis block, yields a noise signal appropriate for modeling consonants. The deterministic plus residual model [160] provides flexible control over the resynthesis process. Spectral envelopes can be extracted from the harmonics and modified before resynthesis, allowing simple resynthesis, independent pitch shifting and time shifting without spectral distortion, or cross-synthesis.

1.2 Formant Based Models

In identifying dissimilar sounds such as vowels, the ear is most sensitive to peaks in the signal spectrum [167][176][172]. Resonant peaks in the spectrum are called formants, typically indexed as $F_1, F_2, \cdots, F_n$, in ascending order of frequency. Figure 1.1 shows a typical voice spectrum. Some source/filter systems allow direct control over the formants, thus providing a parameterization well suited to the perceptual features most important the human ear. A popular speech synthesis system using such controls is the Klatt synthesizer [33], which allows control over parallel or cascade arrangements of resonant filters. The frequency modulation voice synthesis method of Chowning [29][28] uses a sinusoidal carrier for each formant, and a single sinusoid modulating all carriers at the desired fundamental frequency. As long as the carrier frequencies are selected to be integer multiples of the fundamental frequency, the side bands surrounding each carrier fuse into a single harmonic spectrum exhibiting peaks near the desired formant locations. A successful formant based system for voice synthesis is VOSIM (VOcal SIMulation), which resulted from research into the

minimum data representation of speech phonemes [32]. VOSIM combines sin^2 pulses to directly manipulate formant regions in the spectrum. A method similar to VOSIM is the FOrmant wave Function (FOF) system of Rodet [40]. The FOF system performs parallel synthesis in the time domain of functions which transform to formant spectral regions in the frequency domain. Both analysis/synthesis systems and synthesis by rule [41] [59] systems have been constructed using FOFs.

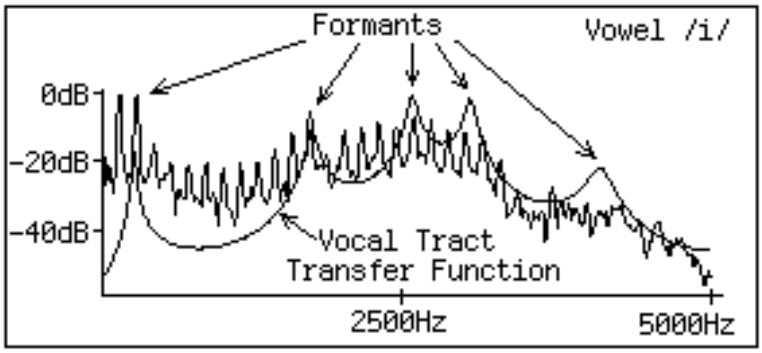

Figure 1.1: Spectrum of vocal utterance of the vowel /i/ (beet). The smooth line on the lower spectrum is of the vocal tract transfer function. The resonant peaks of this curve are called formants.

1.3 Source/Filter Models

Source/filter models of the vocal system take into account the acoustic mechanisms which produce the speech signal. In voiced phonation, the glottal folds open and close roughly periodically, producing a pulsed excitation. The acoustic tube of the oropharynx and the various chambers of the naso-pharynx form a resonant system which filters the glottal pulse, shaping the spectrum of the final output sound. Figure 1.2 shows a midsagittal cross-section of the human head, with the acoustically important features labled. Figure 1.3 shows time and frequency domain plots of a typical glottal waveform, the filter function of the vocal tract, and the resulting output speech waveform and spectrum.

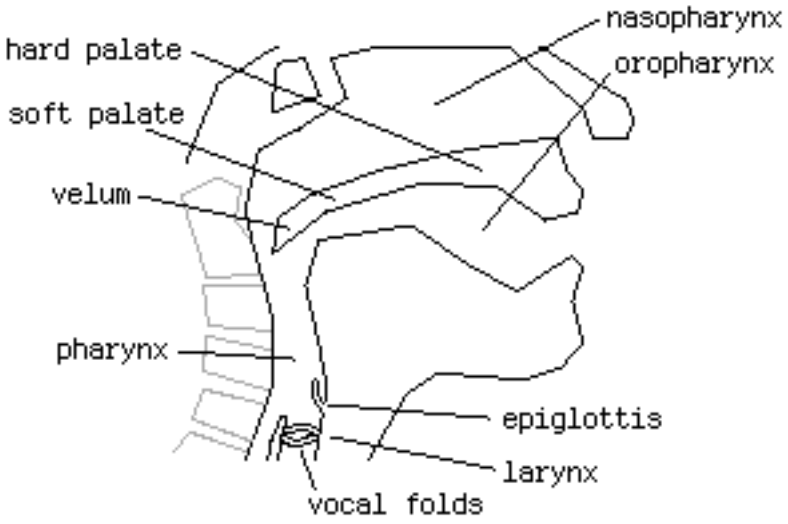

Figure 1.2: A cross-section of the human head, with the acoustically important features labled.

1.3.1 Linear Predictive Coding (LPC)

A highly successful model used for speech and singing synthesis is Linear Predictive Coding (LPC) [56][44][42][45]. This method arises from the mathematical technique of linear least squares estimation, but yields a synthesis system quite closely matched to the physics of the vocal mechanism. In LPC, the sampled periodic speech wave is predicted as a linear combination of past samples:

$$\hat{x}(n) = \sum_{i=1}^{N} x(n-i)c(i) \tag{1.1}$$

The coefficients $c(i)$ from Equation 1.1 are assumed to implement the least squares single step-ahead predictor over the data set, and thus solve the set of linear equations:

$$Rc = r \tag{1.2}$$

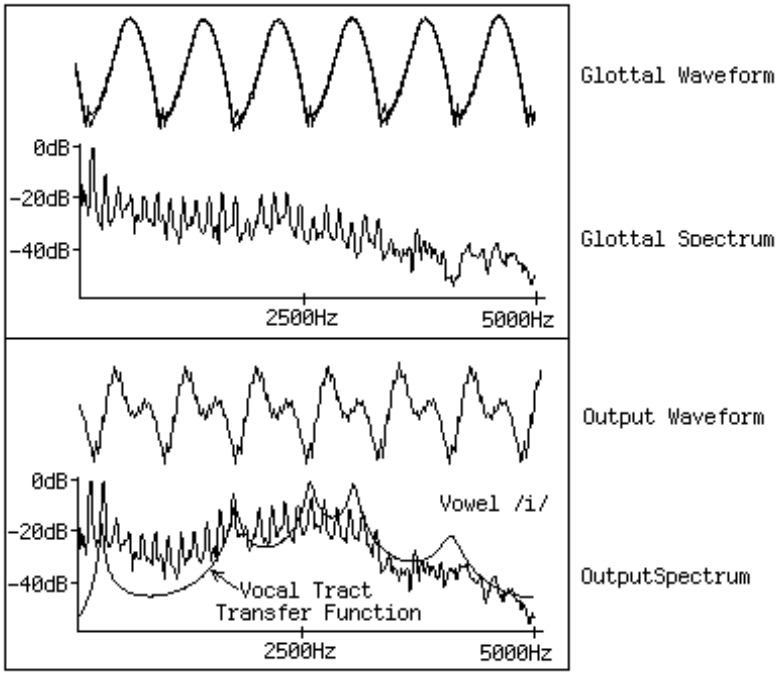

Figure 1.3: Waveforms and spectra of glottal source and vocal tract output for the vowel /i/ (beet). The smooth line on the lower spectrum is the vocal tract transfer function.

where

$$R = E_n \begin{bmatrix} x(n)x(n) & x(n)x(n+1) & \cdots & x(n)x(n+N-1) \\ x(n+1)x(n) & x(n+1)x(n+1) & \cdots & x(n+1)x(n+N-1) \\ x(n+2)x(n) & x(n+2)x(n+1) & \cdots & x(n+2)x(n+N-1) \\ \vdots & \vdots & \ddots & \vdots \\ x(n+N-1)x(n) & x(n+N-1)x(n+1) & \cdots & x(n+N-1)x(n+N-1) \end{bmatrix} \tag{1.3}$$

is the covariance matrix, with E_n denoting the expectation operator over the time variable n, assuming stationarity of x. The single sample delayed covariance vector, r, is given by:

$$r = E_n \begin{bmatrix} x(n)x(n+1) \\ x(n)x(n+2) \\ x(n)x(n+3) \\ \vdots \\ x(n)x(n+N) \end{bmatrix} \tag{1.4}$$

If the matrix R is invertible, the predictor coefficient vector c is given by:

$$c = R^{-1}r \tag{1.5}$$

A signal may be processed in blocks of length M, using the samples of each block to estimate R and r, This yields a set of predictor coefficients for each block of the signal. Other methods of identifying the predictor coefficients employ adaptive algorithms [163][164] and economical methods of estimating the covariance matrix and vector [143][150]. The implementation of a filter using the predictor coefficients c yields an all-pole recursive digital filter of order N. The filter is driven with the predictor error signal ϵ,

$$\epsilon(n) = \hat{x}(n) - x(n) \tag{1.6}$$

to yield an identity resynthesis. Such a synthesis model is called Residual Excited LPC (RELP). RELP does nothing to reduce the data required to represent the signal, nor does this method provide flexible control for resynthesis with modifications. Figure 1.4 shows the waveform and the spectrum of the vowel /a/ (father). The smooth curve on the spectrum is the response of the LPC filter. The residual signal and spectrum is shown. The spectrum of the residual is *whitened*, or flattened, compared with the spectrum of the original vowel, because the spectral *color* is coded into the filter in forming the predictor coefficients.

Noting that the error signal corresponding to periodic speech is dominated by a periodic impulse train, the input to the recursive filter is often replaced by a simple impulse train of the same frequency as the residual pulses. This frequency is commonly called F_0. In the case of unvoiced speech, the LPC analysis yields a least squares fit to the power spectral density of the output signal, and the input to the synthesis filter is white noise.

In both cases, the model of the vocal system is that of a source with a flat spectrum

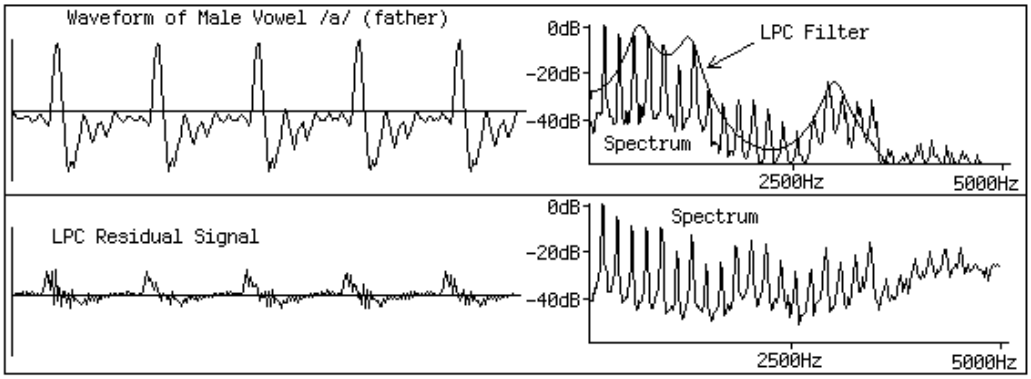

Figure 1.4: Waveforms and spectra of the male vowel /a/ (as in father) and the residual signal from Linear Prediction. The smooth line on the vowel spectrum is the LPC filter response.

corresponding to the glottis or turbulent noise source, and a resonant filter corresponding to the acoustics of the vocal tract. As described, LPC is an analysis/resynthesis system, but does provide flexible control over time and pitch during resynthesis. Factorization of the all-pole LPC filter into complex resonances allows the identification and independent control of formants.

1.3.2 More Physically Based Source Filter Models

The behavior of the vocal tract resonance filter, driven by the weakly coupled glottal source is only partially captured by LPC and some formant-based source-filter models. The mapping of non-physical or pseudo-physical systems onto the actual physical components of the vocal mechanism is usually not sufficient to allow direct synthesis (rather than by analysis of recordings) based on notions arising from vocal pedagogy and speech physiology such as tongue position and glottal effort. More physical models of the vocal tract are possible [55][54], and can be controlled by intuitively natural physical parameters [58][47].

One theory of speech perception, commonly called the motor theory [173], proposes that the human speech perception mechanism includes an articulatory component. This theory contends that humans track likely vocal tract shape trajectories matching trajectories which would produce the speech sounds being heard. Opponents of this theory cite case studies

of subjects who, for physiological reasons, never aquired speech production, but could understand speech. Motor theory proponents argue that the motor mechanism might be more fundamental than the learned speech of a particular society, or that the motor component is only one of many parallel systems performing analysis on the speech signal [175]. Motor theory proponents further argue that if one component is impaired the others adjust to improve performance. The possibility of an articulatory component within the human speech perception mechanism, combined with the desire for more intuitive parameters for controlling synthesis, motivates the use of a more physically based articulatory synthesis model. In the remaining sections of this chapter, a digital filter simulation of the vocal tract filters will be derived from a shape description of the oropharyngeal and nasopharyngeal pathways. The relationship between this model and LPC will be shown. Models of the sources which excite the vocal tract filters will be discussed.

1.4 Derivation of a Digital Simulation of the Acoustic Tube

The WaveGuide Filter (WGF) development of one-dimensional wave propagation in wave guides (in the case of the vocal tract, tubes containing air) provides the framework for controlling the vocal tract filter directly from physical measurements [63][64]. The vocal tract tube is treated as a system of transmission lines [14], yielding closed-form mathematical solutions to the wave equation. The wave equation solutions are easily simulated using digital filters. Given the equations expressing conservation of momentum and mass:

$$a(x)\frac{\partial P(x,t)}{\partial x} = -\rho\frac{\partial U(x,t)}{\partial t} \tag{1.7}$$

$$\frac{\partial U(x,t)}{\partial x} = -\frac{a(x)}{c^2}\frac{\partial P(x,t)}{\partial t} \tag{1.8}$$

where $a(x)$ is the cross-sectional area of the tube at position x, ρ is the density of air, $P(x,t)$ is the pressure at point x at time t, c is the velocity of sound in air, and $U(x,t)$ is the volume velocity past point x at time t, Webster's horn equation can be derived:

$$\frac{\partial}{\partial x}\left[\frac{1}{a(x)}\frac{\partial U(x,t)}{\partial x}\right] = \frac{1}{c^2 a(x)}\frac{\partial^2 U(x,t)}{\partial t^2} \tag{1.9}$$

When $a(x)$ is constant within a section m of the tube, that is, $a_m(x) = a_m$, then Webster's horn equation reduces to the wave equation within each section:

$$\frac{\partial^2 U_m(x,t)}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 U_m(x,t)}{\partial t^2} \tag{1.10}$$

The equivalent pressure expression is:

$$\frac{\partial^2 P_m(x,t)}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 P_m(x,t)}{\partial t^2} \tag{1.11}$$

The solution of Equation 1.11 can be expressed as a decomposition of left and right-going traveling pressure waves,

$$P_m(x,t) = P_m^+\left(t - \frac{x}{c}\right) + P_m^-\left(t + \frac{x}{c}\right) \tag{1.12}$$

where P_m^+ and P_m^- are the right and left-going pressure wave components, respectively. This yields an expression of the wave equation in right and left going traveling pressure waves:

$$\frac{\partial^2 P_m(x,t)}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\left[P^+\left(t - \frac{x}{c}\right) + P^-\left(t + \frac{x}{c}\right)\right] \tag{1.13}$$

To relate pressure to velocity directly, the expression

$$\left[\frac{\partial P_m^+}{\partial x} + \frac{\partial P_m^-}{\partial x}\right] = \frac{\rho c}{a_m}\left[\frac{\partial U_m^+}{\partial x} + \frac{\partial U_m^-}{\partial x}\right] \tag{1.14}$$

can be derived. Define the characteristic impedance of the mth tube section, R_m, as:

$$R_m \equiv \frac{\rho c}{a_m} \tag{1.15}$$

By integrating both sides and ignoring any constant terms as acoustically unimportant D.C. components, the expressions

$$[P_m^+ + P_m^-] = R_m \left[U_m^+ - U_m^-\right] \tag{1.16}$$

$$P_m^+ = R_m U_m^+ \tag{1.17}$$

$$P_m^- = -R_m U_m^- \tag{1.18}$$

can be derived to relate pressure to velocity in each section. Whenever two sections of differing characteristic impedance meet, the boundary conditions to be satisfied are conservation of mass (mass flow, and thus volumetric flow assuming incompressibility, is conserved) and conservation of momentum (pressure is continuous at the junction). These two conditions yield the junction scattering relations:

$$P_1^- = \frac{R_2 - R_1}{R_2 + R_1} P_1^+ + \left(1 - \frac{R_2 - R_1}{R_2 + R_1}\right) P_2^- \tag{1.19}$$

$$P_2^+ = \left(1 + \frac{R_2 - R_1}{R_2 + R_1}\right) P_1^+ - \frac{R_2 - R_1}{R_2 + R_1} P_2^- \tag{1.20}$$

By defining the junction scattering coefficient of the interface between the mth and $m+1th$ sections, k_m, as:

$$k_m = \frac{R_{m+1} - R_m}{R_{m+1} + R_m} \tag{1.21}$$

the scattering relations for pressure and velocity can be written as:

$$P_m^- = k_m P_m^+ + (1 - k_m) P_{m+1}^- \tag{1.22}$$

$$P_{m+1}^+ = (1 + k) P_m^+ - k_m P_{m+1}^- \tag{1.23}$$

$$U_m^- = k_m U_m^+ + (1 + k_m) U_{m+1}^- \tag{1.24}$$

$$U_{m+1}^+ = (1 - k_m) U_m^+ - k_m U_{m+1}^- \tag{1.25}$$

For representation of a tube by a digital filter, the tube is divided into a number of sections, each of the same length determined by the sampling rate, F_S and the speed of sound, c:

$$\text{Section Length} = \frac{c}{F_S} \tag{1.26}$$

This yields a uniform time delay through each section of the tube, equal to the time required for sound waves to propagate through each section. Figure 1.5 shows a smooth acoustic tube, a sampled version of the same tube, the digital filter simulation of the acoustic tube, and the scattering junction connecting adjacent tube sections. The scattering junction is known as the Kelly-Lochbaum junction [54], and can be manipulated algebraically to yield a one-multiply, three-addition structure. The ladder filter structure of Figure 1.5 is discussed and analyzed by Gray [152].

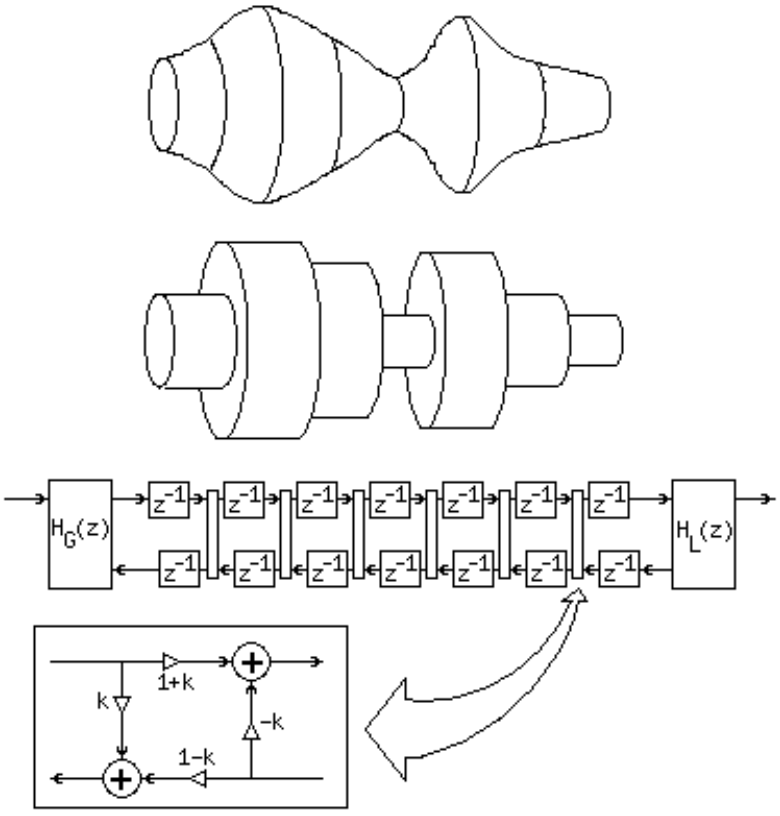

Figure 1.5: A smooth acoustic tube, the sampled version, the digital filter simulation, and the scattering junction connecting adjacent tube sections.

Since the characteristic impedance of each tube section is a function of the cross-sectional

area of a section, and thus the radius, the junction scattering coefficients are computed entirely from the physical tract section measurements. Block $H_G(Z)$ in Figure 1.5 represents the transmission and reflection characteristics of the glottis. The reflection characteristic of the glottis can simply be modeled as a constant positive reflection coefficient (≤ 1) , or more elaborately as a time varying filter. $H_L(Z)$ represents the reflection and transmission characteristics of the lip, which vary with the configuration of the vocal tract. The transmission and reflection functions should be complimentary, that is, in a lossless system, any energy not reflected at the lips is transmitted. A simple model of the lip reflection filter is a low-order low-pass filter, representing the loading of the end of the tube with a piston of air [14]. The cutoff frequency is linearly related to the diameter of the tube end.

Some Vowel Spectrum Examples

As an example of an acoustic tube vocal tract simulation by a digital filter, take an acoustic tube of nine sections. This is consistent with the length of a small female vocal tract of 14 cm. length, sampled at 22.05 kHz. sampling rate. Figure 1.6 shows three vowel configurations, with the smoothed shape of the vocal tract displayed in a cross-section of a human head, the corresponding log-magnitude frequency response of the acoustic tube, and the frequencies of the first three formants.

1.4.1 Propagating Pressure and Velocity

Other derivations, such as the one by Bonder [46], solve the acoustic tube wave equations using the independent variables of pressure and volume velocity. Both variables are propagated in the right-going direction. These formulations yield a transmission matrix model for computing wave propagation, where the pressure and velocity in adjacent sections of the tube are related by:

$$\begin{bmatrix} P_{i-1} \\ U_{i-1} \end{bmatrix} = \begin{bmatrix} \alpha_i & \beta_i \\ \gamma_i & \delta_i \end{bmatrix} \begin{bmatrix} P_i \\ U_i \end{bmatrix} \tag{1.27}$$

where, from the boundary conditions for a uniform tube, the transmission coefficients are:

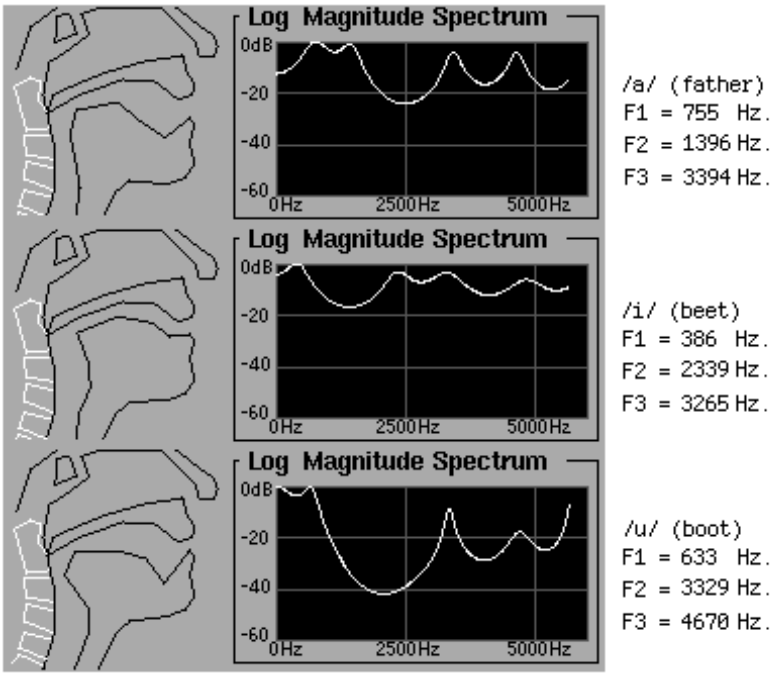

Figure 1.6: Three vocal tract configurations, with the log-magnitude frequency response of the corresponding digital filter, and the frequencies of the first three formant peaks.

$$\alpha = cos(\frac{\omega l}{c}) \tag{1.28}$$

$$\beta = i\frac{\rho c}{a} sin(\frac{\omega l}{c}) \tag{1.29}$$

$$\beta = i\frac{a}{\rho c} sin(\frac{\omega l}{c}) \tag{1.30}$$

$$\delta = cos(\frac{\omega l}{c}) \tag{1.31}$$

where l is the length of each tube section as given in Equation 1.26. It is easily shown that the transmission matrix made up from these coefficients always has a determinant of 1. This expresses the fact that in a lossless acoustic tube, power, the product of pressure and velocity, is conserved across each junction. The scattering matrix is simply a rotation matrix which exchanges energy between the orthogonal variables of pressure and velocity. The relation from input to output of a tube of n cylindrical segments has the form:

$$\begin{bmatrix} P_0 \\ U_0 \end{bmatrix} = \begin{bmatrix} \alpha_1 & \beta_1 \\ \gamma_1 & \delta_1 \end{bmatrix} \begin{bmatrix} \alpha_2 & \beta_2 \\ \gamma_2 & \delta_2 \end{bmatrix} \cdots \begin{bmatrix} \alpha_n & \beta_n \\ \gamma_n & \delta_n \end{bmatrix} \begin{bmatrix} P_n \\ U_n \end{bmatrix} \tag{1.32}$$

$$= \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} P_n \\ U_n \end{bmatrix} \tag{1.33}$$

where

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \prod_{i=1}^{n} \begin{bmatrix} \alpha_i & \beta_i \\ \gamma_i & \delta_i \end{bmatrix} \tag{1.34}$$

By assuming that there is no pressure emmission from the rightmost (lip) end of the tube ($P_n = 0$), the relations of volume velocity from input to output are:

$$U_0 = DU_n \tag{1.35}$$

$$\frac{U_n}{U_0} = \frac{1}{D} \tag{1.36}$$

So the complex roots of D are the resonant frequencies (poles, formants) of the acoustic tube whose transmission characteristics are described by:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{1.37}$$

1.4.2 Power Normalized Form of Waveguide Structure

The decomposition of waves into left and right-going components can be performed on power waves. This representation also results in a ladder filter structure, with somewhat different scattering coefficient relationships. Power is computed as the product of pressure and velocity:

$$\mathcal{P} = PU \tag{1.38}$$

$$\mathcal{P}^+ = P^+U^+ \tag{1.39}$$

$$\mathcal{P}^- = P^- U^- \tag{1.40}$$

If root-power is propagated, denoted by $\tilde{\mathcal{P}}$, the junction scattering relations of Equation 1.19 and 1.20 are transformed using the relationship for root-power:

$$\tilde{\mathcal{P}}^+ = \frac{P^+}{\sqrt{R}} \tag{1.41}$$

yielding:

$$(1-k)\sqrt{\frac{R_1}{R_2}} = \sqrt{1-k^2} \tag{1.42}$$

With a trigonometric substitution of

$$k = sin(\theta) \tag{1.43}$$

the root-power form of the scattering junction equations can be derived:

$$\begin{bmatrix} \tilde{\mathcal{P}}_2^+ \\ \tilde{\mathcal{P}}_1^- \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & cos(\theta) \end{bmatrix} \begin{bmatrix} \tilde{\mathcal{P}}_1^+ \\ \tilde{\mathcal{P}}_2^- \end{bmatrix} \tag{1.44}$$

The rotation matrix nature of the scattering equations is obvious, and indicates the losslessness of the scattering operation. Figure 1.7 shows the scattering junction associated with equation 1.44.

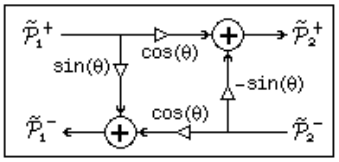

Figure 1.7: Root-power scattering junction. This form corresponds to the application of a rotation matrix applied to left and right-going root-power waves.

1.4.3 Relation of WaveGuide Acoustic Tube to LPC

If the appropriate restrictions are placed on the boundary conditions at the lip and glottis ends of the tube, the structure of the one-dimensional single acoustic tube filter can be manipulated to yield an all-pole response. This structure directly relates to the structure of the canonical Linear Predictive Coder (LPC) realization filter discussed in Section 1.3.1. Figure 1.8 shows the graphical manipulations of "pushing through" delay elements to yield a canonical all-pole filter block diagram. Assuming that the reflection conditions at the lips and glottis are simple (constants for example), the transfer function contains polynomials in Z^{-2}. This motivates the replacement of Z^{-2} by Z^{-1}, effectively halving the sampling rate of the filter. After all of these manipulations have been performed, a filter more efficient than the waveguide ladder filter is realized, but the filter topology is no longer physically based. That is, the direct propagation of wave variables is no longer inherent in the filter operation, and thus the left and right-going wave variables are made difficult to aquire for analysis or further derivations. The phase delay characteristics of the filter are changed in the reduction process if caution is not exercised in defining the input and output points. In fact, it is impossible to make the reduced filter realized at one-half sampling rate of Figure 1.8 exhibit the exact same phase delay as the original filter, because a delay of 1.5 samples is required in the numerator of the transfer function at the reduced sampling rate.

If the conditions at the lips and glottis are modeled more realistically, such as a sophisticated filter at the lips and a time varying or non-linear reflection coefficient at the glottis, the transfer function manipulations shown in Figure 1.8 are not guaranteed to be valid, and the reduction of the acoustic tube filter to an all-pole filter is not possible.

1.4.4 Multiple Waveguides and N-way junctions

The boundary conditions of pressure continuity and flow conservation determine the relationship between pressure and volume velocity at the junction of any number of tubes. Given a junction where n tubes meet, there are n incoming waves whose values are known, and n outgoing waves to be calculated. Denote the incoming pressure and velocity waves in tube i as P_i^+ and U_i^+, and the outgoing waves in tube i as P_i^- and U_i^-. Pressure and velocity are related according to Equations 1.17 and 1.18. The boundary conditions are:

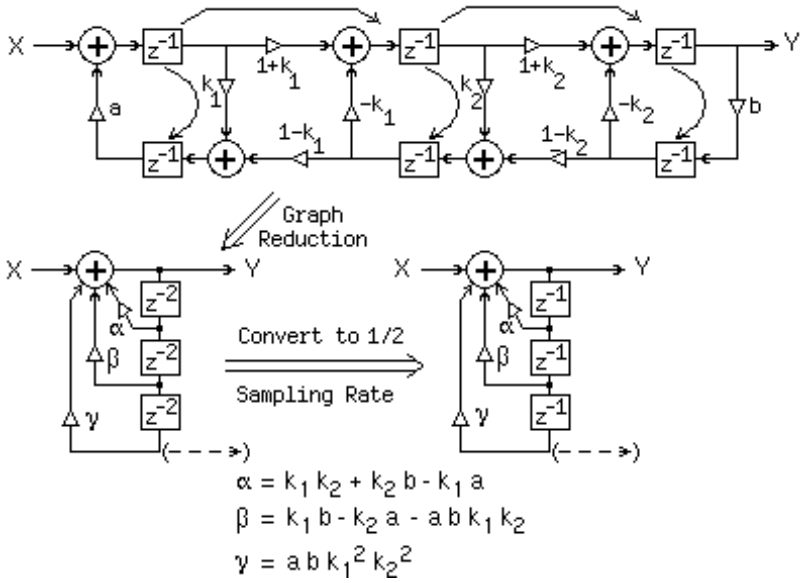

Figure 1.8: A ladder filter realization of an acoustic tube, and an equivalent digital filter realization. The arrows on the ladder filter show how to "push through" single delay elements to combine them into lumped delay elements. Note that the wave propagation nature of the ladder filter is lost in reducing the filter topology. To approximate the phase delay of the original filter most closely, the output of the reduced filter is taken from the lower (dashed) output.

$$P_1 = P_2 = P_3 = \cdots = P_n = P_J \tag{1.45}$$

$$U_1 + U_2 + U_3 + \cdots + U_n = 0 \tag{1.46}$$

where P_J is the junction pressure. Define the characteristic admittance of the i_{th} tube section as the inverse of its characteristic impedance:

$$\Gamma_i = \frac{1}{R_i} = \frac{a_i}{\rho c} \tag{1.47}$$

It can be shown that:

$$P_J = \frac{2\sum_{i=1}^{n} \Gamma_i P_i^+}{\sum_{i=1}^{n} \Gamma_i} \tag{1.48}$$

$$= \sum_{i=1}^{n} \alpha_i P_i^+ \tag{1.49}$$

where

$$\alpha_i = \frac{2\Gamma_i}{\sum_{i=1}^{n} \Gamma_i} \in [0, 2] \tag{1.50}$$

Since $P_i = P_J$ and $P_i = P_i^+ + P_i^-$ for all i, the reflected pressure in any tube is simply the difference between the incoming pressure from that tube and the junction pressure.

$$P_i^- = P_J - P_i^+ \tag{1.51}$$

The reflected volume velocity is given by the product of the characteristic impedance of the tube and the reflected pressure.

1.4.5 The Nasal Tract and Junction

The bifurcation that exists at the velum in the vocal tract can be modeled as a three-way junction. At the velum location, some of the wave energy coming from the glottis might be diverted into the nasal airway, some may continue on to the lips, and the rest will reflect back toward the glottis. A dual acoustic tube model with one three-way scattering junction is shown in Figure 1.9. Figure 1.10 shows the vocal/nasal tract configurations and transfer functions of three nasal vowels. $H_N(Z)$ is the reflection/transmission filter for the nose, which is fixed under normal speech and singing conditions. The reflection function at the nostrils is well modeled by a fixed cutoff low-pass filter.

At sufficiently high sampling rates, extra tubes can be used to model the space below the tongue [55]. More elaborate models of the tongue have been proposed and tested [52], but fall outside of the waveguide acoustic tube model of the vocal tract.

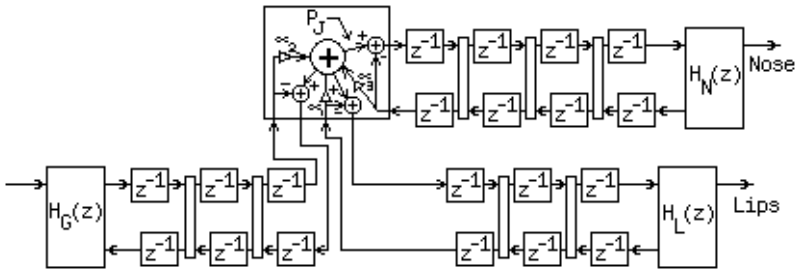

Figure 1.9: The waveguide digital filter block diagram of a system comprised of two acoustic tubes, joined with a three-way scattering junction. This is a filter structure which models the oral and nasal airways of the vocal tract.

1.4.6 Transcutaneous Throat Radiation

A small but significant amount of acoustic energy is radiated from the vocal mechanism through the throat wall. This is especially important in cases of voiced plosives and other times when all other paths out of the vocal tract are closed. Figure 1.11 shows the amplitude envelope of one utterance of a voiced plosive, and the average spectrum of the radiation from the throat of a male speaker during the closed portion of a number of voiced plosives. The microphone was placed six inches from the lips, and the speaker uttered the sounds; "Bee, Dee, Gee, Boo, Doo, Goo, Baa, Daa, Gaa". Spectra were calculated for the closed portions of all nine utterances, then averaged. The average power of the closed portions of the utterances was -13.9 dB referenced to the average power of the open vowel portion of the utterances. The spectrum shows a low-passed response, and suggests that the throat radiation is modeled to a good approximation by a low-order recursive low-pass filter. To account for phase delay effects, a delay line is added to simulate the sound path length from the throat to the point in space where the mixed vocal sound is desired. Figure 1.12 shows the digital filter realization of the vocal tract, including the three-way scattering junction at the velum, and the low-pass filter and delay line which model radiation through the throat wall.

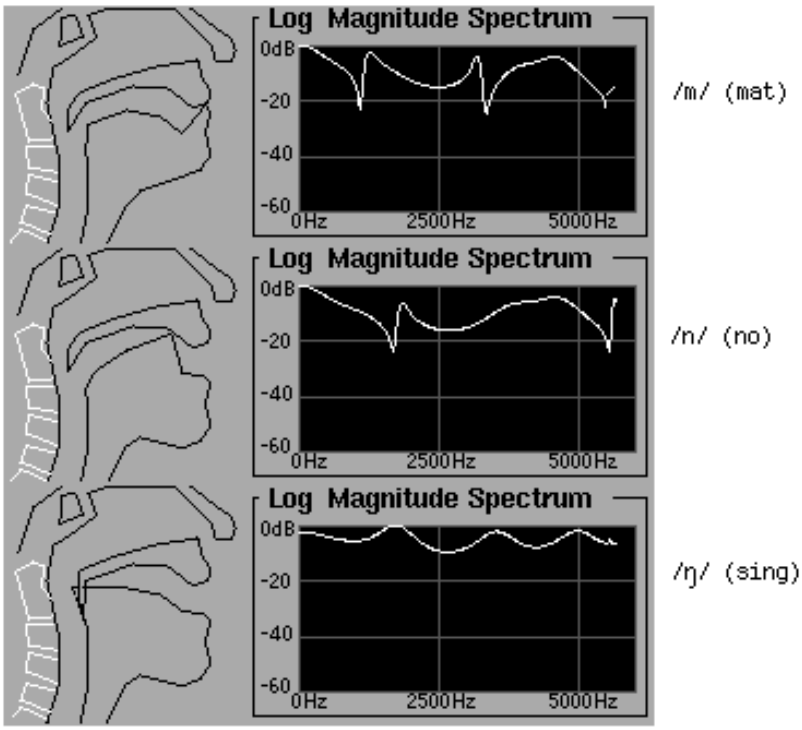

Figure 1.10: Vocal tract shapes and the corresponding log-magnitude transfer functions for three nasal vowels.

1.5 The Periodic Glottal Source

1.5.1 Synthesis Models of the Glottal Waveform

Impulse and Multi-Pulse LPC

Linear Predictive Coding (LPC) was discussed in Section 1.3.1. In such systems, the output signal is modeled as a linear combination of a number of previous samples, and a linear predictor is designed. Since any component of the signal which is linearly predictable is modeled by the filter, the predictor residual exhibits a white (flat) spectrum. The filter source is modeled as having a flat spectrum, and two sources which satisfy this spectral description are impulses and white noise. The predictor filter contains the desired spectral properties of the output wave. Atal, Chang, Mathews, and Tukey [66] showed that there

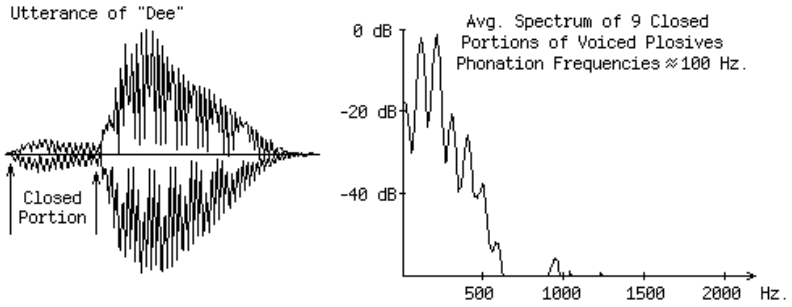

Figure 1.11: One utterance containing a voiced plosive, and the average spectrum of 9 closed voiced portions of the utterances "Bee, Dee, Gee, Boo, Doo, Goo, Baa, Daa, Gaa".

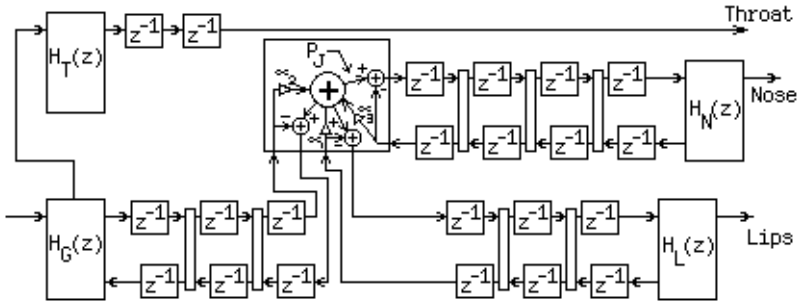

Figure 1.12: A digital realization of the vocal tract, showing waveguides for the oral and nasal passages, the three-way scattering junction at the velum, and the low-pass filter and delay line which model radiation through the throat wall.

are many combinations of sources and filters which exhibit the same output spectrum. A source/filter model using a white excitation source simplifies the filter solution. The model of the source as either a periodic impulse train or white noise fits well within a model of speech in which the signal is considered to be either voiced (periodic) or unvoiced (noise-like) within a certain analysis frame.

A common complaint about the quality of LPC speech/singing systems is that the synthesis sounds 'buzzy'. The buzziness is especially evident in the time varying case, and results from the impulsive quality of the excitation function and the rapid switches from voiced to unvoiced excitation modes. Many sounds produced by the vocal mechanism are both

pitched and noisy, such as the voiced fricative /z/ (as in zoo). As shown in Section 1.4.3, the acoustic tube filter model can be used to implement the all-pole filters yielded by LPC analysis. The "vocal tract shapes" directly corresponding to LPC analysis, however, are somewhat unnatural, specifically because of the white-spectrum nature of the LPC source and the fact that the filter is responsible for all spectral coloration. Another problem with the synthesis of speech/singing by exciting an all-pole filter with parametric impulses or noise is that the determination of whether the sound is unvoiced or voiced must be made, and in the latter case the frequency of phonation (F_0) must be determined.

In the human vocal system the source is a pulsatile signal generated by the opening and closing of the vocal folds. The folds open quite slowly as pushed open by the subglottal pressure, and are rapidly "sucked" closed by the Bernoulli effect [87] resulting from air flow. This generates a quasi-periodic voice source with a spectrum which rolls off roughly exponentially with frequency. The filter, which is controlled by the shape of the vocal tract, does not contain all of the spectral information of the final output signal, but rather the spectral features are distributed between the source and the filter.

To improve the representation of the source, LPC with multi-pulse excitation has been investigated [43][61]. In multipulse LPC, a few non-zero samples represent the residual signal of Equation 1.6. Typically the number of residual samples is reduced by a factor of 8 to 10. In normal voiced speech, this yields a source signal which exhibits several pulses around the glottal closure epoch, and a few pulses elsewhere in the period. Multipulse LPC automatically models the voiced/unvoiced/mixed nature of actual speech, and also models the noise present in voiced speech.

LPC and its variants are principally analysis/resynthesis (speech coding) systems, and as such must yield unique solutions in applications involving automatic speech coding and resynthesis. For research in the synthesis of singing, or the synthesis of singing for compositional purposes, the model parameters should not necessarily depend on analysis of sung tones or performances. In such a system, synthesis by direct control of the model parameters is desired, and constraints of mathematical uniqueness are less important. In Sections 1.4.6 through 1.4, digital models of the vocal tract filters were developed from the principles of physics. The next sections discuss and develop physically motivated models of periodic excitation sources and noise sources in the vocal mechanism.

Time Domain Wave-Shape of Parametric Glottal Pulses

Rosenberg [79] investigated the perceptual aspects of the replacement of the glottal waveform with an abstract (parametric) glottal pulse. He found that simple polynomial or trigonometric functions can be used to build glottal excitation functions without significant perceptual degradation of the speech signal. In investigating the parameters of glottal opening and closing times, Rosenberg discovered three things of importance:

- There is a large tolerance for different combinations of relative opening and closing times.
- Very small opening or closing times are not favored perceptually.
- Opening times which are equal to or less than the closing time are not ranked favorably.

These experiments suggest minimum requirements for parametric glottal pulse control; a fixed opening shape is allowed, but control over closing time is necessary.

Sundberg and Gauffin [83] studied flow glottograms (the waveform of the airflow through the glottal folds), and determined that there are two acoustically important features of the glottogram. Those two features are the amplitude and the closing rate. They mapped these two features onto more subjective descriptions of vocal quality. They denote vocal quality as ranging from *pressed*, in which efficient use of the air through the vocal folds is realized by a high tension between the folds themselves, through *normal*, *flow*, *breathy*, to *whisper*, in which the tension between the folds is so low as to allow the passage of some air through the folds without complete interruption. The researchers found that control along the dimension of *breathy* to *pressed* phonation is related to amplitude, and loudness (or vocal effort) is related to the closing rate of the glottis. These simple controls allow great flexibility of control over the intuitive vocal notions of effort, emotion, mode of phonation, musical dynamics, and pitch range.

Work by Cummings and Clements [68] studied glottal waveshape under differing voice stress conditions. The researchers chose six parameters to describe the glottal waveform; opening slope, opening duration, top duration, closing slope, closing duration and closed duration. Eleven speech conditions were investigated; *angry*, *50 %*, *clear*, *70 %*, *fast*, *loud*, *Lombard*,

normal, *question*, *slow*, and *soft*. The results showed a sufficiently different profile based on the six parameters to allow identification of each of the ten speech styles. Except for the *questioning* speech condition, the closed duration was roughly constant. Closing slope and duration exhibited wide variation, and opening slope only deviated significantly in the *loud* speech case.

Such work suggests a parametrically controlled time domain description of glottal waveshape as a source for the source/filter model of the vocal mechanism. The glottal pulse can be pre-synthesized from the parameters and stored in a wavetable. If multiple wavetables are stored, interpolation between wavetables simulates variations in the source. Wavetables are discussed in Section 1.5.1. A simple parametric glottal pulse consists of a raised cosine until the specified closing edge start point, then a line segment from the cosine curve down to zero at the closing edge end point, then zero for the remainder of the period. The control of closing edge beginning and ending points provides the minimal parametric glottal pulse, with fixed opening slope and time, and controls affecting closing slope and time. If glottal closure beginning and ending points (e_1 and e_2, respectively) are specified as a fraction of the period of the raised cosine, the form for the frequency-normalized continuous-time parametric glottal pulse is:

$$x(t) = \begin{cases} 0.5 - 0.5cos(2\pi t) & t < e_1 \\ -\frac{0.5-0.5cos(2\pi e_1)}{(e_2-e_1)}(t-e_2) & e_1 \leq t \leq e_2 \\ 0.0 & t > e_2 \end{cases} \tag{1.52}$$

Where

$$0.0 \leq e_1 \leq e_2 \leq 1.0 \tag{1.53}$$

To control the bandwidth of the pulse to prevent aliasing, to compress the representation (a few numbers representing magnitudes and phases vs. the sampled data representation of the wavetable), and to provide some spectral parameterization for further processing, the specified pulse can be converted into Fourier series coefficients. A continuous time periodic function can be decomposed into a sum of sinusoids [146],

$$x(t) = C_0 + \sum_{n=1}^{\infty} A_n cos(2\pi F_0 nt) + B_n sin(2\pi F_0 nt) \tag{1.54}$$

Where F_0 is the fundamental frequency, which is the inverse of the fundamental period T_0. The Fourier series coefficients are computed over one period of the periodic wave, and are defined by:

$$C_0 \equiv \frac{1}{T_0}\int_{t=0}^{T_0} x(t)dt \tag{1.55}$$

$$A_n \equiv \frac{2}{T_0}\int_{t=0}^{T_0} x(t)cos(2\pi F_0 nt)dt \tag{1.56}$$

$$B_n \equiv \frac{2}{T_0}\int_{t=0}^{T_0} x(t)sin(2\pi F_0 nt)dt \tag{1.57}$$

In the case of the parametric glottal pulse, the integrals are divided into the cosine portion and the line segment portion. The normalization of the glottal wave to a period (and thus frequency) of 1 simplifies Equations 1.54 through 1.57. If $e_1 = e_2$, corresponding to an instantaneous glottal closure, only the coefficients corresponding to the cosine portion are computed.

$$C_0 = \int_{t=0}^{e_1} x(t)dt \tag{1.58}$$

$$A_n = 2\int_{t=0}^{e_1} x(t)cos(2\pi nt)dt \tag{1.59}$$

$$B_n = 2\int_{t=0}^{e_1} x(t)sin(2\pi nt)dt \tag{1.60}$$

From linearity of the Fourier Series, if the sloping line segment portion is needed ($e1 \neq e2$), the Fourier coefficients can be computed for the line segment alone, and the resultant coefficients added to the raised cosine coefficients obtained in Equations 1.58 through 1.60 to arrive at the total coefficient value.

$$C_0^{closure} = \int_{t=e_1}^{e_2} x(t)dt \tag{1.61}$$

$$A_n^{closure} = 2\int_{t=e_1}^{e_2} x(t)cos(2\pi nt)dt \tag{1.62}$$

$$B_n^{closure} = 2\int_{t=e_1}^{e_2} x(t)cos(2\pi nt)dt \tag{1.63}$$

$$C_0 = C_0^{cosine} + C_0^{closure} \tag{1.64}$$

$$A_n = A_n^{cosine} + A_n^{closure} \tag{1.65}$$

$$B_n = B_n^{cosine} + B_n^{closure} \tag{1.66}$$

The final closed form for computing the Fourier coefficients for the simple parametric glottal pulse is:

$$C_0 = \begin{cases} \frac{e_1}{2} - \frac{sin(2\pi e_1)}{4\pi} & e_1 = e_2 \\ \frac{e_1}{2} - \frac{sin(2\pi e_1)}{4\pi} + \frac{(1-cos(2\pi e_1))(e_2-e_1)}{4} & e_1 < e_2 \end{cases} \tag{1.67}$$

$$A_n = \begin{cases} \frac{sin(2\pi e_1)}{2\pi} - \frac{sin(4\pi e_1)}{8\pi} - \frac{e_1}{2} & n = 1, e_1 = e_2 \\ \frac{sin(2\pi e_1)}{2\pi} - \frac{sin(4\pi e_1)}{8\pi} - \frac{e_1}{2} + & \\ \quad \left(\frac{1-cos(2\pi e_1)}{2\pi}\right)\left(\frac{cos(2\pi e_1)-cos(2\pi e_2)}{2\pi(e_2-e_1)} - sin(2\pi e_1)\right) & n = 1, e_1 < e_2 \\ \frac{1}{2\pi}\left(\frac{sin(2\pi n e_1)}{n} - \frac{sin((n-1)2\pi e_1)}{2(n-1)} - \frac{sin((n+1)*2\pi e_1}{2(n+1)}\right) & n > 1, e_1 = e_2 \\ \frac{1}{2\pi}\left(\frac{sin(2\pi n e_1)}{n} - \frac{sin((n-1)2\pi e_1)}{2(n-1)} - \frac{sin((n+1)*2\pi e_1}{2(n+1)}\right) + & \\ \quad \left(\frac{1-cos(2\pi * a)}{n}\right)\left(\frac{cos(2\pi n e_1)-cos(2\pi n e_2)}{2\pi n(e_2-e_1)} - sin(2\pi n e_1)\right) & n > 1, e_1 < e_2 \end{cases} \tag{1.68}$$

$$B_n = \begin{cases} \frac{3+cos(4\pi e_1)}{8\pi} - \frac{cos(2\pi e_1)}{2\pi} & n = 1, e_1 = e_2 \\ \frac{3+cos(4\pi e_1)}{8\pi} - \frac{cos(2\pi e_1)}{2\pi} + cos(2\pi e_1) + & \\ \quad \left(\frac{1-cos(2\pi e_1)}{2\pi}\right)\left(\frac{sin(2\pi e_1)-sin(2\pi e_2)}{2\pi(e_2-e_1)}\right) & n = 1, e_1 < e_2 \\ \frac{1}{2\pi}\left(\frac{1-cos(2\pi n e_1)}{n} + \frac{cos((n-1)2\pi e_1)-1}{2(n-1)} + \frac{cos((n+1)2\pi e_1)-1}{2(n+1)}\right) & n > 1, e_1 = e_2 \\ \frac{1}{2\pi}\left(\frac{1-cos(2\pi n e_1)}{n} + \frac{cos((n-1)2\pi e_1)-1}{2(n-1)} + \frac{cos((n+1)2\pi * e_1)-1)}{2(n+1)}\right) + & \\ \quad \left(\frac{1-cos(2\pi e_1)}{n}\right)\left(\frac{sin(2\pi n e_1)-sin(2\pi n e_2)}{2\pi n(e_2-e_1)} + cos(2\pi n e_1)\right) & n > 1, e_1 < e_2 \end{cases} \tag{1.69}$$

Once the Fourier coefficients are computed, the waveform of a single cycle may be synthesized digitally by sampling the Fourier Series formula of Equation 1.54 at the appropriate sampling rate. Other features of the parametric glottal pulse can be added to the Fourier series representation, yielding closed form relations between the time-domain parameters and the spectrum of the resultant glottal pulse.

With a representation of the glottal time domain wave as a set of frequency domain sinusoidal components, modification of the glottal source in the frequency domain is simple. Bloothooft and Plomp [4] applied multidimensional scaling to singer voices to determine sets of spectral basis vectors (power spectra in $\frac{1}{3}$ octave bands) which described well the spectral differences between singers, the change in spectrum with pitch, and the spectral change with intensity. The spectral regions where these basis vectors are large are the

principal regions modified by the singer, and the regions which differ between individual singers. Manual or rule-based controls for modifying the parametric Fourier coefficients in specific regions could be developed and applied for natural synthesis-by-rule.

Selection of Wavetable Size and Number of Harmonics for Synthesis

For reasons of economy, wavetables are often employed for synthesis of periodic waveforms [155]. To minimize quantization effects, the wavetable is synthesized using the entire dynamic range available, and the gain control is applied multiplicatively to the output of the wave table during resynthesis. If one period of the wave is stored in the wavetable, the wavetable length is N, the increment step through the wavetable is δ (a floating point number), the desired fundamental frequency is F_0, and the sampling frequency is F_S, the increment is given by:

$$\delta = \frac{NF_0}{F_S} \tag{1.70}$$

yielding an output wave $x(n)$ whose nth sample is the element from the table whose location is:

$$n\delta - mN \tag{1.71}$$

where m is the greatest integer yielding a non-negative result in Equation 1.71.

The selection of the wavetable size is driven by memory and distortion considerations. A wavetable that is too small contains a coarse sampling of the waveform, and results in aliasing and quantization errors. Aliasing occurs if the highest frequency harmonic is not sampled at a rate which is above the Nyquist frequency (at least twice the frequency of the harmonic). This is determined by wavetable length, sampling frequency, and playback frequency. If one period of the waveform is stored in the wavetable, the aliasing constraint yields a minimum wavetable length given a desired maximum number of harmonics:

$$N > 2 * \text{ Maximum Number of Harmonics} \tag{1.72}$$

The maximum frequency and sampling frequency determines the maximum number of harmonics:

$$\text{Maximum Number of Harmonics} < \frac{F_s}{2F_{0\ \text{Maximum}}} \tag{1.73}$$

These guidelines assume ideal interpolation [159][162] of fractional-time samples from the wavetable. Interpolation is sometimes not done in wavetable lookup. This corresponds to the zero-order (rectangular interpolation function) condition of digital signal reconstruction, which implies that the replicated spectrum of the wavetable signal is multiplied by a sinc function in the frequency domain, yielding a peak error of -13 dB in the spectrum of the reconstructed signal for a wavetable which satisfies Equation 1.72 with equality. If linear interpolation is performed, the spectrum of the non-interpolated signal is multiplied by the transform of a triangular window, yielding a spectrum with a maximum error of -26 dB. Higher order interpolation schemes, corresponding to higher order interpolation windows, yield smaller errors. By oversampling the wavetable, the error is decreased 6 dB for each doubling of the wavetable size. Thus 96 dB of signal to noise ratio is achieved with linear interpolation of a single sinusoid stored in a table of length 4096, or no interpolation of a sine stored in a table of length 65,536. Figure 1.13 shows the spectra of low frequency synthesized waveforms using non-interpolated sinusoid tables of length four and eight, and a linearly interpolated table of length four.

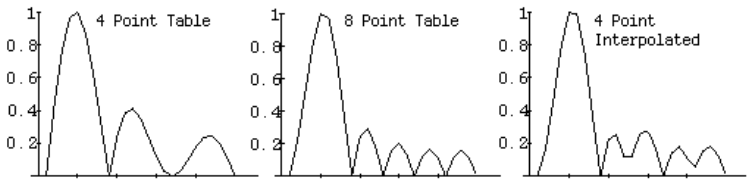

Figure 1.13: Power spectra of low frequency synthesized waveforms using non-interpolated single sinusoid tables of length four and eight, and a linearly interpolated table of length four. The main lobe is the desired sinusoidal component, and the other lobes are distortion components.

1.5.2 Physical Models of the Vocal Folds

One model of the output of the vocal folds involves placing a source of constant pressure or velocity at the input to the vocal tract model, and modulating the opening at the glottal end of the vocal tract tube [57]. The glottal aperture is opened and closed according to a specific waveshape, generated by using the wavetable techniques discussed in the previous section. This model is slightly more physical than simple waveform synthesis, and yields some of the source/filter interactions observed in the vocal mechanism.

A highly physical model of the vocal folds can be constructed using masses and springs. Such models have been proposed by Ishizaka and Flanagan [73], Titze [84][85], Kacprowski [74], and others. The parameters of mass values, spring constants, and breath pressure are used to control the model. The differential equations governing such systems are solved in discrete time to yield a glottal waveform. The waveguide ladder filter realization of the vocal tract, because of the phase delay characteristics and bidirectional wave propagation form, allows coupling of a mass-spring oscillator in a physically meaningful way. The effects of the Bernoulli force and other physical phenomena can be included into the model for a highly accurate simulation. There are two practical problems with using this type of model for sound synthesis, however.

The first problem of a physically based mass-spring model is that of parameterization and control. Once the model has been constructed, the parameters are values of spring constants, masses, glottal breath pressure, and air flow. While many people are mildly familiar with control of the glottis via breath pressure and flow, they certainly are not consciously familiar with control of the vocal source via the individual masses and springs. Singers, composers, students of speech, and other potential users of articulatory vocal simulation systems view glottal control more in the domains of effort, emotion, and the resultant spectrum. Much of fine instantaneous glottal control is completely involuntary [25]. This implies that much of the control stream of parameter changes going to the model must be generated by yet another model, possibly involving some abstract control matrix to map a set of m parameters onto n controls [144], or a physiological model involving more muscles (masses and springs) and some model of neurological feedback at two levels [25][166][181]. Once this model was in place, meaningful parameters could be used to control it, but the model would grow to prohibitive computational size.

The second problem of using a mass-spring coupled oscillator model of the vocal folds is that of computational burden versus sound quality. One of the more successful mass-spring models of the glottal folds is that of Titze [84][85], using 16 mass and spring elements. Another model by Titze and Talkin [86] provides control of the glottal folds via parameters of glottal opening, ligament stress, and vocalis stress. This model provides great insight into the workings of the glottal oscillator, but is computationally complex. The computational burden of using these models is quite large. Further, a number of non-linear behaviors (for example when the vocal folds are closed) must be computed using logical decisions, and such decisions are often the most costly to compute in a Digital Signal Processor (DSP) chip architecture. As discussed above, the control stream to such a model is necessarily quite high, because of the number of discrete components (masses and springs) to control.

Finally, the sound quality of synthesis examples using such models is compelling, but not yet of sufficiently high quality to suggest their use in musical applications, especially when compared to more classical linear techniques of direct synthesis. Titze and Talkin state that their simulations yield results which deviate much more from actual human larynxes than individual human larynxes differ from each other. Time-varying solutions of boundary value problems and construction of variable non-linear mass-spring systems yield insight, and hold much promise for the future of synthesis, but current hardware prohibits the use of all but the simplest of such models for real time synthesis.

1.6 Sources of Noise in the Vocal Tract

Second to glottal fold oscillation, turbulence is the next most important source of sound in the vocal tract. The passage of air at sufficient velocity through an aperture causes turbulent streaming, and thus noise is generated [136][113][114]. The turbulence ceases if the aperture opens sufficiently or the flow decreases. The possibility of turbulent flow is indicated by the value of the Reynolds number, which is a unitless quantity expressing viscous force within the fluid. The Reynolds number is computed from the dimensions of the aperture and the magnitude of the flow by:

$$Re = \frac{Vd}{\nu} \tag{1.74}$$

where V is the particle velocity and d is the effective diameter of the aperture. The kinematic viscosity of the fluid, ν, is defined as the ratio of the dynamic viscosity to the density, and is about 0.15 cm^2/s for dry air [7]. The Reynolds number can be computed in terms of A, the area of the aperture, and U, the volumetric flow, by using the following relationships for flow through a circular aperture:

$$A = \frac{\pi d^2}{4} \tag{1.75}$$

$$U = VA \tag{1.76}$$

yielding:

$$Re = \frac{2U}{\nu\sqrt{A\pi}} \tag{1.77}$$

Turbulent streaming is likely if the Reynolds number is greater than a critical quantity, Re_{crit}, which is about 1,000 for a rectangular slit, and larger for circular apertures. If turbulence is present, noise is generated with a power which is proportional to V^8. The radiated sound power is related to the volumetric flow by:

$$P \propto (\frac{U}{A})^8 \tag{1.78}$$

The center frequency of the principal peak in the spectrum of the turbulent noise is given by:

$$f = \frac{SV}{d} = \frac{SU\sqrt{\pi}}{2\sqrt{A^3}} \tag{1.79}$$

where S is the Strouhal number, which is 0.15 for the center frequency of noise spectral density.

1.6.1 Fricative Consonants

In the case of fricative consonants, some region of the oropharyngeal tube is constricted, air blowing through the constriction causes a turbulent jet to form, and the jet radiates sound energy. One or two resonant peaks characterize the power spectra of most fricative consonants [122]. Figure 1.14 shows the power spectra of four fricative consonants.

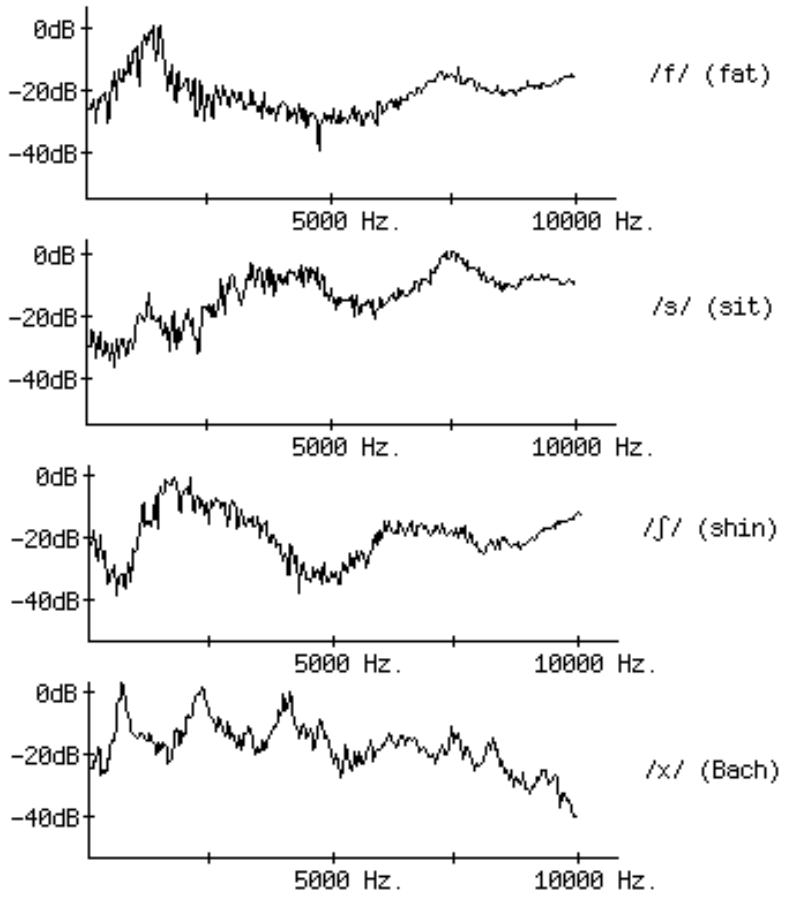

Figure 1.14: Power spectra of four fricative consonants.

The spectral properties of most fricative consonants are due to the acoustical behavior of the turbulent jet [140]. In the case of some fricative consonants, one or more resonances of the acoustic tube are evident in the spectrum, or tube loading affects jet dynamics [112][124]. The regularly spaced peaks in the /x/ (German fricative as in Bach) fricative spectrum of

Figure 1.14 are typical of the odd-harmonic resonances of a pipe which is closed at one end (the constriction point) and open at the other [2]. Some researchers suggest that detailed studies and simulations of flow throughout the vocal tract are necessary for any significant improvements to be made in the quality of voice synthesis [128][137][138]. Other researchers assert that mathematical and computer models of turbulent flow are so primitive at this time that sound synthesis by actual modeling of unsteady flow is impossible [124].

The noise spectrum nature of fricative consonants suggests that white noise, filtered by a low-order resonant filter would comprise a successful synthesis model. This technique is used in LPC, where the detection of an unvoiced (noisy) sound causes the resynthesis filter to be driven with a white noise source. In the waveguide acoustic tube model of the vocal tract, the noise source can be injected into the vocal tract at the correct location. Thus, any spectral properties of the consonant due to linear tube acoustics are modeled naturally by the acoustic tube simulation filter. Any spectral properties due to turbulence can be modeled by an additional low-order resonant filter. Figure 1.15 shows four vocal tract configurations and the spectra of four synthetic fricative consonants. All consonants were synthesized by injecting filtered white noise into the correct location in the acoustic tube digital simulation filter. A four-pole filter was used to model the spectral properties of the jet.

1.6.2 Noise in the Glottis

Flow through the pulsating aperture of the glottis produces noise. This noise is automatically modeled by the residual signal in residual-driven LPC and it is approximated in multipulse LPC systems. In parametric speech and singing synthesis systems, this noise component is often ignored. Section 2.5 explores noise generation in the glottal source both theoretically and experimentally.

1.7 Identification of Filter and Source Control Parameters

Because of the physical nature of the control parameters, many high quality singing voice synthesis examples can be done experimentally using the multiple acoustic tube waveguide filter model of the vocal tract. Configuring the tube into intuitively derived shapes and

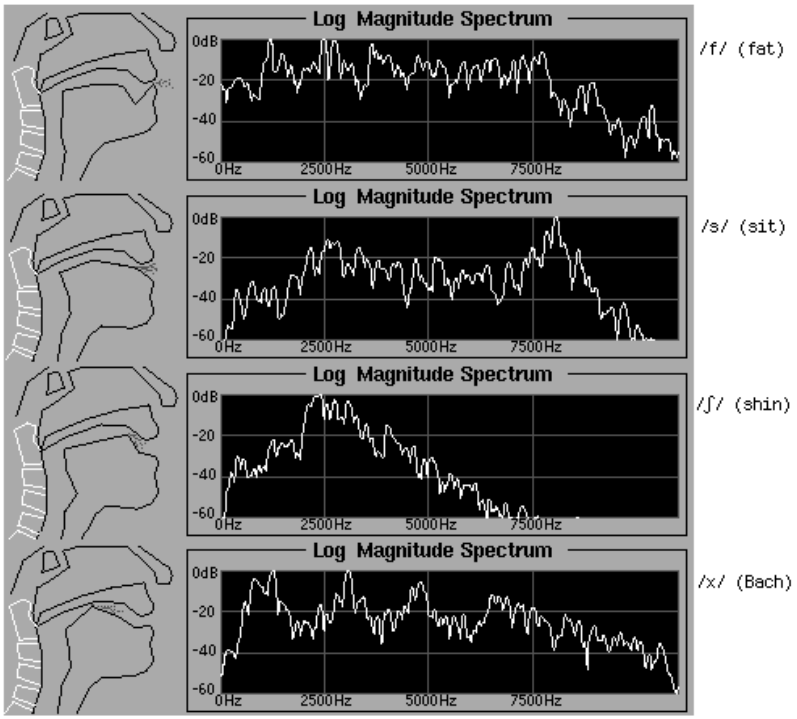

Figure 1.15: Vocal tract configurations and power spectra of four synthetic fricative consonants.

driving the model with the parametric glottal pulse often yields synthesized sounds which are perceptually close to the expected sound. This type of synthesis has been called "analysis by synthesis" and "synthesis by art". In order for synthesis to proceed past the ad hoc experimental stage, however, some methods must be made available for identifying the parameters of vocal tract shape, velum opening, glottal waveform (and thus the parameters of glottal wave amplitude and closure rate), voice pitch and pitch deviation, noise power, and noise spectrum. The principle task is to separate the source from the filter, which is non-trivial. As it is difficult to isolate the source from the filter, it is similarly difficult to separate the processes and techniques of filter identification and source identification. The wide variety of sources and filters yielding the same result [66][46] means that either:

1. Assumptions must be made about the source, the filter, or both. In the simple LPC case, the assumption is that the filter is responsible for all of the spectral

properties, and thus the source exhibits a flat spectrum.

2. Some on-line or off-line measurements must be performed on the subject. This can be as invasive as inserting a sensor through a hole in the neck, or as non-invasive as merely asking the subject to phonate in particular ways (such as whispered speech).

In the remaining sections of this chapter, methods of identifying the vocal tract shape and glottal waveform are presented. Identification of the glottal waveform is discussed, followed by a discussion of methods of identifying the vocal tract filter, methods of mapping a filter to a vocal tract shape, and direct methods of identifying the vocal tract shape. Chapter 2 concentrates on the principal focus of the experimental research, specifically the identification of pitch deviation and glottal noise characteristics in singer voices.

1.7.1 Identifying the Vocal Tract Filter

The waveguide acoustic tube model of the vocal tract provides a stable and physical method of synthesizing vocal sounds, so the aquisition of sampled vocal tract measurements for digital simulation is of particular importance. Methods of identifying the vocal tract filter involve the direct identification of the vocal tract shape, or identification of a filter transfer function, then if desired, the conversion of the filter parameters to physical measurements. X-ray techniques for identifying vocal tract shape were used extensively throughout the 1950's and 1960's [50][8]. The desire for less invasive and more accurate methods of shape identification led to the development of methods involving the use of acoustical information.

The direct determination of the vocal tract geometry from formant locations [81][77] has been investigated. Bonder [46] showed, however, that non-uniqueness conditions exist in the n-tube formulation of the single acoustic tube vocal tract. Incorporation of more spectral information than simple formant locations, and the use of a priori information about the human vocal tract yields more accurate solutions, and can solve the uniqueness problem [71]. *Off-line* measurement of the vocal tract geometry is accomplished by measuring the acoustical impedance at the lips by sinusoidal, noise, or impulse injection methods [81][82]. Since the source is injected into the vocal tract, off-line techniques make no assumption about the source in solving for the filter characteristics. If the computed vocal tract transfer functions are later used to solve for source characteristics, however, there is an assumption made that the vocal tract shape is the same while phonating as it is when measuring the

lip transfer function.

The reflection coefficients of an acoustic tube model can be derived from the coefficients of a digital filter realization. Low frequencies in the voice spectrum are often de-emphasized before performing LPC to reduce dynamic range [11], but the de-emphasis also serves to cancel the natural spectral roll off of the glottal source. LPC provides a set of filter coefficients, and the vocal tract area functions can be obtained from the coefficients using recursive procedures such as the Levinson [148], Schur [145], and Durbin [151] recursions. Algorithms such as these are derived from (or related to) scattering formulations from seismology, optics, and acoustics, and are summarized by Yagle [165]. Such recursive algorithms can be viewed physically as *section peeling* algorithms, since the pieces of the acoustic tube are *peeled off* one section at a time. Given a transfer function with M poles:

$$H(Z) = \frac{1}{A(Z)} \tag{1.80}$$

an all-pole step-down recursion for identification of reflection coefficients from a digital filter transfer function can be defined. Define:

$$\tilde{A}(Z) = Z^{-m} A(Z^{-1}) \tag{1.81}$$

$$A_M(Z) = A(Z) \tag{1.82}$$

recursively as m ranges downward from M to 1, the steps performed are:

$$k_m = \tilde{A}_m(\infty) \tag{1.83}$$

$$A_{m-1}(Z) = \frac{A_m(Z) - k_m \tilde{A}_m(Z)}{1 - k_m^2} \tag{1.84}$$

$A(m)$ can be obtained by applying LPC to vocal tract signals, where the order of the LPC filter is the same as the number of sections in the acoustic tube model. Possible vocal tract signals for identification are a high-frequency emphasized speech waveform, whispered speech, a glottal fry source (extremely low frequency pulse mode of phonation), or a transfer function aquired from the lips. Figure 1.16 shows the vocal tract shapes for the vowel /i/ (beet) obtained from the step-down recursion applied to the LPC filter derived from

whispered speech, glottal fry, and normal voiced phonation with 6 dB per octave emphasis applied.

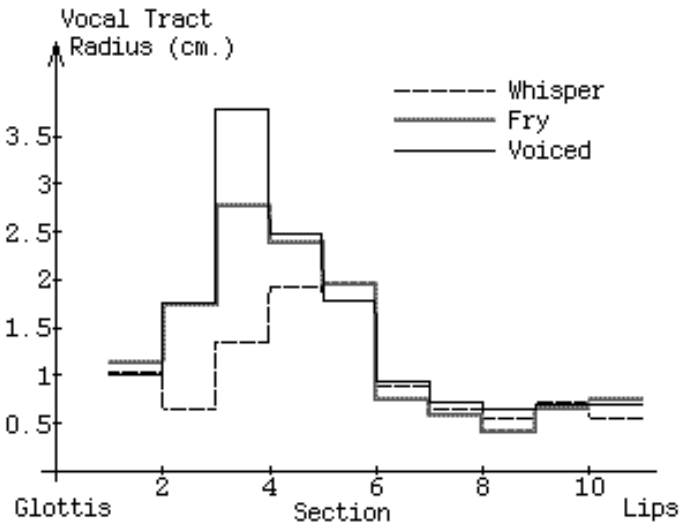

Figure 1.16: Vocal tract shapes for the vowel /i/ (beet). Shapes were obtained by applying step-down recursion to the LPC filter derived from whispered speech, glottal fry, and normal voiced phonation with 6 dB per octave emphasis applied.

1.7.2 Adapting the WGF Vocal Tract Model From the Voice Signal

In this section a new method of speech tracking is presented based on the waveguide filter acoustic tube model of the human vocal tract. This method is similar to that discussed by Fant, Lin, and Badin [51], but is more general and is focused on vocal tract shape identification and speech tracking. A gradient descent method of directly moving the articulators is presented, driven by spectral features of the speech input signal. The least-squares movement which yields the desired formant pattern is imposed on the articulators. The resulting vocal tract shape is compared to a library of shapes, and the shape which is closest is recorded as the vowel estimate. Tracking examples of real speech signals are presented.

The FAST Vowel Tracking Algorithm

A simple WGF acoustic tube model of the oropharyngeal airway of vocal tract is used for the adaptation algorithm. The tube is terminated at the glottal end with a single reflection coefficient, representing the average glottal reflection. At the lip end, an inverting one-zero low-pass filter is used to approximate the reflection characteristics of the open end of a tube. For these experiments, the glottal reflection coefficient was chosen to be 0.9, and the lip filter was chosen to be a simple one-zero filter with a gain of -0.4.

Many features of the input signal could be used to drive the articulators. Formant heights, formant locations, the entire spectral envelope as defined by the harmonic peaks, or a combination of these could be used. The system described here uses only the formant frequency locations to adjust the parameters of the vocal tract, and is thus called the Formant driven Articulatory Speech Tracker (FAST). Formant peaks are detected in the speech signal, then the vocal tract model is adjusted to yield matching formant locations. The radii of the vocal tract model are adjusted in a least-squares fashion, following the 'laziest' trajectory to the desired shape. In this study, the model was made up of nine sections, corresponding to the length of a small female vocal tract (14 cm.) at 22,050 Hz. sampling rate. For tracking of the male voice, the spectrum of the vocal tract model was shifted downward by 15% to more closely match male formant locations. The vector of radii specifying the vocal tract shape is of length q, and will be called C:

$$C = [r_1 \; r_2 \; \cdots \; r_q] \tag{1.85}$$

Detecting Formants in the Input Speech Signal

Given a sampled data signal $x(n)$, the signal is processed in blocks of length N, by windowing, transforming, and computing the magnitude spectrum:

$$x_k(n) = w(n) * x(kM + n), \qquad 0 < n < N \tag{1.86}$$

$$|X_k(m)| = |\mathcal{F}[x_k(n)]| \tag{1.87}$$

where k is the block index, and M is the hop size. $\mathcal{F}$ is a frequency transform operator (Fourier or Hartley as described in Appendix A), and the magnitude operator is defined appropriately for the transform being used [146][147]. The window used in these experiments was the second order Blackman-Harris window [153], defined as:

$$w(n) = \left(0.42 - 0.5 * cos\left(2 * \pi * \frac{n}{N}\right) + 0.08 * cos\left(4 * \pi * \frac{n}{N}\right)\right) \tag{1.88}$$

For each magnitude spectrum, the formants are determined. In this study, the following technique was used:

1. Apply high frequency emphasis of 6 dB per octave to the signal. The emphasis operation approximately cancels the roll-off of the glottal source, and deemphasizes any components near zero frequency.

2. Find the largest peak in the spectrum and record the position as a formant frequency.

3. Multiply the spectrum by a prototype zero spectrum (inverse of prototype pole spectrum) located at the detected formant. This inverse-filters the effect of the formant, and emphasizes other formants for detection. The prototype zero is a 0.95 radius complex conjugate pair at 1/4 sampling rate (for minimal interaction of the roots). The magnitude of the transform of the impulse response of the prototype zero is stored in a table.

4. Repeat 2. and 3. until several formants are detected.

5. Sort the formants in order and remove any duplicate formants, formants which are too low in frequency (less than 200 Hz.), or formants which are too close together (less than 350 Hz. difference).

6. If any detected formant exhibits a level lower in amplitude than -60 dB from the highest spectral point, it is of questionable use. In this study, any such formant was replaced with the closest matching formant location of the current vocal tract.

This procedure yields the length p vector of input signal formant locations, S, defined by:

$$S = [F_1 \; F_2 \; \cdots \; F_p] \tag{1.89}$$

Adapting the Vocal Tract

The magnitude transfer function of the impulse response of the vocal tract model exhibits an all-pole response, and is thus easily searched for local maxima. These peaks are the formants of the vocal tract, and their locations are stored in the p length vector T. The purpose of the adaptation algorithm is to move the vocal tract formants to coincide with the formants detected from the input speech signal. Define the vector of desired formant changes as:

$$\Delta = S - T \tag{1.90}$$

To adapt the vocal tract articulator coefficients, the sensitivity of each tract formant location to each coefficient is measured. Define the gradient matrix, which expresses the sensitivity of each formant location to changes in each of the vocal tract radii, as:

$$\nabla = \begin{bmatrix} \frac{\partial F_1}{\partial r_1} & \frac{\partial F_2}{\partial r_1} & \cdots & \frac{\partial F_p}{\partial r_1} \\ \frac{\partial F_1}{\partial r_2} & \frac{\partial F_2}{\partial r_2} & \cdots & \frac{\partial F_p}{\partial r_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial F_1}{\partial r_q} & \frac{\partial F_2}{\partial r_q} & \cdots & \frac{\partial F_p}{\partial r_q} \end{bmatrix} \tag{1.91}$$

The actual ∇ is measured discretely. This is accomplished by perturbing a coefficient by a small amount, recomputing the magnitude transform of the impulse response of the perturbed vocal tract, and recording the change of each formant position. By perturbing each coefficient individually, the gradient surface of the formant locations is measured. Thus the measured ∇ is made up of row vectors whose components are given by:

$$\nabla_i = \frac{S - \tilde{T}_i}{\Delta r} \tag{1.92}$$

where $\tilde{T}_i$ is the formant position vector of the vocal tract with the ith radius r_i perturbed by Δr.

In a linear system, the Newton's method solution is the least-squares movement of each of the radii yielding the desired formant shifts, given by:

$$C_\Delta = \Delta(\nabla\nabla^t)^{-1}\nabla \tag{1.93}$$

provided that the inverse exists.

If the system is assumed linear in a small region around the current location, the gradient can be measured with a small Δr, and the radii can be moved by a small fraction μ of the computed movement vector:

$$C_{new} = C + \mu C_\Delta \tag{1.94}$$

The steps of adaptation for each block of the input signal are:

1. Identify the input signal formants.
2. Compute the vector T of vocal tract formants.
3. Compute the desired formant delta vector S.
4. Measure the gradient ∇ (Equation 1.91).
5. Form the radius movement vector C_Δ (Equation 1.7.2).
6. Move the radii by some fraction μ of the movement vector.
7. Repeat steps 2 through 6 until the vector $\Delta = T - S$ is sufficiently small or cannot be improved.

This algorithm takes advantage of two properties of speech and the speech production mechanism, and depends on one assumption about vocal tract motion.

- The first property is the slowly varying quasi-stationarity of speech signals, which

allows the algorithm to track slowly varying changes in the formant features of the signal.

- The second property is continuity in the vocal tract shape, ensuring that the vocal tract moves from one shape to the next by occupying intermediate shapes in between.
- The single assumption made is that the vocal tract follows the least-squares motion (in this case, least-squares on the radii), when moving from one shape to the next.

These conditions imply that if the vocal tract model is in the correct shape for a given speech sound, and the sound varies, the vocal tract model varies accordingly.

1.7.3 FAST Experiments on Real Speech Signals

A normally phonated /i/ (as in beet) sound at 150 Hz. was presented to the FAST tracker with an initially neutral shape (all radii = 1 cm.). Figure 1.17 compares the tract shape aquired by application of the FAST algorithm to the tract shapes of Figure 1.16, aquired by applying the section-peeling algorithm presented in Section 1.7.1.

To test the FAST algorithm on continuous vowel trajectories, nine vowels and shapes were selected for the reference library. Table 1.1 shows the vowels and their first three formant frequencies.

Vowel Dictionary				
IPA Symbol	Reference Word	F1	F2	F3
/i/	(beet)	250	2290	3010
/I/	(bid)	390	1990	2550
/ε/	(bed)	530	1840	2480
/ae/	(bad)	660	1720	2410
/$\wedge$/	(bud)	520	1190	2390
/a/	(father)	730	1090	2440
/$\supset$/	(bought)	570	850	2410
/U/	(book)	440	1020	2240
/μ/	(boot)	490	1350	1690

Table 1.1: A table of vowels and the corresponding frequencies of the first three formants.

The system was tested on two utterances; one utterance was the simple connected vowel

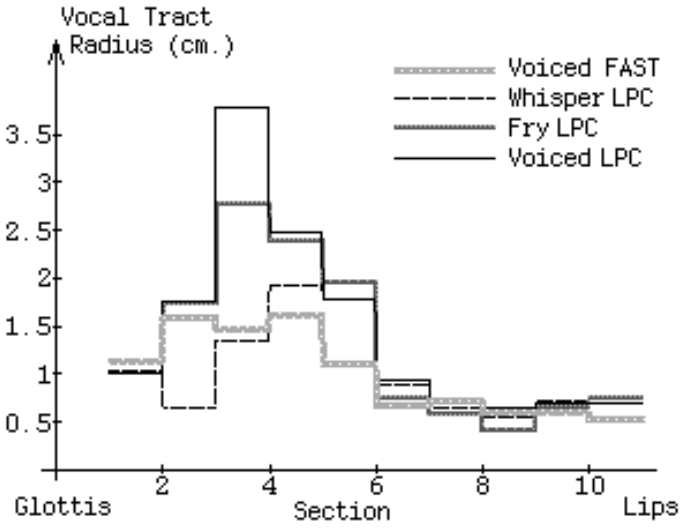

Figure 1.17: Vocal tract shapes for the vowel /i/ (beet) obtained by applying; FAST algorithm to regular phonation, step-down recursion to the LPC filter derived from whispered speech, step-down recursion to the LPC filter derived from glottal fry, and step-down recursion to the LPC filter derived from regular phonation with 6 dB per octave emphasis applied.

trajectory /μ//i//a/ (ooo eee ahh), the other was the rainbow vowel passage "We were away a while ago." Both were uttered by the author with natural inflection over a pitch range of 100 to 150 Hz. Through experimentation, the norm selected for identifying the closest library shape to the adapted vocal tract shape was $L^{0.5}$, where:

$$L^x = \sum_{k=1}^{p} |C_k - \tilde{C}_k|^x \tag{1.95}$$

Where $\tilde{C}_k$ is a library shape for comparison. The higher the number x in Equation 1.95, the more the largest component is accentuated. Selecting a lower norm deemphasizes the outliers, and causes selection of the closest shape based on all coefficients.

Nine trials were run on the /μ//i//a/ utterance, with the FAST tracker initialized to each of the nine library vowels. Figure 1.18 shows the tracking behavior of all trials. To aid in identifying the individual paths, the trajectories in the plot are artifically 'fanned out' to their starting vowel positions at the end of the plot.

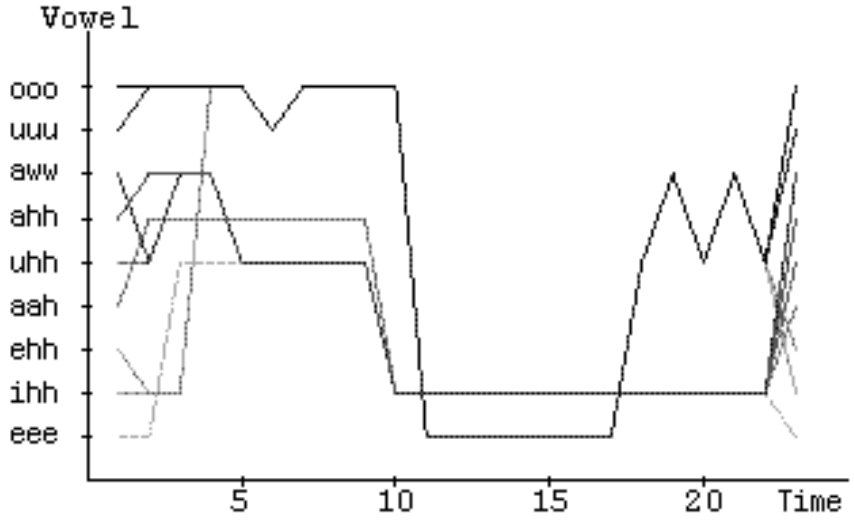

Figure 1.18: Tracking results for the utterance /μ//i//a/ (ooo eee ahh) with nine initial starting shapes.

The initial shapes /I/,/ε/,/U/, and /μ/ tracked correctly, and within four analysis blocks these four trajectories joined to follow the path:

/μ//U//μ//μ//μ//i//i//i//i//i//i//$\wedge$//$\supset$//$\wedge$//$\supset$//$\wedge$/

The other five shapes converged to /I/ and remained at that state through the /μ/ and /a/ phases of the input speech signal. This demonstrates a multi-modal nature to the the vocal tract shape space, consistent with the multiplicity of vocal tract shapes yielding the same acoustical properties [46][66]. Further, it shows that the /i/ and /I/ vowels are not as close in the shape space as in the formant space, and that the ordering of shapes is likely quite different from the formant based order of table 1.1.

Figure 1.19 is a vocal tract shape vs. time display of the trajectory with initial position /μ/. The different shapes of the three vowels are evident in the plot.

The FAST tracker was given the correct initial shape of /μ/ to process the utterance "We were away a while ago", and yielded the trajectory of Figure 1.20. The vowels identified were:

/μ//I//μ//U//μ//I//μ//ae//a//ae//$\wedge$/
"We w ere aw ay aw hile a g o"

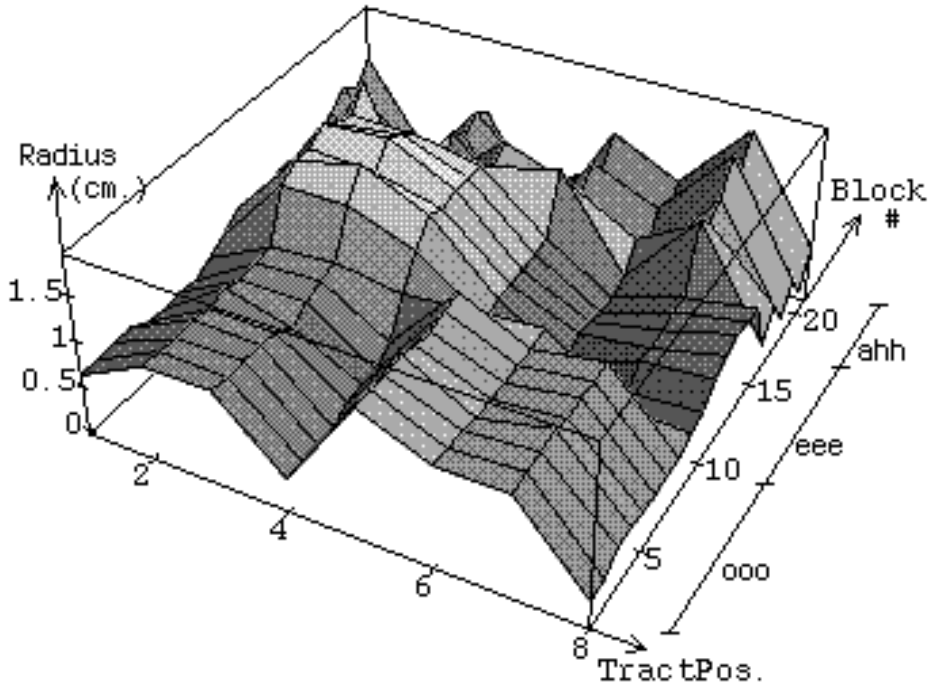

Figure 1.19: Vocal tract shape vs. time display for utterance /μ//i//a/.

An important question to discuss at this time is, "If the formants are detected as part of the FAST tracking algorithm, and if the dictionary of Table 1.1 contains the formant locations, why not simply match the measured formants to the library formants and quit?" One reason is that formants are not the only features which could drive the algorithm. For this study the formant frequency positions alone were selected to drive the articulators. The ear is most sensitive to the formant locations [167][172], but is also capable of detecting the bandwidth and gain of the formants. Other cues might lie in the noise underlying the formant shaped

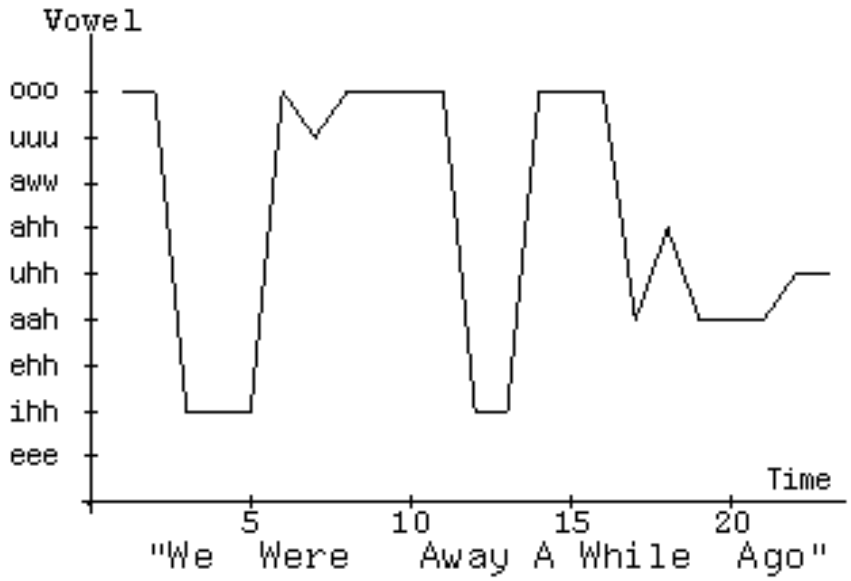

Figure 1.20: Trajectory for Utterance "We were away a while ago".

harmonic peaks. The smallest set of features giving best perceptual recognition is one likely set. Other likely sets might not be restricted to human hearing constraints, but could be entirely performance based. Given the assumption (or identification) of a glottal pulse shape, the FAST algorithm could be modified to the TFAST (Time domain Feature driven Articulatory Speech Tracker) algorithm, where the vocal tract is modified to best match the time domain wave shape of the input speech signal.

Even if formants are the features which drive the tracking algorithm, there are still good reasons to track in the shape space:

- Segmentation of the speech signal is accomplished by observing the current vocal tract shape estimate and comparing to shapes in a library. As the current shape changes continuously, the closest shape in the stored library gives the estimate of the vowel being uttered. The points where the closest library shape changes are indicators of potential boundaries between phonemes. If the shapes in the library are ordered, the continuous slowly-varying nature of the shape description implies that if the shape library is searched outward from the previous shape, the current

shape will be located quickly.

- Improved separation of consonants is realized by using information about the vocal tract shape and trajectory. One frequently occurring problem in both human and machine speech recognition is that consonants within a family (fricatives, plosives, voiced plosives, etc.) sound much the same. The duration of such events is often so short that frequency domain analysis to determine the phoneme is difficult, if not impossible. If the vocal tract shape and trajectory is known, the formation of a constriction could indicate the likelihood of a consonant in the future. If the consonant occurred, the shape would identify the consonant, and the use of spectral processing could increase confidence if needed. Such predictive capabilities are powerful, and not as easily obtained from frequency domain techniques alone.

Investigation of the control and feature detection aspects would improve the basic FAST algorithm. Specifically:

- Investigation of the norms used for vocal tract adaptation and identification. The issues of laziness of movement and closeness of fit of the vocal tract articulatory controls are highly influenced by the norm used to define distance in the articulatory vector space. If the same norm is used to determine shape as is minimized in the adaptation, the closest shape decision is more likely to yield the starting shape. The experiments showed that the norm used for defining closeness of vocal tract shapes clearly influences the ordering, and thus would influence the decisions of any system which tracks vocal tract shapes. The norm which is minimized piecewise in the adaptation process is naturally selected to be least-squares on the radii, but an equally likely candidate would be least-squares on the areas. A somewhat less physical but computationally efficient criterion is least-squares on the reflection coefficients themselves. Norms other than least-squares should also be investigated, as well as other schemes which attach penalties and weightings based on physical constraints of the human vocal tract.

- Vector quantization of vocal tract shapes. This reduces search complexity and memory usage. If the set of vocal tract shape vectors is quantized to some set of discrete vectors, efficient data and complexity reduction is possible. Further, there is a compelling physical motivation for selecting a subset of all possible vectors. Since the

vocal tract shapes that humans can use vary only over a specific range, and over an even smaller range in normal speech, the restriction of shapes to these ranges is natural. Such a physically based subspace selection not only reduces the complexity of the search, but further steers the tracking process by indicating when the adapted shape is nearing a physically impossible region (an indication that the system is lost).

- Library construction. Selection is aided by perceptual and physical guidelines. Related to the vector quantization of shapes is the selection and ordering of shapes which best fit the perceptual boundaries of phonemes and diphones. This provides a further reduction in the complexity of the space to be searched, and thus more efficient look-up and decoding. A library of diphones (transition rules from one state to another), often considered important in speech analysis and synthesis, could be constructed. It is likely that the smoothly tracking nature of the system would negate the need for such a lexicon, however.

- Use of general and specialized hardware. The model used for this study was implemented entirely in software, and thus required significant time to perform the gradient measurement and adaptation. Each articulator coefficient requires an impulse response and log-magnitude transform calculation, dominating the computation time. The vocal tract model is running in a synthesis-only system [49][48] on a Motorola DSP56001 digital signal processing chip, and could be used for the FAST tracking algorithm. If the DSP model were reduced to the components required for FAST, and the log-magnitude frequency transform operations were implemented on the DSP, an improvement in computation time of two to three orders of magnitude is projected. Further computational improvements in the searching and adaptation algorithms would bring the system to real-time capability.

1.7.4 Identifying the Glottal Wave

Once a reliable estimate of the vocal tract filter is obtained, the glottal wave can be estimated by applying a technique known as inverse filtering, or deconvolution. Given the assumption that the voice wave $y(t)$ is generated by the linear filtering of source wave $x(t)$ by convolution with vocal tract impulse response $h(t)$, the frequency domain representation of this linear

operation is:

$$Y(Z) = X(Z)H(Z) \tag{1.96}$$

The source is obtained by:

$$X(Z) = Y(Z)/H(Z) \tag{1.97}$$

provided that H(Z) has no zeroes (roots of the numerator polynomial) outside the unit circle in the Z plane. Equation 1.97 is the defining equation for inverse filtering. In the LPC case,

$$H(Z) = \frac{1}{A(Z)} \tag{1.98}$$

so the inverse filtering operation is guaranteed stable. The process of inverse filtering in actual practice is often part science and part art. Early techniques involved the use of analog filters [72] , and called for initial coarse adjustment of the filters given a priori information about the speech wave, then the filters were adjusted further while watching the output on an oscilloscope until the output wave met some criteria. Typically the goal was thought to have been achieved when the output waveform 'looked' like a glottal waveform. So the process involved the assumption of a glottal pulse shape, then filtering was applied until the result looked enough like the assumed glottal shape. Automatic inverse filtering was investigated by Miller and Mathews [78].

Digital filtering improves the stability and method of solution of the deconvolution problem, and was proposed as a method of identifying the vocal tract filter in the time-varying case [76]. Again, however, the filter was used in a hand guided mode of operation. Wakita defined a recursive method of inverse filtering [88][89], but with the stated goal of identifying vocal tract shape rather than source characteristics. Rothenberg proposed and investigated a method of inverse filtering which uses a special apparatus to measure volume velocity from the lips, rather than acoustic pressure [80]. The inverse filtering problem is simplified if the flow is deduced from pressure gradient measurements performed very near the glottal folds [67].

Somewhat more invasive methods involve the insertion of probes into the vocal tract, sometimes through the throat wall [134], or the use of Electroglottographs (EGG) to measure electrical resistance of the larynx [75]. Such methods are generally not considered suitable for the study of singer voices.

The inverse filtering method used for the singing research and resynthesis examples herein involves using LPC to fit the spectra of multiple signals made by a single singer using a single vocal tract shape. The LPC filter is then factored into complex resonator pairs (formants) by finding the complex roots of the polynomial and grouping them into second order real filter polynomials. The parameters of center frequency and Z plane radius are made available for adjustment by hand, such as increasing or decreasing filter resonance. In a normal mode of use for identifying glottal shape, a vowel is selected and the singer phonates at the specific pitch and volume under investigation. Taking care not to change his/her vocal tract shape, the singer then produces whispered speech. If the singer is able (some are not), they are then asked to phonate in a glottal fry mode (extremely low frequency glottal pulses), again with the same vocal tract shape. The whisper and fry signals can be processed with LPC analysis in multiple blocks and averaged, yielding smoothed estimates of the vocal tract transfer function. If the whisper and fry filter transfer characteristics differ greatly, the process can be repeated. If the transfer functions agree, the estimate is assumed to be reliable and is used to inverse filter the normally phonated sound. The resulting waveform may be inspected for similarity to typical glottal waveforms. Figure 1.21 shows the spectra of the normal phonation mode, whispered speech, and glottal fry of the vowel /i/ (beet), with the LPC transfer functions superimposed. High frequency emphasis of 6 dB per octave was applied to the normal phonation signal only. The numbers to the right are the center frequencies and Z plane radii of the resonances of the corresponding LPC filters.

Figure 1.22 shows the waveform of the vowel /i/, and the waveforms aquired by inverse filtering using the voiced, whispered, and fry LPC filters shown in Figure 1.21.

Noting that the spectra of all inverse filtered glottal waveforms in Figure 1.22 exhibited peaks at the second and fifth harmonics, the fry inverse filter was modified by increasing the resonance of the first formant to 0.985, and a fifth resonance was added at 750 Hz. (the location of the fifth harmonic). Figure 1.23 shows the glottal waveform estimate achieved by application of the hand adjusted filter. By aquiring many glottal waveforms by inverse filtering many vowels, a reliable estimate of the time domain glottal shape is obtained.

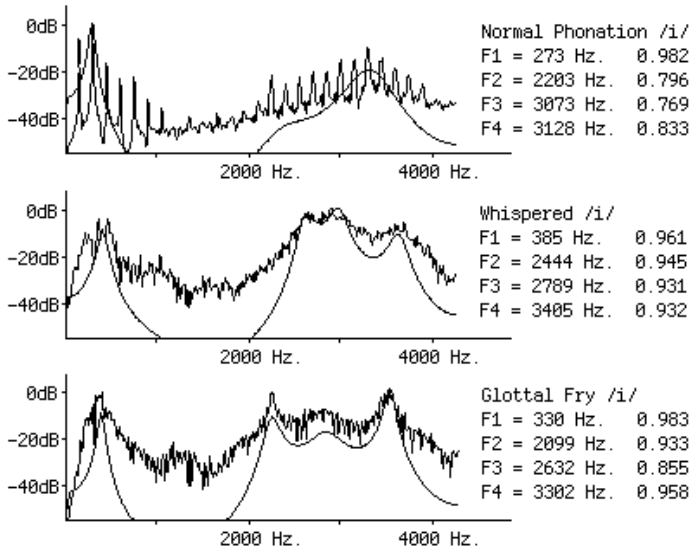

Figure 1.21: Spectra of normal mode phonation (top), whispered speech (center), and glottal fry (bottom) of the vowel /i/ (beet). The smooth curves are the LPC filter spectra, and the numbers to the right are the resonances and Z plane radii of the LPC filters. The normal mode phonation signal was processed with 6 dB per octave high frequency emphasis prior to application of LPC.

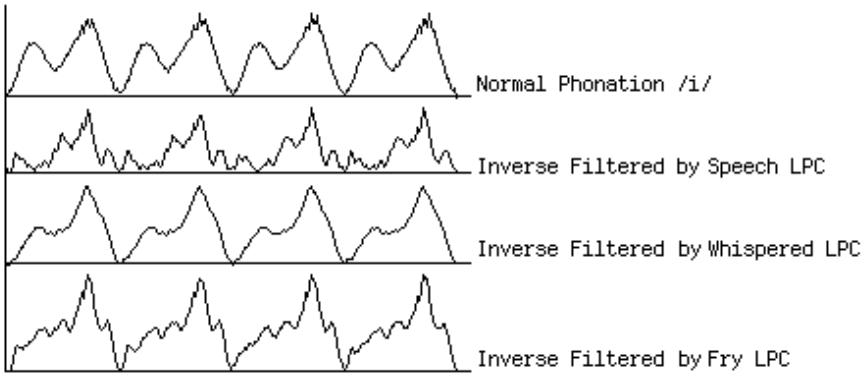

Figure 1.22: From top to bottom: Waveforms of normal mode phonation of the vowel /i/, and normal mode phonation inverse filtered by its own LPC filter, a whispered speech LPC filter, and glottal fry LPC filter.

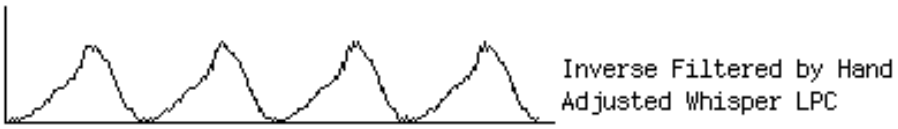

Figure 1.23: Waveform of normal mode phonation of the vowel /i/, after inverse filtering by hand adjusted whispered speech LPC filter.

Chapter 2

Identification of Glottal Source Deviations

Deviations of pitch and timbre in the voice are extremely important perceptual features. Some amount of pitch deviation is present in the voice at all times, no matter how much the speaker/singer endeavors to remove it. This involuntary pitch deviation comes about automatically, as a result of the neurological system controlling the physical elements of the voice source oscillator [94]. There is even a component of modulation of the fundamental pitch of the voice which is caused by the heartbeat [103]. Other components of pitch deviation are controlled consciously, such as pitch inflection, and to a lesser extent, singer vibrato. Synthesis of vowels without pitch deviation sounds mechanical [96][3]. If a singing tone is synthesized without pitch deviation, then deviation is added slowly, the percept suddenly becomes that of a human singer [28][174]. Some deviations in spectrum come about because of the source/filter model of the vocal tract. The spectrum changes shape as the changing source pitch causes the harmonics to move under the relatively fixed spectral envelope of the vocal tract filter. Other spectral changes are due to deviations of the source spectrum, caused by modulation of subglottal pressure and noise in the voice source.

Pitch deviation and noise in the quasi-periodic voice source is discussed and analyzed in this chapter. First, a brief overview of singing voice pitch deviation research is presented, followed by an overview of methods of extracting the pitch signal. A new method of pitch detection is defined which provides accurate smoothed or unsmoothed estimates of the pitch

at any sampling interval. An experimental study of singer pitch deviation using the new pitch detection method is presented. Using a large sample of vocal tones from a number of trained professional singers, rules are formed about the behavior of the deviation signal as a function of pitch and vocal effort. A discussion of spectral and waveshape deviations emphasizes noise generation in the glottal source. A fluid dynamics analysis of the glottal source is performed. Methods for extracting the periodic and residual parts of the voice signal are presented. Studies of noise in singer voices are presented, with conclusions and definitions of rules for synthesis of glottal source additive noise as a function of pitch, voice type, and mode of phonation.

2.1 Pitch Deviation in the Voice Source

For discussion and analysis of pitch deviation, it is important to define pertinent terms. The intentional sinusoidal modulation of the fundamental pitch is called *vibrato*, and occurs at a frequency of 5-7 Hz. in trained western BelCanto singing voices. The commonly believed notion that a singer is capable of controlling the rate of his/her vibrato is a subject of controversy. One compelling study used auditory masking and vibrato side tones, and found that singers were not able to vary vibrato to match a particular rate [166]. There is no debate that singers can be trained to control the amount of vibrato to a large degree, as is demonstrated in barbershop and early-music singing. The perceived pitch of vibrato tones was investigated by Sundberg, and Shonle and Horan [21][179]. These studies found that the perceived pitch was closest to the geometric mean of the extreme frequencies (rather than the linear mean), but the difference between the geometric and linear means at normal modulation values is too small to be of musical significance. The studies also showed that the perceived pitch was relatively independent of the total modulation amount, and modulation rate (when varied between four and eight Hz.).

Modulation components at frequencies higher than the vibrato are called *jitter*. These high frequency components are also often called *flutter*. The production of jitter is generally regarded as an involuntary process, caused by random neural firing and a low level feedback mechanism which, in the singing voice, can be trained to cause the periodic oscillation of vibrato [178].

Modulation components at frequencies lower than the vibrato rate have commonly been

called *wow*. The terms wow and flutter stem from the wow and flutter specifications used to evaluate motor-driven analog audio playback equipment. In the case of these specifications, it is desirable to reduce the wow and flutter to zero. In contrast, some non-zero amount of frequency modulation is necessary to simulate natural vocal sounds. To avoid the negative connotations associated with the terms wow and flutter, the terms *drift* and *jitter* will be used for the remainder of this dissertation. Drift components of very low frequency are directly related to tuning, or the long term average of the fundamental pitch. Drift is generally considered to be consciously controllable by means of an auditory feedback loop [181][168], but it is not possible to completely remove the drift component at will.

Most synthesis models of singer (and instrument) pitch deviation involve a single sinusoid to model the vibrato, mixed with simple low-pass filtered noise to model both the drift and jitter components [111]. Maher and Beauchamp [97] proposed a more elaborate model of vocal pitch control, involving one sinusoidal oscillator, three sources of lowpass filtered noise, various summing elements, and a multiplier. The multiple noise sources isolate the random variations in average frequency, vibrato rate, and vibrato depth. The authors stated, however, that the extra parameterization is of questionable perceptual significance. The vibrato study research conducted herein is to investigate the behavior of the jitter and drift regions of the pitch signal spectrum as a function of sung pitch and intensity, to formulate a set of rules for pitch deviation control, and to suggest a suitable set of control parameters.

2.1.1 A Brief Summary of Pitch Detection Methods

Pitch detection is of interest whenever a single quasi-periodic sound source is to be studied or modeled, specifically in speech and music [95][104]. Pitch detection algorithms can be divided into methods which operate in the time-domain, frequency-domain, or both. One group of pitch detection methods uses the detection and timing of some time-domain feature. Other time-domain methods use autocorrelation functions or difference norms to detect similarity between the waveform and a time lagged version of itself. Another family of methods operates in the frequency-domain, locating sinusoidal peaks in the frequency transform of the input signal. Other methods use combinations of time and frequency-domain techniques to detect pitch.

Frequency-domain methods call for the signal to be frequency-transformed, then the frequency-domain representation is inspected for the first harmonic, the greatest common divisor of all harmonics, or other such indications of the period. Windowing of the signal is recommended to avoid spectral smearing, and depending on the type of window, a minimum number of periods of the signal must be analyzed to enable accurate location of harmonic peaks [153][156]. Various linear pre-processing steps can be be used to make the process of locating frequency-domain features easier, such as performing linear prediction on the signal and using the residual signal for pitch detection. Performing non-linear operations such as peak limiting also simplifies the location of harmonics.

In a time-domain feature detection method the signal is usually pre-processed to accentuate some time-domain feature, then the time between occurrences of that feature is calculated as the period of the signal [91][92][110]. A typical time-domain feature detector is implemented by low pass filtering the signal, then detecting peaks or zero crossings. LPC is often used as a pre-processing step [90]. Since the time between occurrences of a particular feature is used as the period estimate, feature detection schemes usually do not use all of the data available. Selection of a different feature yields a different set of pitch estimates [93]. Since estimates of the period are often defined at the instant when features are detected, the frequency samples yielded are non-uniform in time. To avoid the problem of non-uniform time sampling, a window of fixed size is moved through the signal, and a number of detected periods within each window are averaged to obtain the period estimate. For reliable and smooth estimation, the window must be at least a few periods long. For best accuracy, the signal should be interpolated between samples in order to locate the feature occurrence time as accurately as possible.

Related to the time-domain feature detector is the autocorrelation method. The autocorrelation of the signal is first formed:

$$x \otimes x(m) = \sum_{i=q}^{q+N-1} x(i)x(i+m) \tag{2.1}$$

The main peak in the autocorrelation function is at the zero-lag location ($m = 0$). The location of the next peak gives an estimate of the period, and the height gives an indication of the periodicity of the signal.

All three of the above methods usually require a number of periods of data to form a reliable estimate, and thus some averaging of the frequency signal is unavoidable. The methods often exhibit difficulty in detecting the period of a periodic signal which is missing the fundamental harmonic in the harmonic series. Periodic but pathological signals can be devised to cause nearly any pitch detection algorithm to fail [149].

2.1.2 The Period Predictor Pitch Tracker (PPPT)

The pitch detector/tracker presented here is a refinement of the Average Magnitude Difference Function (AMDF) detectors [108], the earliest of which is that of Miller and Weibel [99]. Methods of this type have also been called comb-filter methods [100]. The AMDF pitch detector forms a function which is the compliment of the autocorrelation function, in that it measures the difference between the waveform and a lagged version of itself. The generalized AMDF function is:

$$AMDF(m) = \sum_{i=q}^{q+N-1} |x(i) - x(i+m)|^k \tag{2.2}$$

The quantity k is set to 1 for average magnitude difference, and other values for other related methods. The zero lag ($m = 0$) position of the AMDF function is identically zero, and the next significant null is a likely estimate of the period. Other nulls will occur at integer multiples of the period. The signal is preprocessed to aid in detection of the first null. The difficulties of using this pitch detection method arise from the issues of finite sampling rate, noise, and signal stationarity. If the signal is truly periodic with period T_0, and T_0 is an integer multiple of the sampling period T_s, then all nulls at integer multiples of T_0 are identically zero. If the period is not an integer multiple of T_s, however, the first null ($m \neq 0$) actually exists between two values of m. A coarse estimate of pitch is tolerable for many speech applications, but is not acceptable for analysis and synthesis of music. Compared to the small computational burden of computing the AMDF, there is no economical method of accurately interpolating between samples to find the true period. This implies that the sampling rate must be sufficiently high to yield the high accuracy required for musical applications. If the signal is quasi-periodic (amplitude modulated, corrupted by noise, etc.), the nulls will never be zero, even if T_0 is an integer multiple of

T_s. The problem of interpolation between lag samples to obtain an accurate pitch estimate is even further complicated in the case of a frequency modulated signal.

A method of pitch detection which uses the phase delay of a periodic predictor to form the pitch estimate is presented in the next sections. This pitch detector accurately tracks a quasi-periodic signal, and will be called the Periodic Predictor Pitch Tracker (PPPT). The PPPT provides a method of automatically and adaptively determining the optimum continuous-time lag, and also provides an estimate of the reliability of the pitch estimate. The PPPT system as initially described is not a complete pitch detector, in that it relies on some other scheme for an initial estimate of the period. Once the detector locks onto the correct period, the method provides accurate estimates of the instantaneous period using all samples of the input signal, provides an estimate of the periodicity of the signal, and provides controls which affect the dynamics and accuracy of the pitch detector.

2.1.3 FIR Filter Methods of Periodic Prediction

Given a quasi-periodic signal $x(n)$, and an integer estimate P of the initial period, periodic prediction is implemented by:

$$\hat{x}(n) = \sum_{i=-M}^{M} x(n - P + i)c(i) \tag{2.3}$$

where M is some appropriately chosen small number and $c(i)$ are the predictor coefficients. Algorithms for obtaining the predictor coefficients are discussed in Section 2.1.4. Backward prediction is implemented by replacing P with $-P$ in Equation 2.3. Figure 2.1 shows a block diagram of the predictor.

The phase (relative to the Pth delayed sample) of the FIR filter implemented by the predictor coefficients is computed by:

$$\theta = \arctan\left(\frac{\sum_{i=-M}^{M} c(i) * \sin(\omega * i)}{\sum_{i=-M}^{M} c(i) * \cos(\omega * i)}\right) \tag{2.4}$$

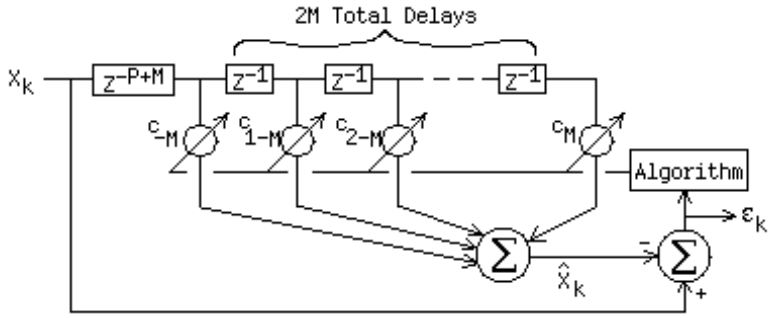

Figure 2.1: Linear FIR period predictor.

The frequency ω is the frequency at which the phase delay of the filter is calculated. The frequencies of interest in calculating the phase delay are the harmonics of the fundamental that it is desired to detect, so some uncertainty is present in the initial calculations. A coarse calculation of ω can be performed by using the value of P, in which case the value of ω is stored and reused as long as the integer value of P does not change. If a more accurate estimate is required, the last predicted period is used to compute ω, or the calculation of ω and the pitch estimate is iterated until a desired accuracy is reached. The relation between the pitch estimate and ω is:

$$\omega = \frac{2\pi}{T_0} \approx \frac{2\pi}{PT_s} \tag{2.5}$$

where T_s is the sampling period in seconds and T_0 is the period estimate. Computation is reduced by exploiting the evenness and oddness of the cosine and sine functions, and the symmetry of the filter definition. Equation 2.4 thus reduces to:

$$\theta = \arctan\left(\frac{\sum_{i=1}^{M}(c(i) - c(-i)) * \sin(\omega * i)}{c(0) + \sum_{i=1}^{M}(c(i) + c(-i)) * \cos(\omega * i)}\right) \tag{2.6}$$

For further computational savings, sine, cosine, and arctan values can be calculated by interpolated table lookup as discussed in Section 1.5.1. The phase delay of the filter is computed by:

$$Phase\ Delay = \frac{\theta}{\omega} \tag{2.7}$$

By adding the computed time delay to the time delay of the P length delay line, the net time delay of the predictor is computed. This total delay is then used to compute a period and frequency estimate:

$$Period = T_0 = \frac{P}{Sampling\ Rate} + Phase\ Delay \tag{2.8}$$

$$Frequency = F_0 = \frac{1}{T_0} \tag{2.9}$$

2.1.4 FIR Period Predictor Implementation Algorithms

The three methods of implementation of the adaptive FIR predictor discussed here are Covariance method Least Squares (CLS), Recursive Least Squares Adaptive (RLS), and Least Mean Squares Adaptive (LMS). All of these methods minimize the Mean Square Error:

$$MSE = \frac{1}{N}\sum_{k=0}^{N-1} \epsilon_k^2 \tag{2.10}$$

where the instantaneous error ϵ_k is defined as the difference between the signal sample and the predicted sample at time k:

$$\epsilon_k = x(k) - \hat{x}(k) \tag{2.11}$$

The Covariance Method

This method of prediction was defined in the LPC descriptions of Section 1.3.1. In this case, the linear prediction is performed P samples ahead. The coefficients $c(i)$ from Equation 2.3 are assumed to implement the least squares P step-ahead predictor over the data set, and thus solve the set of linear equations:

$$Rc = r \tag{2.12}$$

where

$$R = E_i \begin{bmatrix} x(i)x(i) & x(i)x(i+1) & \cdots & x(i)x(i+2M) \\ x(i+1)x(i) & x(i+1)x(i+1) & \cdots & x(i+1)x(i+2M) \\ x(i+2)x(i) & x(i+2)x(i+1) & \cdots & x(i+2)x(i+2M) \\ \vdots & \vdots & \ddots & \vdots \\ x(i+2M)x(i) & x(i+2M)x(i+1) & \cdots & x(i+2M)x(i+2M) \end{bmatrix} \tag{2.13}$$

is the covariance matrix, and

$$r = E_i \begin{bmatrix} x(i)x(i+P) \\ x(i)x(i+1+P) \\ x(i)x(i+2+P) \\ \vdots \\ x(i)x(i+2M+P) \end{bmatrix} \tag{2.14}$$

the P-delayed covariance vector. If the matrix R is invertible, the coefficient vector c is given by:

$$c = R^{-1}r \tag{2.15}$$

The signal can be processed in blocks of length N, yielding a set of predictor coefficients, and thus a period estimate, for each block. Each sample, calculating the coefficients requires $2NM$ multiplies to compute the covariance matrix and vector, and $O(M^3)$ operations for the matrix inversion. Recall that M is small, however, so the autocorrelation operation usually dominates computation.

Recursive Least Squares (RLS)

The Recursive Least Squares algorithm implements a recursive approximate Newton's method solution of Equation 2.15 [143][154]. The matrix R is approximated by an exponentially

decaying sample mean:

$$\hat{R}_k = (1-\alpha)\hat{R}_{k-1} + \alpha X_k X_k^T \tag{2.16}$$

The matrix at this point can be inverted directly. To reduce computation, however, the matrix inverse can be updated recursively using the matrix inversion lemma:

$$(A+BC)^{-1} = A^{-1} - A^{-1}B(A^{-1}B + C^{-1})A^{-1} \tag{2.17}$$

so:

$$\hat{R}_k^{-1} = \frac{\hat{R}_{k-1}^{-1}}{\lambda} - \frac{\frac{\alpha}{\lambda}\hat{R}_{k-1}^{-1}X_k X_k^T \hat{R}_{k-1}^{-1}}{\lambda + \alpha X_k^T \hat{R}_{k-1}^{-1} X_k} \tag{2.18}$$

where $\lambda = 1-\alpha$. With this estimate of the inverse of the autocovariance matrix, the predictor coefficients are updated adaptively:

$$c_{k+1} = c_k + \alpha \hat{R}_k^{-1} X_k \epsilon_k \tag{2.19}$$

The operations required for RLS are $O(M^2)$, but since M is typically low for the periodic predictor, other computational factors usually dominate. To avoid numerical problems, methods of recursively updating $\hat{R}^{-1}$ involving matrix factorization are often used [150].

Least Mean Squares (LMS)

The Least Mean Squares adaptive [163][164] algorithm is a gradient steepest descent algorithm using the instantaneous error to estimate the gradient of the error surface. Each coefficient is adjusted each sample by an amount proportional to the instantaneous error and the signal value which is associated with the coefficient being adjusted. This corresponds to setting the $\hat{R}_k^{-1}$ matrix equal to the identity matrix in the RLS coefficient update Equation 2.19, and yields the LMS update equation:

$$c_{k+1} = c_k + 2\mu X_k \epsilon_k \tag{2.20}$$

The adaptation constant 2μ replaces α by convention from the Newton's method derivation (the 2 comes from taking the multi-dimensional derivative of the error function), and controls the dynamics (and stability) of adaptation. Stability is ensured if:

$$\mu < \left((2M+1)\overline{x^2}\right)^{-1} \tag{2.21}$$

where $\overline{x^2}$ is the signal power. The adaptation parameter μ can be adapted dynamically, yielding the Normalized LMS algorithm:

$$c_{k+1} = c_k + \frac{\alpha}{(2M+1)\overline{x^2}} X_k \epsilon_k \tag{2.22}$$

where the signal power is computed over the last 2M+1 (or greater) samples. The parameter α is any positive number less than 1.

Adapting the Delay Parameter P

The integer period estimate P is variable, and there are new issues of filter dynamics in the LMS and RLS systems caused by on-line adaptation of the delay-line length. Ideally, the filter should experience no transients because of the adaptive modification of P. A proposed method for the on-line adaptation of P follows.

In the proposed algorithm, three LMS/RLS predictors are implemented, one with the current delay P and one each with delays $P-1$ and $P+1$. This provides optimum filter coefficients to be used in the P predictor when the delay line length is modified. If the value of the FIR filter delay is computed to be greater than 0.5 samples, a sample is added to the delay line (P is increased by 1). The coefficients of the P predictor are copied to the $P-1$ predictor. The coefficients of the $P+1$ predictor are copied to the P predictor. Thus the coefficients placed in the $P-1$ and P predictors are optimal for continued prediction. The coefficients of the $P+1$ predictor are time-reversed:

$$c(i) \leftrightarrow c(-i), \quad |i| \leq M \tag{2.23}$$

This coefficient swap operation takes advantage of the symmetry of the filter formulation

and the resulting phase delay expression of Equation 2.7. By flipping the coefficients of the filter, the phase delay is negated. A phase delay of 0.5 samples becomes -0.5, and the $P+1+0.5$ net delay of the $P+1$ delay predictor becomes $P+2-0.5$ when P is increased by 1, yielding the same net delay. Thus the coefficients of the $P+1$ predictor exhibit the correct delay (but are not guaranteed optimal for minimum MSE prediction). A complementary set of operations is performed to shrink the delay line:

- copy the P coefficients to the $P+1$ predictor
- copy the $P-1$ coefficients to the P predictor
- time-reverse the $P-1$ predictor coefficients

After the delay line length has been modified, no further changes to the delay line length are allowed until a number of samples have been processed. This avoids rapid oscillation of delay line length when the period is near a 0.5 sample boundary, and allows the time-reversed coefficients to adapt to optimal values.

A more economical scheme involves running only one predictor, and simply swapping the filter coefficients according to Equation 2.23. This method also requires a settling time during which no changes are made to the delay line length after time-reversal of the filter coefficients.

2.1.5 Adaptive Sampling Rate and Delay Method

One other method of implementing a periodic predictor is similar to the pitch detector of Ney [102], which uses dynamic time-warping of the sampled data signal. The method presented here involves sinc interpolation [162][159] of the sampled data signal $x(n) = x(nT)$ to yield a continuous time signal $x(t)$. The value of a period P (no longer restricted to be an integer) is then selected which minimizes the MSE of Equation 2.10, where the error in this case is defined as:

$$\epsilon_k = x(kT) - x(kT + P) \tag{2.24}$$

This method essentially implements the continuous-time (arbitrary sampling rate) average squared-magnitude difference function pitch detector. Sinc interpolated resampling to

minimize a vector difference norm is discussed in more detail in Section 2.6.3.

2.1.6 Demonstration of Performance of the LMS PPPT

The Normalized LMS PPPT ($\alpha = 0.01$) was selected for testing because of its computational efficiency. The number of coefficients was selected to be 5 ($M = 2$). The pitch estimate was averaged and written out each 50 samples. The sampling rate used for this example was 22.05 kHz, which corresponds to a pitch sampling rate of 441 Hz. Figure 2.2 shows the test signal, which is a 500 Hz. sine wave synthesized with additive white noise at -30 dB and 10% sinusoidal vibrato at 10 Hz.

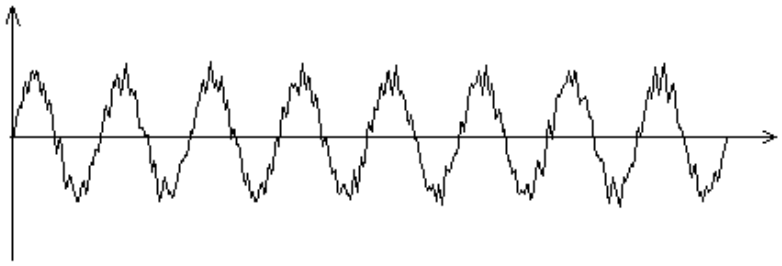

Figure 2.2: Test signal for pitch detection: noisy sinusoid with sinusoidal vibrato

Figure 2.3 shows the time-domain pitch signal extracted by the NLMS PPPT. Below the time-domain pitch signal display is the Log-Magnitude spectrum of the pitch signal (with the mean removed). The smoothing window of 50 samples is about one period of the input signal, meaning that the pitch estimates are obtained at regular intervals using all of the signal data, but no averaging of multiple periods takes place. The signal was also analyzed using linearly interpolated low pass filtering and zero crossing detection, with an analysis window of 200 samples (about four periods) and a hop size of 50 samples.

Figure 2.4 shows the pitch signal and the Log-Magnitude spectrum results of LPFZCD analysis. The time-domain feature detection spectrum exhibits a measurement noise floor of about -42 dB, while the LMS detector exhibits one lobe of second harmonic distortion at -39 dB, and a noise floor of about -55 dB. The Zero Crossing detector overestimated the maximum frequency by as much as 6.1 Hz. (1.1%), and underestimated the minimum

frequency by 8.3 Hz. (1.8%). The LMS detector overestimated the maximum frequency by 0.14 Hz. and underestimated the minimum frequency by 0.13 Hz. (0.03% in both cases).

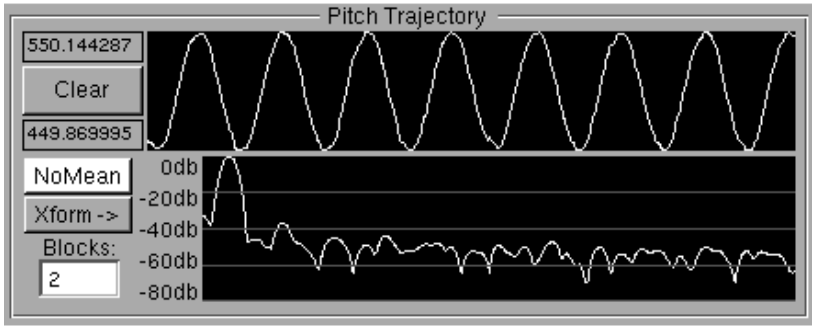

Figure 2.3: Time-domain and frequency-domain plots of LMS-Extracted pitch trajectory. Signal was noisy sine with sinusoidal vibrato.

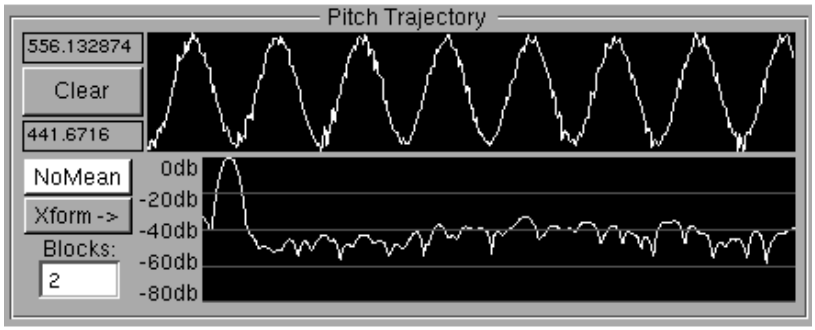

Figure 2.4: Time-domain and frequency-domain plots of Zero-Crossing-Extracted pitch trajectory. Signal was noisy sine with sinusoidal vibrato.

2.1.7 PPPT Relation to Maximum Likelihood Estimator

In the case of a deterministic signal with additive gaussian noise, the total signal can be viewed as a Gaussian noise process with a time-varying mean. Rife investigated the identification of single tones and multiple tones in noise [106][107]. In the case of the PPPT, the signal model is that of a periodic signal with additive noise. The PPPT coefficients are adjusted to yield the minimum error. If the additive noise is Gaussian, the minimum error

predictor is equivalent to the maximum likelihood predictor, yielding an error power equal to the variance of the additive Gaussian noise. If the additive noise is not Gaussian, the PPPT is the minimum variance predictor.

2.1.8 An Extension to the PPPT

The PPPT predicts any integer number of periods ahead, and thus does not yield a unique pitch estimate for a given periodic signal. An extension to the PPPT is proposed which uses multiple predictors to eliminate the problem of harmonic prediction error. The task of pitch detection of brass instrument signals is essentially one of determining which harmonic is being played, since the valve (or slide) position uniquely determines the harmonic series of notes which are easily played [2][101]. In the system constructed for this study, a family of 8 fixed length μ-LMS PPPT's was implemented with $M = 1$ (three predictor coefficients). The adaptation parameter μ was calculated according to Equation 2.21, where $\overline{x^2}$ was selected to be twice the maximum possible signal power. The delay line lengths were nearest-integer fractions of an assumed fundamental period. The predictors were configured to run constantly, with each predictor keeping track of its average squared error over the last 80 samples (10 ms. in this case). Since the lengths were fixed and the predictors ran constantly, the need to run parallel predictors at $P + 1$ and $P - 1$ was eliminated. The power of the input signal over the last 80 samples was also computed, for determination of the presence of a signal. The sampling rate of the system was 8012 Hz.

By inspecting the error power signal of each PPPT, the harmonic being played is determined. If the signal power is sufficiently low, the instrument (or voice) is assumed to be silent. If all error power signals are high (compared to the power in the input signal), it is likely that the instrument is playing a noisy or inharmonic tone. Such is the case in the attacks of instruments, or in the case of the voice, during the utterance of a fricative consonant. If some error power signals are high while others are low, the most likely period is that of the low error PPPT with the smallest integer period. Thus the signal power and error powers determine whether a signal is silent, whether the signal is periodic, and the period of the signal. Estimates of the instantaneous period are calculated (and compared, averaged, etc.) from the coefficients of one or more of the predictors which exhibit a low error signal.

Figures 2.5 and 2.6 show graphs of the harmonic detection PPPT predicting two multi-note

events. Eight normalized LMS PPPT's were used to predict eight harmonics. All eight error signals are displayed. The instantaneous period is displayed for the harmonic which is the most likely pitch, and is circled for ease of location in the graph. Each sample (plot pixel) of the error signal represents 10 ms., and each vertical grid line on the graph represents 100 ms.

The graph of Figure 2.5 is the display of the brass instrument PPPT predicting the musical notes B♭5 (415 Hz.), F5 (622 Hz.), and B♭6 (830 Hz.). The notes were sampled from a real trumpet. Each attack was approximately 40 ms, and the PPPT required 50 ms., 40 ms., and 10 ms., respectively, to lock onto each of the three notes. No erroneous estimates occurred once the correct note was identified. Continuous tracking of pitch is evident in the vibrato of the first and last notes.

The graph of Figure 2.6 is the result of prediction of small pieces of the same three notes randomly edited together. No attacks were included. This experiment was conducted to determine the average time required to identify a new note, with no noise present in the transition. The average time required to identify a new note was 13 ms., with none longer than 20 ms..

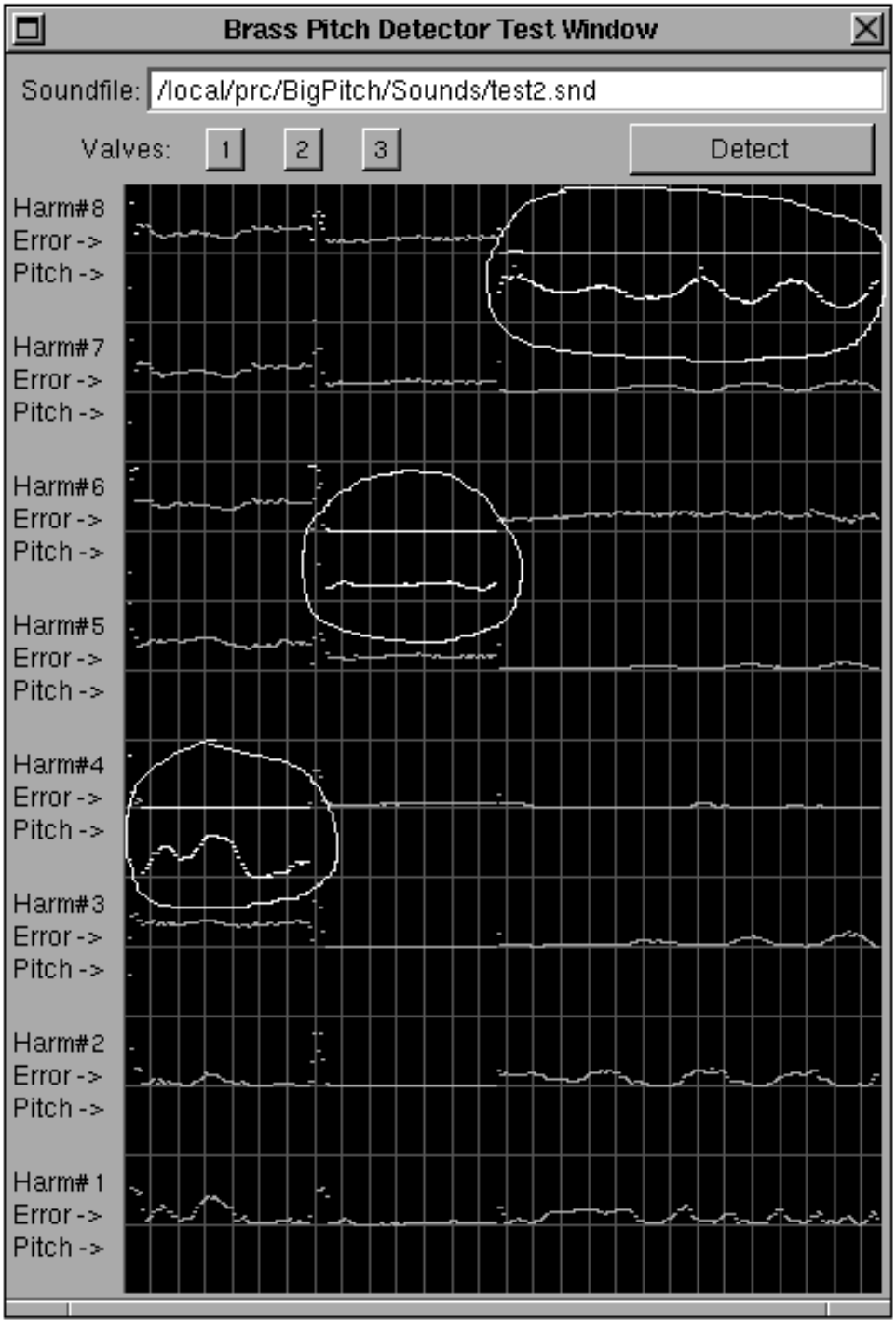

Figure 2.5: Harmonic detector PPPT results for the note sequence B♭5, F5, B♭6.

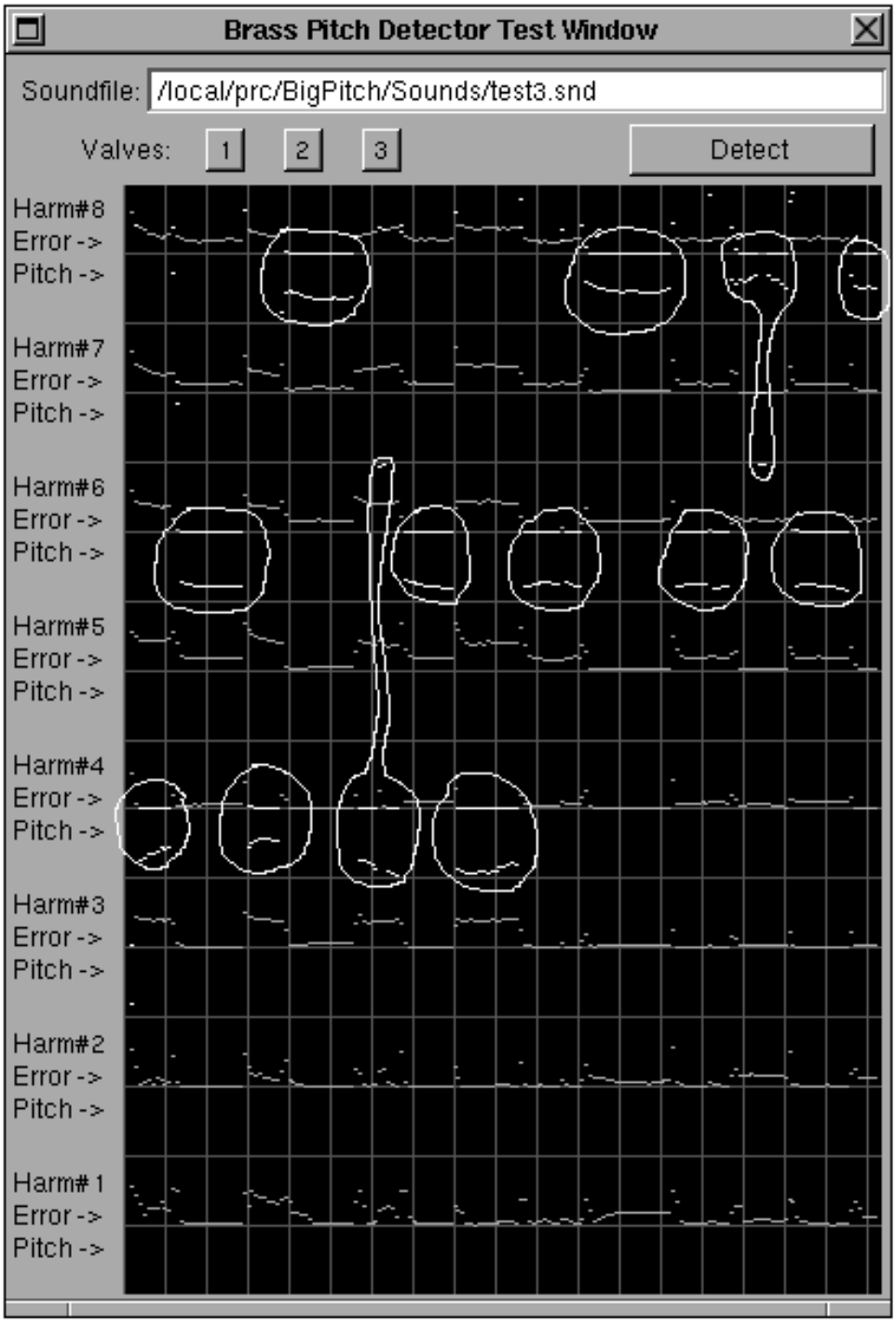

Figure 2.6: Harmonic detector PPPT results for randomized notes B♭5, F5, B♭6.

2.2 A Study of Singer Jitter and Drift

Studies of jitter and drift in western BelCanto singing voices have been conducted in the past, but typically on tones produced by singers instructed to sing with no vibrato [111]. One reason for this is that the jitter and drift components are easier to isolate and study when the vibrato is absent. Another reason is that many pitch detection methods yield noisy pitch estimates. Signal processing on low amplitude components in the presence of a large vibrato peak is difficult, because the jitter and drift components are often below the noise floor of the pitch detection algorithm.

The PPPT was shown to exhibit a noise floor of less than -55 dB in the presence of a single sinusoidal modulation signal and noise. This pitch detector was used to extract singer pitch signals, for the purpose of studying the behavior of jitter and drift as a function of pitch, loudness, and the behavior of jitter and drift in the presence and absence of vibrato.

Four singers, one each of the voice parts Soprano, Alto, Tenor, Bass, were selected for the initial study. All of the singers are judged to have excellent vocal quality, have received extensive private instruction, and each singer has over 10 years each of choral and solo singing experience. The singers were instructed to sing 30 long tones on the vowel /a/ (father). The singers breathed between each note, and were allowed to repeat any notes which they felt were uncharacteristic of their ability. The frequencies produced were selected individually for each singer, to ensure that 5 samples were available across the entire comfortable singing range. The notes sung are shown in Figure 2.7. The first set of five notes was performed at the dynamic level of Mezzo Forte (medium loud), with vibrato. This sequence was repeated at the same dynamic level, but without vibrato. Next, the same five notes were performed with vibrato at the dynamic level of Pianissimo (very soft), then without vibrato. Then the same sequence was performed at the dynamic level of Fortissimo (very loud), first with vibrato, then without. Even though some component of vibrato is present in many of the non-vibrato tones, the sung-tones produced under the non-vibrato requirement will hereafter be referred to as non-vibrato tones.

The sound files were digitized directly to 16 bit samples at a rate of 44.1 kHz. using a B&K 4006 microphone, an IMS MPA-4 microphone preamp, and a Sony DTC 1000ES Digital Audio Tape (DAT) machine. The files were transferred to computer disk using an Ariel DM-N digital microphone. The files were then down-sampled to a sampling rate of 5512.5

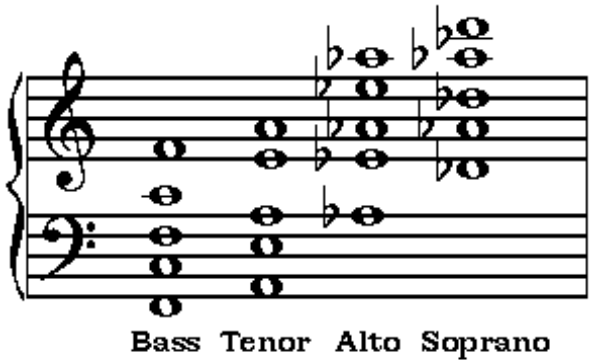

Figure 2.7: Musical notes sung by the singer vibrato test subjects.

Hz. by digital low-pass filtering at 2.5 kHz. cutoff frequency and decimating the resultant signal by a factor of eight. The low-pass filter used was designed with -96 dB stop-band rejection. Pitch signals were extracted from these files using a sampling rate of 100 Hz. by taking the average of each set of 55 PPPT pitch samples. This pitch signal sampling rate ensures that modulation information up to 50 Hz. was available for analysis. Figure 2.8 shows time-domain pitch signals extracted from the sung tones of the soprano subject KH and the bass subject PC. Vibrato and non-vibrato cases are shown, and the vibrato component is clearly visible in the non-vibrato tone of subject KH.

Figures 2.9, 2.10, 2.11, and 2.12 show the plots of the power spectral densities (PSD) of the pitch signals of the four singers. The upper plots show all PSD's for all samples of a given singer plotted on the same graph. The left plots are of the singer singing with vibrato, and the right plots are of the singer singing as instructed to produce no vibrato. The lower plots are the averages and standard deviations of all of the spectra. The jitter region (8-50 Hz.) of the mean spectrum was fit with a linear function, and intercepts of that line at 8 and 32 Hz. are marked for comparison of vibrato/non-vibrato characteristics.

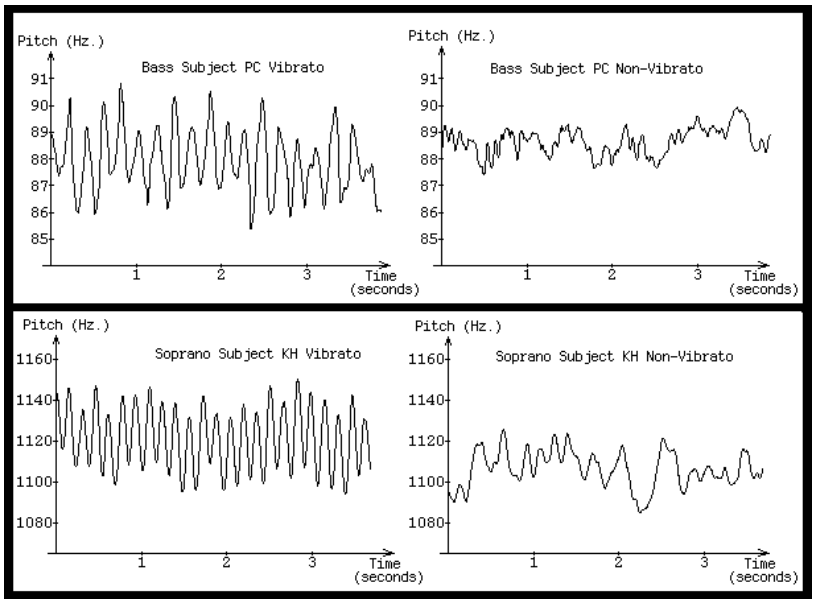

Figure 2.8: Time-domain pitch signals extracted from singer tones. The vibrato component is clearly visible in the non-vibrato tone of subject KH

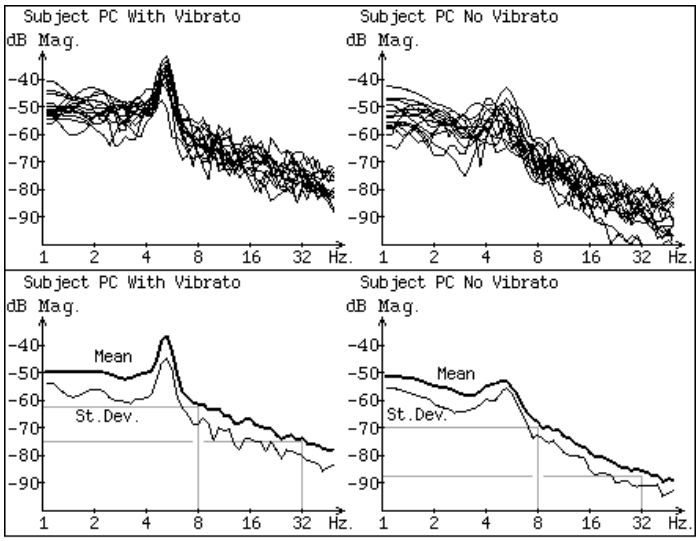

Figure 2.9: Power spectral densities of vibrato (left) and non-vibrato (right) tones of bass singer subject PC. The lower plots are the averages and standard deviations of the vibrato and non-vibrato PSD's.

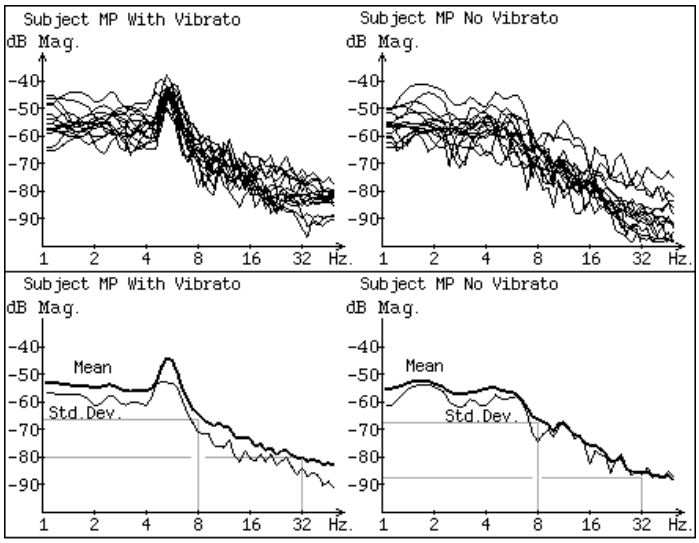

Figure 2.10: Power spectral densities of vibrato (left) and non-vibrato (right) tones of tenor singer subject MP. The lower plots are the averages and standard deviations of the vibrato and non-vibrato PSD's.

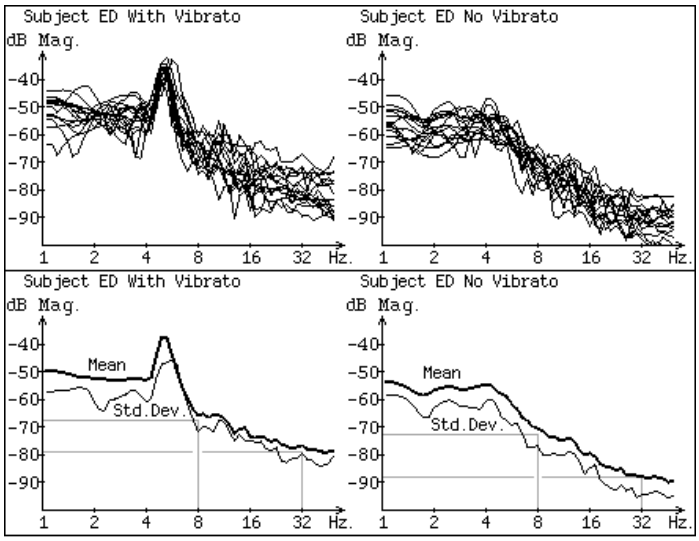

Figure 2.11: Power spectral densities of vibrato (left) and non-vibrato (right) tones of alto singer subject ED. The lower plots are the averages and standard deviations of the vibrato and non-vibrato PSD's.

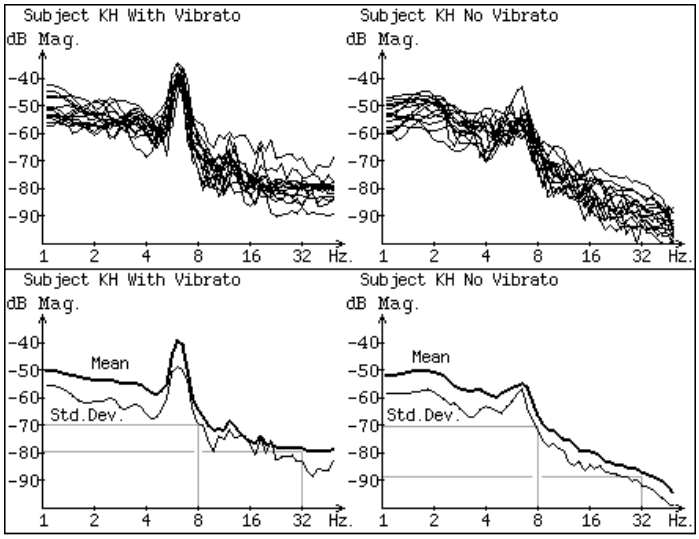

Figure 2.12: Power spectral densities of vibrato (left) and non-vibrato (right) tones of soprano singer subject KH. The lower plots are the averages and standard deviations of the vibrato and non-vibrato PSD's.

Each singer exhibited a peak in the non-vibrato tones near their natural vibrato frequency, as was found by Ternström and Friberg [111]. The location of this peak was relatively consistent within a particular singer across dynamic level and pitch, but not as consistent as the location of the peak produced when intentionally introducing vibrato into a tone. The consistency of vibrato rate in vibrato tones is shown by the low standard deviation directly under the average vibrato peak in each of the graphs. The vibrato peak in the tenor and alto non-vibrato tones is so low that location of a clear center frequency is difficult. The average location of the peak in the non-vibrato tones was significantly different than that of the vibrato tones in each particular singer, with the peak locations of the vibrato tones and non-vibrato tones being, respectively, 5.21 Hz. and 5.1 Hz. for the bass, 5.38 Hz. and 4.6 Hz. for the tenor, 5.27 Hz. and 4.0 Hz. for the alto, and 6.28 and 6.5 Hz. for the soprano.

Also consistent with Ternström's and Friberg's study was that the overall amplitude of jitter decreased with vocal range. That is, sopranos exhibit less jitter than basses. The four subjects studied here exhibited ordering according to this hypothesis. In the vibrato case, singers exhibited jitter spectra of about -65 dB (0.97 cents average) at 8 Hz, and rolled off at about 6 dB per octave. In the non-vibrato case, the jitter spectra were about -70 dB (0.55 cents average) at 8 Hz, and exhibited an average 8 dB per octave roll off. The standard deviations were consistently smaller in the drift region than the jitter region, which is consistent with the conscious/unconscious control difference between these two modulation components, and the fact that singers are highly trained in pitch control. The drift spectrum fell off slowly (roll off of about 1.5 dB / octave) from -50 dB (5.5 cents average) at 1 Hz. out to the vibrato peak at -50 dB average in the vibrato tones, and showed a decrease in the non-vibrato tones to -53 dB at 1 Hz. rolling off at about 2 dB per octave. This decrease implies that singers can hear their voices and control them better in the non-vibrato case than in the vibrato case, and is consistent with the model of drift as a random mechanism with control input from auditory feedback.

Ternström and Friberg investigated the non-vibrato case of eight singers on different vowels, but on one note and one dynamic level only. To investigate the dependence of jitter and drift on loudness, the PSD's of all pitch signals at a particular dynamic level were averaged in the vibrato and non-vibrato case. Figure 2.13 shows the plots of the power spectral density of the pitch signals of all singers in the vibrato case, arranged by dynamic marking. The broad dual peak nature of the aggregate vibrato peak shows the variability of vibrato rate

between different singers. Figure 2.14 shows the plots of the power spectral density of the pitch signals of all singer subjects in the non-vibrato case, arranged by dynamic marking. One spectrum, the tenor subject MP singing his lowest note pianissimo, was discarded in the computation of the average spectrum because of its extreme difference from all other entries.

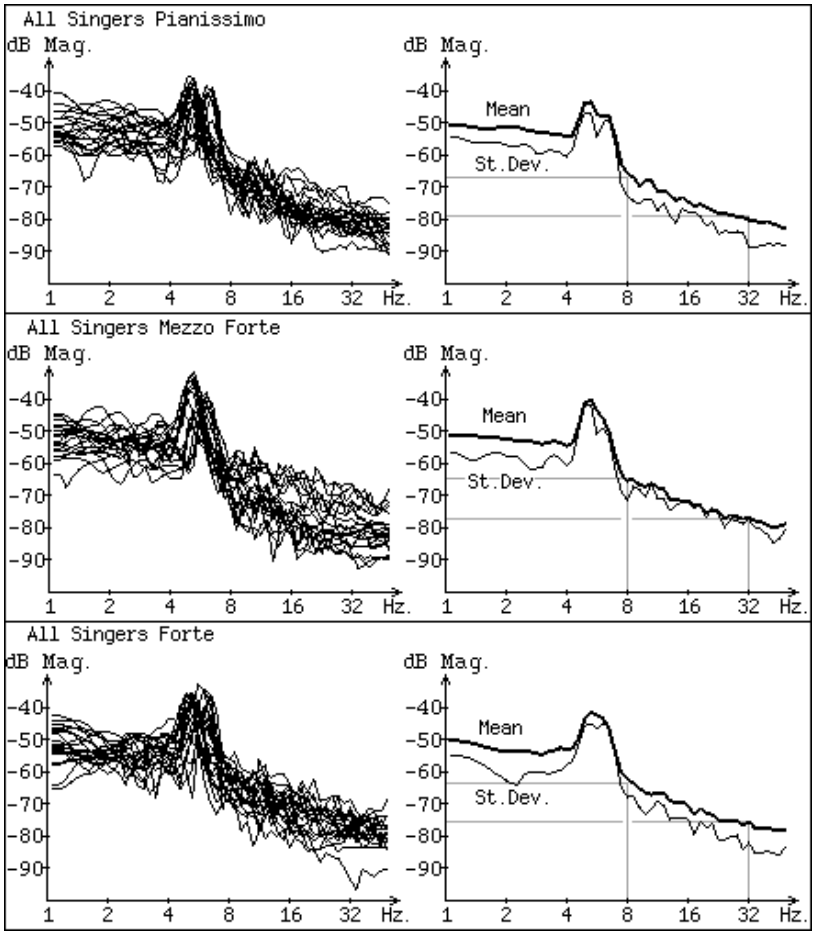

Figure 2.13: Spectra of non-vibrato pitch signals of all singers arranged by dynamic level.

The average PSD jitter curves show only a slight dependence on dynamic level in both vibrato and non-vibrato tones. An overall increase in jitter with increasing dynamic level was shown, with the deviation being about 4 dB between pianissimo and fortissimo. No significant change in spectral slope was observed, with all vibrato curves exhibiting a 6

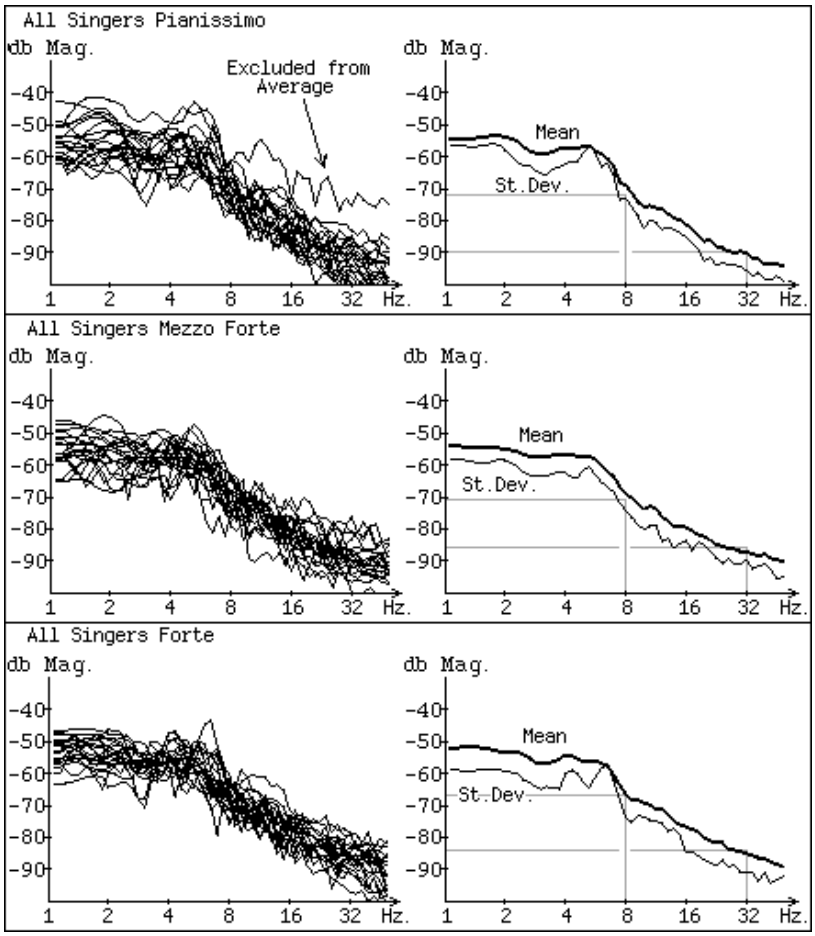

Figure 2.14: Spectra of vibrato pitch signals of all singers arranged by dynamic level.

dB/octave roll-off, and all curves without vibrato exhibiting an 8 dB/octave roll-off. The drift regions of the spectra showed no clear dependence on dynamic range, implying that the singers in this study could hear themselves and tune well at all dynamic levels. It should be noted, however, that in this study the singers were singing alone in an extremely quiet recording studio control room. In a noisier environment, singers hear themselves less well at soft dynamic levels, and a corresponding increase in drift should be expected.

A final investigation was conducted to determine whether jitter and drift depend on the position within the particular singer's range, the absolute pitch of phonation, or both. Two sets of spectral averages were formed. The PSD's of all singers at a particular region

in their vocal range were averaged in the vibrato and non-vibrato case. Averages were also done within 4 one-octave frequency ranges; 90-179 Hz., 180-359 Hz., 360-719 Hz., and 720-1439 Hz. Figure 2.15 shows the plots of the power spectral densities of the pitch signals of all singers for both vibrato and non-vibrato tones, arranged by position within the singer's range. The standard deviations of all of these plots are significantly larger than the mean spectra, indicating that grouping spectra in this way is an unreliable method of classification. Figure 2.16 shows the plots of the power spectral density of the pitch signals of the singers for both vibrato and non-vibrato tones, arranged by the absolute pitch. The standard deviations for these plots are quite small, indicating that the grouping of spectra by absolute pitch is a reliable method of classification.

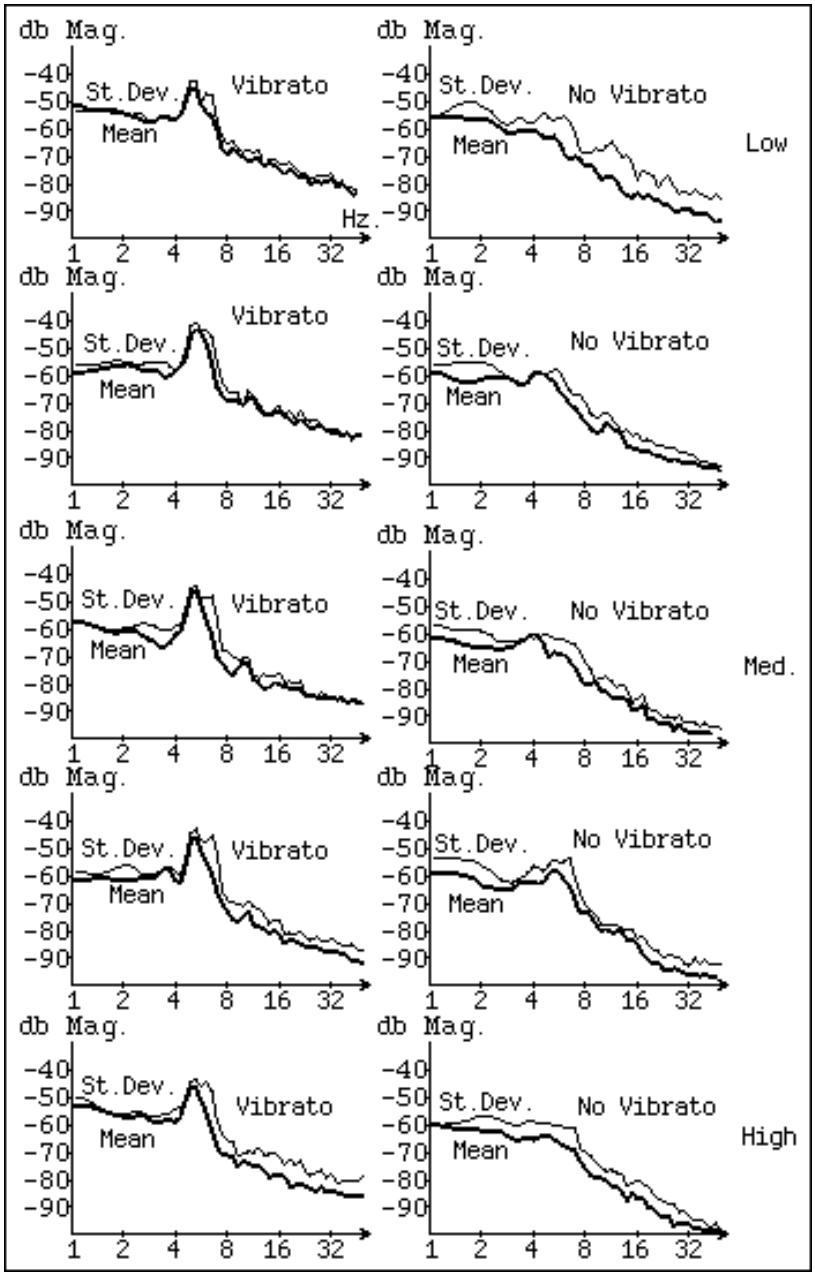

Figure 2.15: Singer pitch spectra averaged according to position within each singer's range (low to high). The high standard deviations show that averaging across relative range is unreliable.

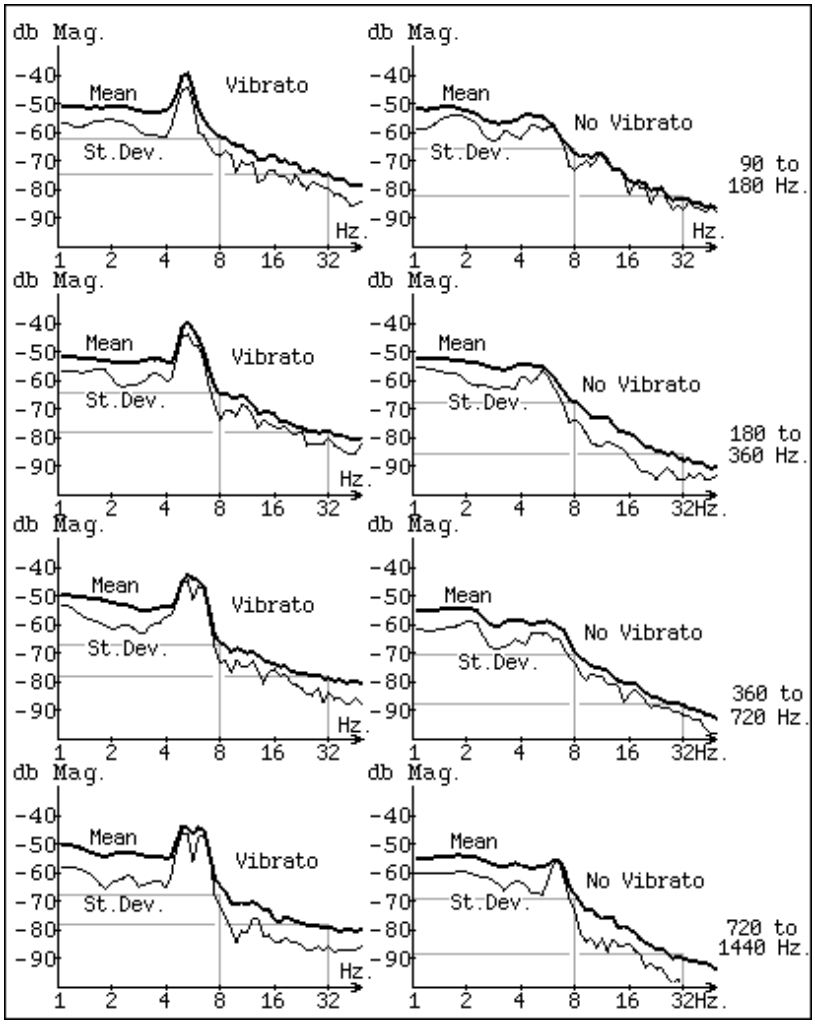

Figure 2.16: Singer pitch spectra averaged according to absolute pitch in one octave bands.

The jitter spectra showed a slight dependence on pitch, decreasing 2 dB per octave from low pitch to high pitch. The jitter curves exhibited a consistent slope for all ranges of 8.5 dB per octave in the non-vibrato case and 6 dB per octave in the vibrato case. The drift curves showed a weak dependence on pitch, decreasing about 1 dB per octave of increasing pitch.

2.3 Rules for Synthesis of Jitter and Drift

Figure 2.15 shows the line segment approximations to the jitter and drift spectra, in the vibrato and non-vibrato cases, arranged by pitch and dynamic level. The data indicates that a suitable control space for jitter must allow control over spectral height and slope as a function of dynamic level, phonation pitch, and presence/absence of vibrato. The minimum jitter is exhibited with no vibrato, at high pitch, and low dynamic level. This jitter is about -70 dB at 8 Hz., rolling off at 8.5 dB per octave. The maximum jitter is exhibited with vibrato, at low pitch, and high dynamic level. This jitter is about -60 dB, rolling off at 6 dB per octave. In both the vibrato and non vibrato case, increases in dynamic level account for about 4 dB increase in jitter across the entire dynamic range, and decreases in pitch account for about 2 dB per octave of jitter increase.

From the data and the model of drift production, the drift modulation component is most strongly affected by the singer's ability to hear him/herself. An extremely simple but nearly complete model of drift is a flat spectrum at -50 dB extending to the vibrato peak. The only significant deviations from this model found in this study were in the vibrato/non-vibrato comparison, which indicated a small increase in spectral roll-off in the non-vibrato case.

2.4 Spectral Deviation in the Voice

Shimmer is the term used to describe the fluctuations of amplitude and spectral shape in sounds. The voice exhibits significant shimmer, especially in tones sung with vibrato. For pure spectral or waveform synthesis methods, Maher and Beauchamp [97] suggest a solution to the requirement of a time-varying synthesis spectrum. The suggested solution is to use two wavetables, and interpolate between them synchronous to the vibrato oscillator. This

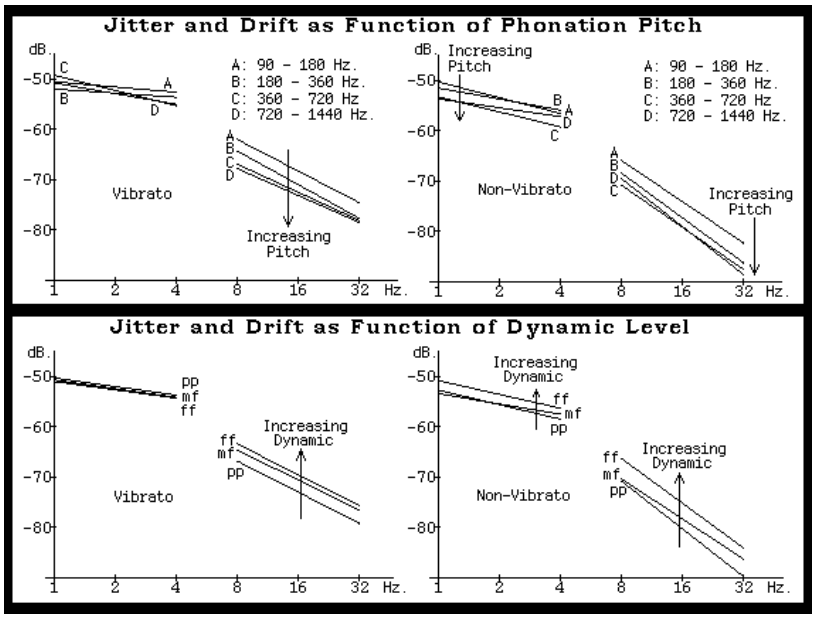

Figure 2.17: Line segment fits to jitter and drift spectra as function of phonation pitch and dynamic level.

way, the time variation due to vibrato is modeled.

The source filter model of the vocal mechanism provides many characteristics of the vibrato synchronous shimmer component automatically, because of the spectral modulation which comes about as the spectrum of the voice source moves under the relatively constant spectrum of the vocal tract filter. Figure 2.18 shows the time domain envelope of a synthesized vocal tone with vibrato. Two spectra are shown, one spectrum was computed at a high frequency point in the vibrato cycle, and the other spectrum was computed at a low frequency point. The shape of the vocal tract resonance envelope and resultant modification of the levels of particular harmonics is clearly evident in the region around the tenth harmonic.

There is at least one other component of vocal shimmer, however, due to variations in the source itself. If more vibrato synchronous deviation is needed than is provided by the source/filter model, the interpolated wavetable method outlined by Maher and Beauchamp can be used to model the glottal source. The dual wavetable interpolation is controlled by the vibrato oscillator.

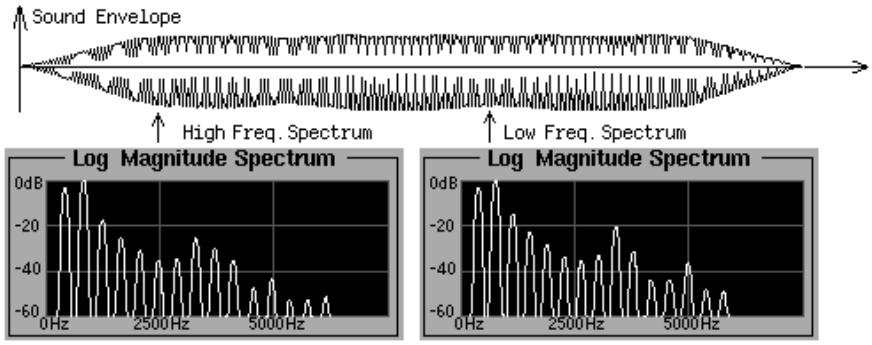

Figure 2.18: Time domain envelope (top) and two spectra (bottom) of a synthesized vocal tone with vibrato. The left spectral plot was calculated from a high frequency point in the vibrato cycle, and the right plot from a low frequency point.

Slow time-varying control of glottal source interpolation models the changes in the glottal wave shape with register, intensity, and mode of phonation. Figure 2.19 shows an example of slow interpolation of the glottal source, in the case of a tone increasing in power and vocal effort (a musical crescendo). The amplitude envelope shows the increase in volume over the event. The left spectral plot is the spectrum near the beginning (soft phonation) of the event, and shows decreased high frequency energy. The right plot shows the increased high frequency energy at the other extreme of the event.

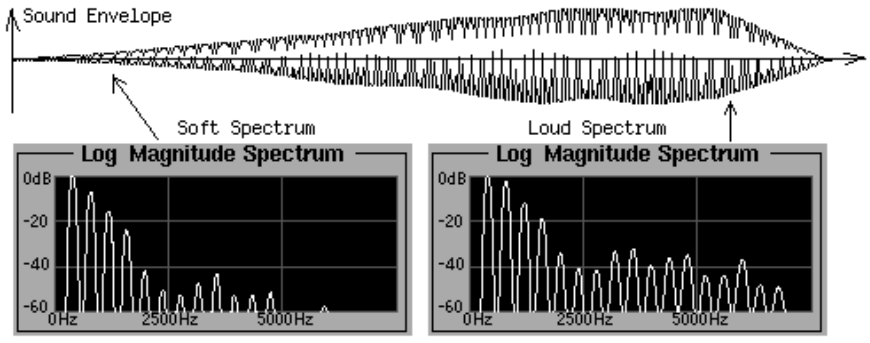

Figure 2.19: Time domain envelope and two spectra of a synthesized musical crescendo. The left and right spectra were computed at the soft and loud portion of the event, respectively.

2.4.1 Non-Linearities in the Vocal Tract

Other sources of spectral and amplitude deviations are the results of non-linearities in the vocal tract. One source of non-linearity is the non-stiff nature of the vocal tract walls, which are composed of tissue. The pliable nature of the vocal tract walls causes losses to be distributed throughout the vocal tract. Other distributed losses in the vocal tract are heat conduction losses and the energy loss due to the viscosity of air. The basic model assumes that the tract is both linear and lossless, and thus does not model non-linear vocal tract losses. Models accounting for this have been proposed [57][65], but are significantly more complex than the waveguide ladder filter formulation. The waveguide model can be modified to simulate non-linear losses [62]. To simulate the losses due to vocal tract wall flexibility, viscosity, and heat conduction, simple loss coefficients are introduced into all scattering gains in each scattering junction, or lumped at one point in the linear system. For more accurate modeling of non-linear distributed losses, elaborate time-varying filters may be included into each scattering junction. The transcutaneous coupling between the vocal and nasal paths via the velum and soft palate is modeled easily at the three way junction modeling the intersection of these paths. Losses due to the resistive component of glottal impedance can be modeled in the glottal model.

Other sources of non-linearity exist because waves propagate through the vocal tract tubes superimposed on a flow of air. The flow breaks into jets and vortices at boundaries, causing potential sources of sound radiation. The basic one-dimensional wave propagation model of the vocal tract as being composed of n-tubes does not address the problem of acoustics in a flow. Work by Kaiser [128], Teager and Teager [138][137], and Iijima, Miki and Nagai [127] demonstrate some phenomena attributed to non-linear fluid dynamic activity in the vocal tract.

Kaiser [128] proposes a jet-cavity flow model of the vocal tract, in which additional components of sound are generated throughout the vocal tract by flow phonomena. Vortices and jets interacting at boundaries such as the teeth cause sound components which are superimposed on the basic glottal wave. One striking example is at the teeth, where a toroid of turbulence is formed which modulates the effective area at this point. Kaiser concludes with a statement that linear models will continue to be the best available until the interactions and the acoustic consequences of jet-cavity flow phenomena are measured and

verified. Because of the inherent method of controlling the WGF acoustic tube model by a shape description, the phenonemon of time-varying modulated area functions are easily simulated.

The studies of Teager and Teager approached the problem of actually measuring flow in a vocal tract during phonation using hot-wire anemometry. They discovered various turbulent phenomena, including sheets of flow which oscillate between opposite walls of the vocal tract at frequencies of around 500 Hz. The modulating sheets of air found by Teager and Teager are more difficult to model than a simple modulation of area function. The possibility of a modulating path length is an extremely difficult problem to tackle in the WGF acoustic tube framework, because the samples of the vocal tract are assumed to be uniform in space. Also, the one dimensional wave nature of the WGF acoustic tube model breaks down in the presence of multiple possible paths through the same tube.

Iijima, Miki, and Nagai [127] investigated viscous flow in the glottis using finite element techniques to solve the two-dimensional Navier Stokes equations. The simulations showed the formation of vortices, and in the case of time varying flow, the vortices propagated downstream. The authors conclude that vortices at the glottal source should not be ignored in models of vocal fold vibration. Such findings of turbulent behavior motivate the research presented in the remainder of this chapter, specifically noise generation in the glottal source region.

2.5 Pulsed Noise in the Glottis

The quasi-periodic oscillations of the glottis exhibit small period-to-period deviations in the waveform, some of which is brought about by bursts of noise in or near the glottal oscillator. The passage of air at sufficient velocity through an aperture causes turbulent streaming, and thus noise is generated [136][113][114]. The flow is zero when the time varying aperture is closed, and the turbulence ceases if the aperture opens sufficiently or the flow decreases. The basic fluid dynamic equations quantifying turbulent jet formation and noise radiation were presented in Section 1.6. The Reynolds number, which is proportional to flow and inversely proportional to the radius of the aperture, indicates the likelihood of turbulence:

$$Re = \frac{2U}{\nu\sqrt{A\pi}} \tag{2.25}$$

where U is the volumetric flow and A is the effective area of the aperture. The kinematic viscosity of the fluid, ν, is defined as the ratio of the dynamic viscosity to the density, and is about 0.15 cm^2/s for dry air [7]. Turbulent streaming is likely if the Reynolds number is greater than a critical quantity, Re_{crit}, which is about 1,000 for a rectangular slit, and larger for circular apertures. If turbulence is present, noise is generated with a power proportional to V^8. The radiated sound power is computed from the volumetric flow by:

$$P \propto (\frac{U}{A})^8 \tag{2.26}$$

The center frequency of the principal peak in the spectrum of the turbulent noise is given by:

$$f = \frac{SV}{d} = \frac{SU\sqrt{\pi}}{2\sqrt{A^3}} \tag{2.27}$$

where S is the Strouhal number, which is 0.15 for the center frequency of noise spectral density. Tube resonances affect the formation and power radiated by turbulent jets. Vortex shedding is a related but quite different phenomenon which occurs at sharp edges and boundaries, producing sound with a power which depends on lower powers of the flow-to-area ratio. Hirschberg [124] gives power relationships of

$$\frac{U^4}{A} \qquad \text{and} \qquad \frac{U^6}{A} \tag{2.28}$$

for turbulent sound radiation in a tube, and vortex dipole sound radiation in a tube, respectively.

Figure 2.20 shows the characteristics of a typical cycle of oscillation of the glottal folds. Views a) and b) are superior and cross sectional views of the glottal folds. The drawings were made by the author after Hess [95] from the work on electro-glottographs (EGG) of Lecluse [75]. Graphs c) and d) are of the effective area and volumetric flow (flow glottograph). Graph e) is an EGG (electro-glottograph, or laryngograph), which measures laryngeal impedance using electrodes placed on the neck. The glottal impedance is maximum

when the folds are closed, so the EGG curve is inverted from the other curves describing glottal activity.

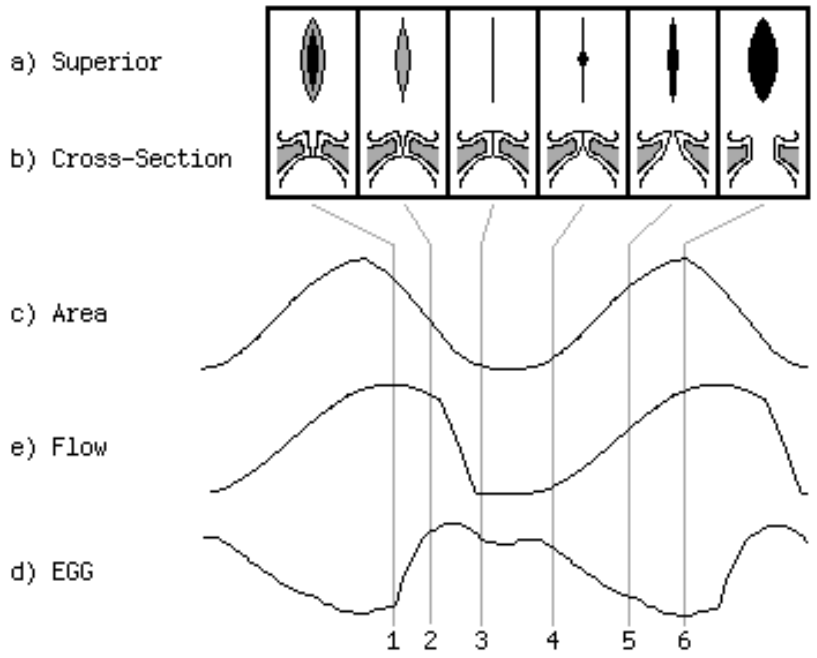

Figure 2.20: Superior and cross section views of the glottal folds at 6 phases of a typical cycle of oscillation. The time-varying area function, a flow glottogram, and a laryngogram are also shown.

From Fant's [8] reference to a Bell Labs film of glottal vibration, the glottal folds are approximately 11 mm wide and achieve a maximum breadth of 2.6 mm. Chiba and Kajiyama [6] observed an average flow of $140 cm^3/s$ at medium intensity phonation at 144 Hz. Assuming the functional form of the glottal flow from Equation 1.52 with $e_1 = e_2 = 0.7$, the maximum flow is about 300 cm^3/s. Kioke and Hirano [131] demonstrated a near-identical relationship between the width and the area of the glottal opening, and provided graphs and data on these functions. Using the numbers of Kioke and Hirano for glottal area and the functional form of glottal flow from Equation 1.52, an analysis of the Reynolds number was done for each of 18 positions within the glottal cycle. Table 2.1 shows the results of this analysis, and predicts the possible turbulent behavior of a typical glottal cycle assuming quasi-static free-space jet conditions. The six phases of oscillation shown in Figure 2.20 are marked in

Glottal Reynolds Number Analysis						
Phase	Area (cm^2)	U (cm^3/s)	R	Turbulent?	Power (dB)	Freq. (Hz.)
1	.27	300	4340	Yes	-44	285
	.225	297	4710	Yes	-39	370
2	.168	282	5175	Yes	-30	545
	.108	256	5860	Yes	-18	960
	.048	139	4773	Yes	-11	1760
3	.015	0	0	No	$-\infty$	0
	ϵ	0	0	No	$-\infty$	0
	ϵ	0	0	No	$-\infty$	0
	.001	4	952	Maybe	0	16815
4	.01	15	1130	Maybe	-34	1995
	.045	40	1420	Yes	-52	560
	.084	74	1920	Yes	-53	405
	.138	114	2310	Yes	-55	295
5	.189	157	2715	Yes	-55	255
	.231	199	3115	Yes	-53	240
	.264	237	3470	Yes	-52	230
6	.291	269	3750	Yes	-51	230
	.3	290	3980	Yes	-49	235

Table 2.1: Analysis of Reynolds number at positions within a typical glottal cycle.

the first column of Table 2.1.

The calculations of Table 2.1 indicate that a likelihood of turbulence exists for the entire open phase, but achieves maximum sound radiation power at the point where the vocal folds begin to close (Phase 1 of Figure 2.20). An extremely high power burst of noise is also likely at the glottal opening instant (Phase 4 of Figure 2.20), corresponding to highly pressurized air rushing through a small slit. As is shown in the Phase 5 cross-section of Figure 2.20, the aperture exhibits a sharp edge at the time of glottal opening, further indicating the likelihood of vortex formation. Thus it is expected that in some cases there are two pulses of noise per cycle, one when the the vocal folds are opening, and one when they are closing. Higher values of flow, common in singing and loud speech, would yield linearly higher Reynolds numbers and center frequencies, and would increase the radiated noise power by the exponentiated flow-to-area relationship. The pulse at the opening phase

exhibits the greatest power, with the pulse corresponding to glottal closure being secondary in power. Similarly, the spectral peak of the radiated noise power spectrum is at the highest frequency at the opening phase, decreases until the point of full glottal opening, and rises again as the glottal folds are closing. Figure 2.21 shows graphs of the Reynolds number, the noise power, the center frequency, and the glottal area and volumetric flow with the same phases of glottal oscillation marked as in Figure 2.20.

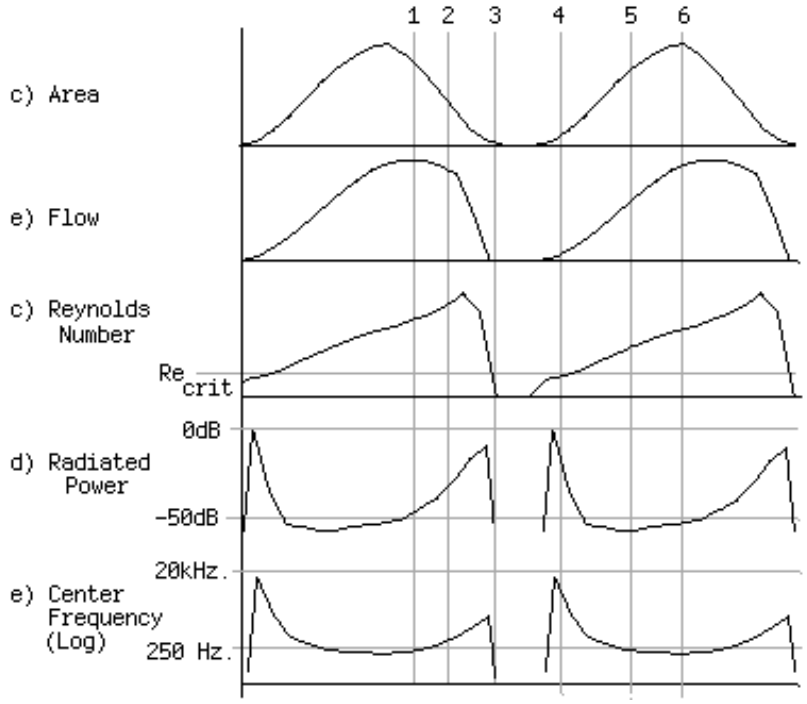

Figure 2.21: Graphs of the time-varying glottal area function, a flow glottogram, the Reynolds number, the radiated power, and the center frequency of the power spectrum. Six Phases of a typical cycle of glottal oscillation from Figure 2.20 are marked.

The simple analysis of Table 2.1 assumes that turbulence is instantly born when the dimensions and flow quantity are suitable, and the disturbance dies as quickly. A more detailed analysis of the behavior of pulsed turbulence was done by Kingston [130], after the work of Brown, Marglois, and Shah [115]. These studies investigated the effects of turbulent jets in tubes driven by pulsating sources of flow. The ratio of normalized pulsation frequency, Ω, to the Reynolds number was identified as an important measure of turbulent behavior. The normalized pulsation frequency is a unitless quantity defined as:

$$\Omega = \frac{r^2 \omega}{\nu} \tag{2.29}$$

where r is the radius of the tube. Another convenient expression for Ω is:

$$\Omega = \frac{2AF_0}{\nu} \tag{2.30}$$

where A is the tube cross-sectional area, and F_0 is the frequency of phonation in Hz. Recalling the large frequency range of the singing voice, and allowing for large deviations in tube cross-sectional area depending on the vowel, Ω can range from 10^2 at 50 Hz. in an /μ/ (boot) vowel, to 10^5 at 2000 Hz. in an /a/ (father) vowel.

Three regimes of pulse-turbulence interaction were observed by Kingston, corresponding to high, medium, and low ratios. For low pulsation frequencies ($\Omega/Re < 0.04$), the flow is quasi-steady and follows the behavior indicated by the analysis of Table 2.1. For high pulsation frequencies ($\Omega/Re > 0.1$), the turbulence is steady and independent of flow pulsations. For intermediate frequencies, the relation between turbulence and pulsation is complex, and is characterized by vortex resonance phenomena. The average value of the Reynolds number from Table 2.1 is 2750. Assuming a minimum vocal tract tube area of 0.15 cm^2, the transition region from pulse-turbulence interaction to non-interaction lies between 55 and 140 Hz. The maximum Reynolds number is 5860, yielding a transition region bounded by 120 and 300 Hz.

From the calculations, pulsed turbulence is expected at phonation frequencies below 200 Hz., which is located within the vocal range. Even allowing for large deviations in the assumed parameters of flow and glottal area, it is still expected that low notes sung by bass singers might exhibit pulses, or perhaps dual pulses. Kingston also studied the relationship between pulsation amplitude and turbulent behavior, but restricted most of his analysis to amplitude modulations of less than 50%. In the case of normal glottal oscillation, the amplitude is 100%, and the assumptions of Brown are less valid. Highly modulated flow could significantly affect the location of the transition region separating steady from pulsed turbulence. The phenomenon of vortex shedding is expected in the glottal area as discussed by Hirschberg [124] and simulated by Iijima, Miki, and Nagai [127].

The presence of pulsed noise in singer voices is investigated in the next sections from a

digital signal processing viewpoint, concentrating on extracting the aperiodic part of the glottal wave, and inspecting the extracted residual for time domain structure.

2.6 Methods for Extraction of Non-Periodic Part of Glottal Waveform

To investigate noise in the any quasi-periodic signal, techniques for identifying and separating the periodic and non-periodic parts are required. Three methods of extracting the non-periodic part are discussed, one operating in the frequency domain, and two operating in the time domain. All of these methods yield similar results, but are slightly different due to the definition of periodicity that each assumes.

2.6.1 Noise Extraction by Frequency Transform

One method of extraction uses the Deterministic plus Residual Model of Serra [160], employing short term Fourier analysis to locate and remove sinusoidal peaks in the frequency domain. By any definition of periodicity, the sinusoidal peaks corresponding to the periodic part of the signal should be spaced nearly evenly, aiding the location process. Any spectral component not harmonically related to the fundamental frequency must be classified as part of the noise signal. After the sinusoidal components are carefully identified and removed, the remaining signal (noise residual) can be resynthesized by inverse transforming. Alternatively, the sinusoidal component is resynthesized and subtracted from the original signal in the time-domain to yield the residual. Figure 2.22 shows a periodic waveform with additive noise and a frequency domain representation showing sinusoidal peaks in the presence of noise.

2.6.2 Noise Extraction by Periodic Prediction

Another extraction method operates in the time domain and uses a least-squares periodic predictor [44][42]. The prediction process yields an error signal which is the non-periodic component of the signal (the component we wish to study). The form of a periodic predictor which predicts the signal $x(n)$ is:

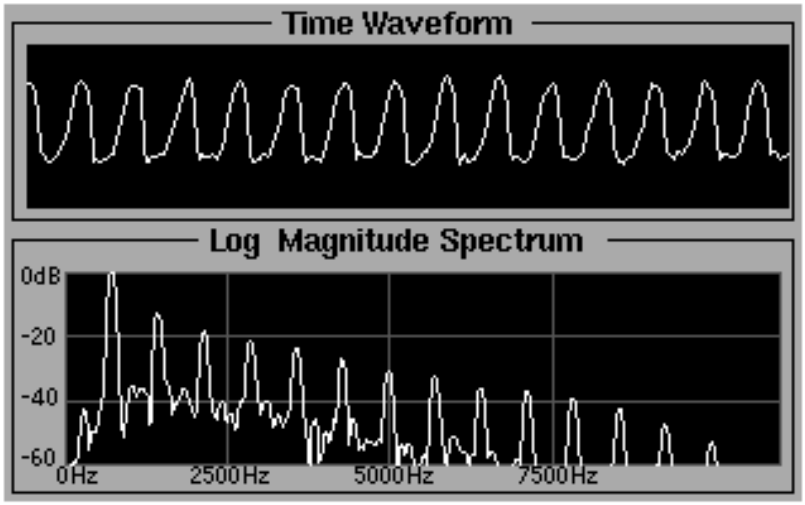

Figure 2.22: Time domain waveform of periodic signal with noise. DFT of signal shows harmonic peaks in the presence of noise.

$$\hat{x}(n) = \sum_{i=-M}^{M} x(n - P + i)c(i) \tag{2.31}$$

where M is some appropriately chosen small number, P is an integer estimate of the period, and $c(i)$ are the predictor coefficients.

The number of filter taps ($2M+1$ adjustable weights) is usually small, and the algorithm for adjusting the weights can be quite simple, since the weights need not change once they are adjusted for a given periodic signal. The Least Mean Squares (LMS) algorithm [163][164] was used for the examples of this study. LMS is a gradient steepest descent algorithm using the instantaneous error as an estimate of the gradient of the error surface. Each coefficient is adjusted at each sample by an amount proportional to the instantaneous error and the signal value which is associated with the coefficient being adjusted. The LMS update equation is:

$$C_{n+1} = C_n + 2\mu X_n \epsilon(n) \tag{2.32}$$

Where C_n is the filter coefficient vector:

$$C_n = \begin{bmatrix} c(-M) \\ c(-M+1) \\ \vdots \\ c(M) \end{bmatrix} \tag{2.33}$$

X_n is the current sample vector in the filter memory:

$$X(n) = \begin{bmatrix} x(n-P-M) \\ x(n-P-M+1) \\ \vdots \\ x(n-P+M) \end{bmatrix} \tag{2.34}$$

and $\epsilon(n)$ is the non-periodic error signal, defined by:

$$\epsilon(n) = x(n) - \hat{x}(n) \tag{2.35}$$

The adaptation constant μ controls the dynamics (and stability) of adaptation. Stability is ensured if:

$$\mu < \left((2M+1)\overline{x^2}\right)^{-1} \tag{2.36}$$

where $\overline{x^2}$ is the signal power. If the adaptation parameter μ is adapted dynamically, the Normalized LMS algorithm results:

$$C_{n+1} = C_n + \frac{\alpha}{(2M+1)\overline{x^2}} X_n \epsilon(n) \tag{2.37}$$

where the signal power is computed over the last 2M+1 (or greater) samples. The parameter α is any positive number less than 1.

2.6.3 Noise Extraction by Period Similarity Processing

Another time-domain method uses period similarity processing [142], with the added improvement of sinc-interpolated sampling rate conversion [159][162]. Each period is resampled by sinc interpolation. This operation is given by:

$$\hat{x}(nT_2) = C \sum_{i=-\infty}^{\infty} x(i) sinc(i + nT_2 + \theta) \tag{2.38}$$

where T_2 is a sampling period, C is a gain constant, and θ is a time offset. Figure 2.23 shows the conversion of a sampled data signal to a continuous time signal using sinc interpolation.

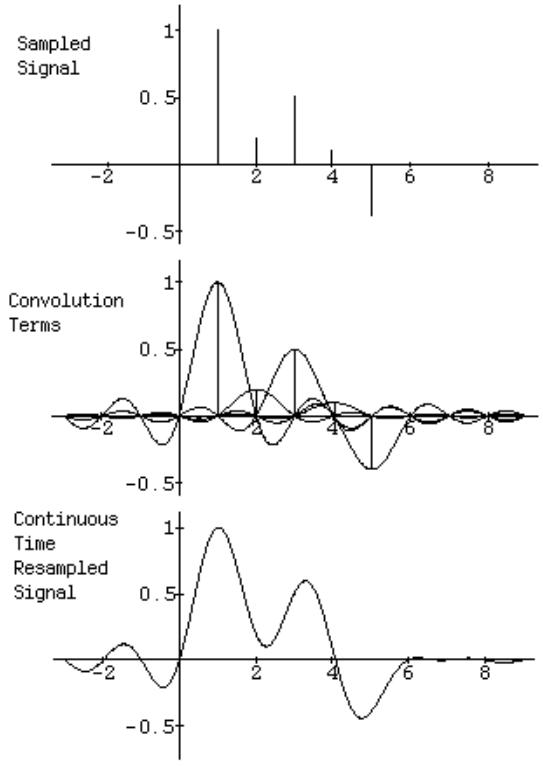

Figure 2.23: Sinc interpolation of the sampled data signal $Z^{-1}+0.2Z^{-2}+0.5Z^{-3}+0.1Z^{-4}-0.4Z^{-5}$, to yield a band-limited continuous-time signal.

The sampling parameters T_2, C, and θ are then selected to yield a least squares difference between the current resampled period and a prototype period. As the resampled periods are successively added to the prototype to yield a smoother prototype period, the noise component in the prototype approaches zero in the limit as the number of periods approaches infinity.

Denote the prototype period as the vector sum of the resampled periods:

$$X_{\text{Proto}} = \sum_{k=1}^{\text{number of periods}} \hat{X}_k \tag{2.39}$$

$$\hat{X}_k = \hat{x}(nT_2, theta, C) \tag{2.40}$$

where $\hat{x}(nT_2, theta, C)$ is the kth resampled period with the parameters T_2, θ, and C selected to yield a least squares difference between the kth resampled period and the prototype period:

$$\sum_{n=1}^{\text{period length}} |X_{\text{Proto}}(n) - \hat{X}_k(n)|^2 = \text{Minimum} \tag{2.41}$$

The prototype is subtracted from each period to yield the period residuals:

$$X_k \text{ Residual} = \hat{X}_k - \hat{X}_{\text{Proto}} \tag{2.42}$$

This process has the advantage that no averaging of multiple periods takes place in forming a period of residual. The time domain connection between the signal and the residual is preserved, simplifying later evaluation of the noise signal. One further advantage of the period similarity method is that any period-to-period correlations in the noise signal (provided the correlation decays with increasing time lag) average to zero in the formation of the prototype. Any short term correlations in the noise signal (perhaps due to formant resonances, for example), average to zero in the prototype, and are reflected in the residual signal. Figure 2.24 schematically shows period similarity processing.

A combination extraction and analysis method which is similar to the period similarity

processing method was proposed by Schumacher and Chafe [139]. This method will be discussed in Section 2.8.

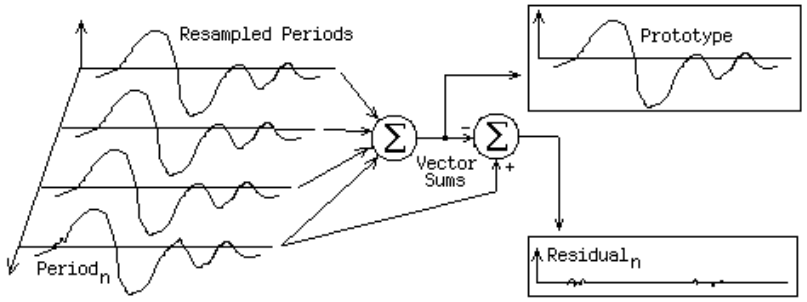

Figure 2.24: Schematic diagram of the process of averaging resampled periods to form a prototype. The prototype is subtracted from each period to yield residual periods.

Noise extraction by period similarity processing was performed on an example vocal tone. A bass singer was asked to sing a 100 Hz. tone on the neutral vowel /∧/ (bug). This vowel corresponds to a vocal tract filter which colors the glottal signal so little that inverse filtering is often not required to find the glottal closing time. Period similarity processing was performed over 200 periods, and the prototype and residual periods were formed. Figure 2.25 shows a few cycles of the original waveform, the prototype waveform, and the amplified residual signal. Noise is clearly evident in the original waveform. Some structure seems evident in the residual signal, but clearly it is difficult to conclude anything by visual inspection of so few cycles.

2.7 Methods for Analysis

Two new methods of analyzing the extracted noise residual signals are presented. The techniques involve identifying the periods of the original signal. Identification of the periods is already accomplished if extraction is performed by the period similarity method. Identification of the periods involves detection of some time domain feature using a method such as low-pass filtering and zero-crossing detection. Each detected period yields a pointer into the time-domain residual signal, and thus segments the signal into 'noise periods'. The

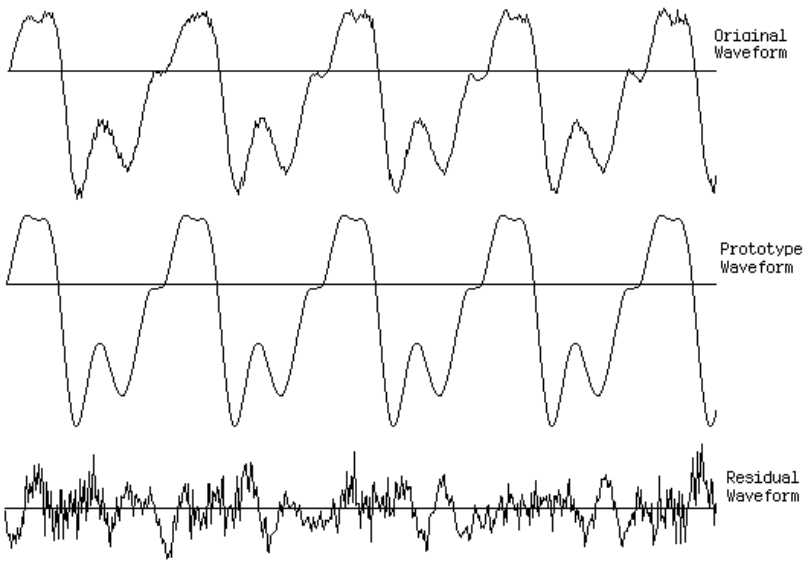

Figure 2.25: Original waveform (top), periodic prototype (center), and amplified residual (bottom) of male vocal tone sung at 100 Hz. on the vowel /∧/ (bug).

noise periods can be inspected for features which occur at particular times within each period. Each residual period can be subdivided and processed in sections containing particular signal events, such as glottal closing and opening epochs [91][90][92].

One other analysis technique involves a simple visual inspection of the time domain signal. This is instructive for certain signals where the noise bursts are clear enough to be easily viewed, but it does not yield quantitative results for use in comparison and modeling.

2.7.1 Period-Synchronous Noise Power Analysis

In the period-synchronous noise power analysis method, the noise power (sum of squared sample values) is computed in each of the sub-period sections. These powers are plotted in three dimensions, with height representing power, one axis representing the period number, and one axis representing the position within the period. Inspection of this 'noise period power surface' shows clear ridges and valleys running in the direction of the period number if the signal contains pulsed noise. The duty cycle of the noise pulses is deduced from the width of the peaks and valleys, and the dynamic range of the noise is deduced from the

heights of the peaks and valleys. The Noise Dynamic Range (NDR), representing the ratio of the average ridge height to the average valley height, is used for analysis of signals within this paper. Figure 2.26 shows the noise period power surface of the vocal tone of Figure 2.25.

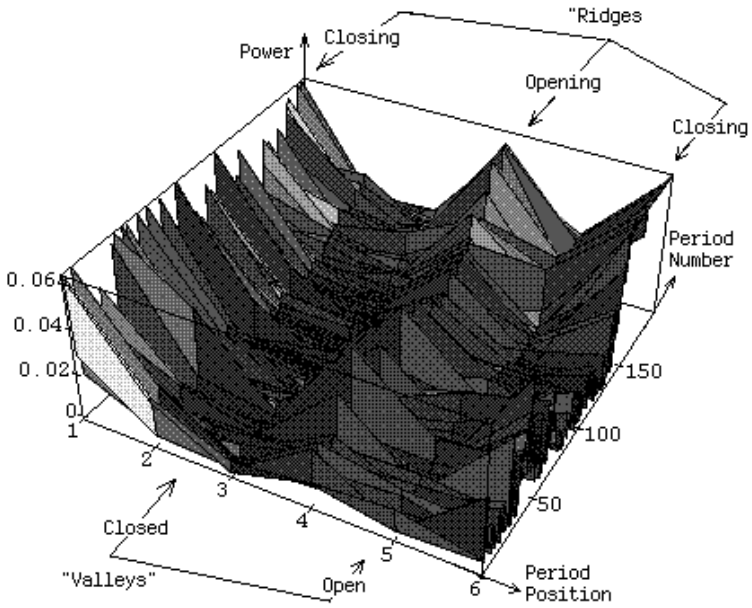

Figure 2.26: A noise period power surface of the residual signal extracted from a male vocal tone sung at 100 Hz. on the vowel /∧/ (bug). Glottal opening and closing phases are marked.

2.7.2 Noise Period Spectrum Analysis

The third analysis technique involves performing Discrete Frequency Transforms (DFT) on each of the noise period subsections. Each DFT is used to compute a power spectrum. The power spectra corresponding to each particular period position are averaged across all periods, yielding smoothed estimates of the power spectrum of the noise at each position within a typical period. This provides information about the time-varying nature of the noise signal spectrum. These spectra can be plotted in 3D with height representing intensity, one axis representing position within the period, and one axis representing frequency. Figure 2.27 shows a plot of the noise period spectra of the vocal tone of Figure 2.25. The average

spectral deviation of the residual component is easily seen.

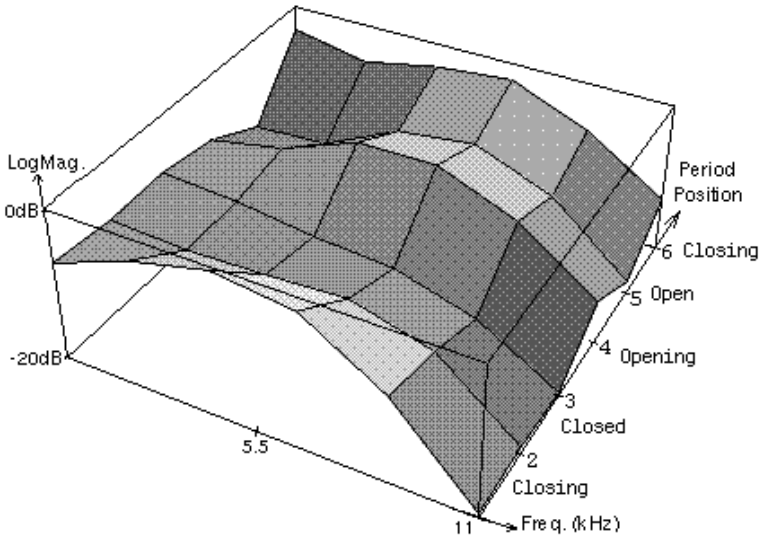

Figure 2.27: A noise period spectral surface of the residual signal extracted from a male vocal tone sung at 100 Hz. on the vowel /∧/. Glottal opening and closing phases are marked.

2.8 Extensions and Similar Methods

A combination extraction and analysis method similar to the period similarity processing method was proposed by Schumacher and Chafe [139]. This method calls for the identification of the periods, then resampling to make all periods equal to the length of the longest. Each period may be subtracted from any other, generating plots which display the portion of the waveforms which vary most. Discrete Fourier Transforms can be performed at one location across many features, yielding information about features which occur periodically, but at multiples of the period of the signal being investigated. Such periodic disturbances are called subharmonics, and are discussed in Section 2.11. The differences between the period similarity processing/power surface/spectral period computation methods and the methods of Schumacher and Chafe are:

1. In period similarity processing, sinc-interpolated resampling is performed to

minimize the least squares difference between the waveform and all other waveforms. This provides the minimum power estimate of the residual signal. In the method of Schumacher and Chafe, linear interpolated resampling is done to match period lengths only. Lengths are defined by zero crossings.

2. In period similarity processing, the formation of the prototype waveform provides an estimate of the least-squares periodic component of the signal. The method of Schumacher and Chafe does not call for the formation of the period prototype.

3. The method of period similarity processing was devised to investigate the time-domain microstructure of noise in quasi-periodic signals. Periods are divided into sub-sections of a few samples to compute power and spectra. The analysis methods of Chafe and Schumacher were devised to search for subharmonics, which are periodic occurrences. Period similarity processing is inferior in performance in the presence of subharmonics, because part of the subharmonic behavior is coded into the prototype.

4. Period similarity processing, power surface computation, and spectral period computation provides characterization of the disturbances within a typical period. The power surface method of visualization is extremely simple to compute and yields an arrangement of the data which is easy to evaluate visually. The spectral period method provides information about the spectrum of the noise component at specific locations within a typical period. The method of Schumacher and Chafe allows browsing and exploration of feature differences and similarities between any two periods, but the cost is increased computational complexity.

2.9 Male Singing Voice Extraction and Analysis Examples

Three voice signals were selected for analysis: a male singer singing the neutral vowel /∧/ (bug) at 100 Hz, the same vowel in falsetto register at 275 Hz, and the voiced fricative /z/ at 100 Hz. Figure 2.28 shows the power surfaces and spectral periods. The dual pulse nature of the chest register tone is easily seen in the power surface plots. The long-term Signal to Noise Ratio (SNR) of the chest register tone was 6 dB less than that of the falsetto tone, but the NDR for the chest register tone was 6 dB greater than that of the falsetto

tone. The power surface of the falsetto residual shows no clear noise ridge, indicating that the breathier quality often associated with the male falsetto voice might come from the flatter less-modulated noise signal. Voiced fricatives generate a pulsed noise component. As the glottal folds open and close, the pressure in the chamber behind the constriction is modulated, and thus the flow rate through the constriction varies synchronous with the glottal oscillation [38][39]. The voiced fricative in the study exhibited broad single pulse surfaces.

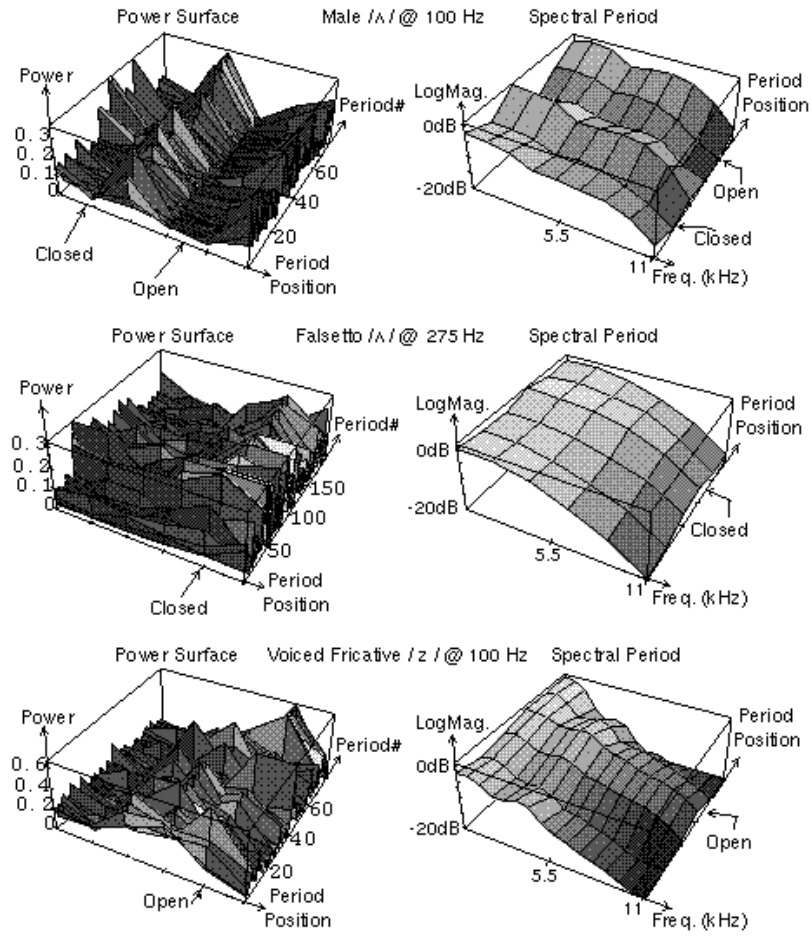

Figure 2.28: Power surfaces (left) and spectral periods (right) of male vocal tones.

2.10 A Study of the Noise Characteristics of Singers

Recordings were made for analysis of a number of trained singers. Four choral singers were studied: one bass, one tenor, one alto, and one soprano. The singers were asked to sing long tones across their entire comfortable range on the neutral vowel /∧/ as in bug. This vowel yields an output waveform close to the glottal waveform, so inverse filtering is usually not necessary. The sequences of notes were sung at three dynamic levels, pianissimo (very soft), mezzo forte (medium loud), and fortissimo (very loud). Figure 2.7 shows the notes sung by the singer test subjects. The sound files were digitized directly to 16 bit samples at a rate of 44.1 kHz. using a B&K 4006 microphone, an IMS MPA-4 microphone preamp, and a Sony DTC 1000ES Digital Audio Tape (DAT) machine. The files were transferred to computer disk using an Ariel DM-N digital microphone. The files were then down-sampled to a sampling rate of 22.05 kHz. by digital low-pass filtering at 11 kHz. cutoff frequency and decimating the resultant signal by a factor of two. The low-pass filter used was designed with -96 dB stop-band rejection. A 200 period sample was extracted from the center of each tone for pulsed noise extraction and analysis. Pulsed noise extraction was performed by the period similarity method. The average Normalized Noise Power (NNP) was calculated for each of the analyzed signals. The NNP in dB of the quasi-periodic signal $x(n)$ of length N is defined as:

$$10log_{10}\left(\frac{\sum_{n=0}^{N-1}\epsilon(n)^2}{\sum_{n=0}^{N-1}x(n)^2}\right) \tag{2.43}$$

2.10.1 Average Noise in Singer Voices

Figure 2.29 shows the average NNP for the four singers as a function of the position within the particular singer's range. The three superimposed curves correspond to the three dynamic levels. The maximum noise level deviation across dynamic level was 10 dB. The average noise level deviation between any two dynamic levels for the same note was 2.35 dB. The maximum deviation across pitch for a given dynamic was 19.6 dB, and the minimum was 3.1 dB. The average deviation from the lowest to the highest note was 10.6 dB.

The results indicate that the NNP depends little on dynamic, and much more on pitch, and further that this dependence is consistent across subjects of different phonation frequency ranges as well as within a particular subject.

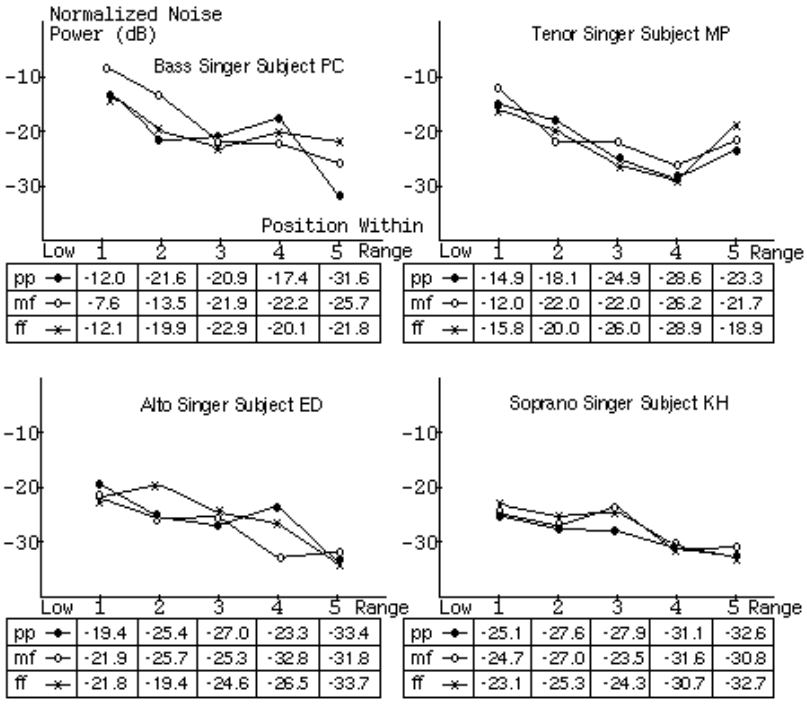

Figure 2.29: Average Normalized Noise Power (NNP) as a function of frequency. The superimposed curves correspond to three dynamic levels of singing.

Figure 2.30 shows all plots together, aligned by pitch. A line proportional to inverse frequency is fit through the curves, and shows that the NNP rolls off inversely in frequency. The data of post operative laryngeal surgery patients of Muta, Baer et.al. [132] is consistent with the inverse relationship of NNP with pitch.

There is one consistent exception to the monotonic decrease in NNP, and this occurs at the highest note of the tenor subject MP in all three dynamic levels. This was theorized to be due to the changing of the mode of phonation from the chest (normal) to the head (falsetto) register. To investigate this theory, four additional male singers were asked to sing the same sequence of five notes across their comfortable range. They were instructed to sing the highest note in falsetto register immediately after singing in in chest voice, then to sing

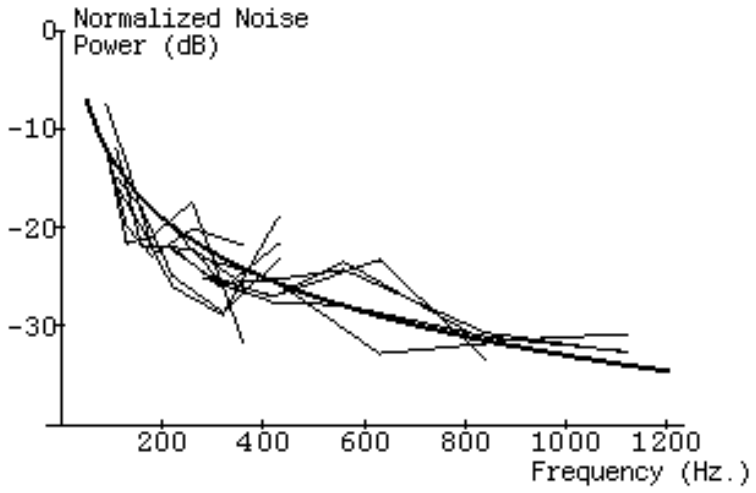

Figure 2.30: Average Normalized Noise Power (NNP) of all singers as a function of frequency. The bold curve is a log plot of $\frac{k}{frequency}$, to show that the noise component rolls off roughly according to this relationship.

one more falsetto note above the duplicated chest/falsetto tone. The NNP curves of Figure 2.31 show that the falsetto signals averaged 2.4 dB greater in NNP than the corresponding chest register tones.

To verify the 1/f hypothesis of NNP, data was taken from eight additional singers, two of each voice part. The extraction was performed using linear periodic prediction as described in Section 2.6.2. The average NNP as a function of frequency is plotted in Figure 2.32, along with a plot of Log(k/f). A curve of the form $a * f^b$ was fit to the data of all twelve singers. The least-squares fit was $59 * f^{-1.2}$, which closely agrees with the hypothesis of an inverse proportionality relationship of NNP with frequency.

Given Equation 2.26, which predicts that the radiated noise power varies as the eighth power of flow, it may seem contradictory that measured noise in singer voices was largely independent of dynamic level, and inversely proportional to frequency. A study of airflow in singer voices [134] found that airflow increases slightly with both increasing pitch and loudness, but often airflow decreases in higher tones. This is also consistent with the findings of Cavagna and Margaria [116]. Higher tones often are produced with a more 'pressed' voice, and the overall glottal resistance changes as a result. The nature of noise

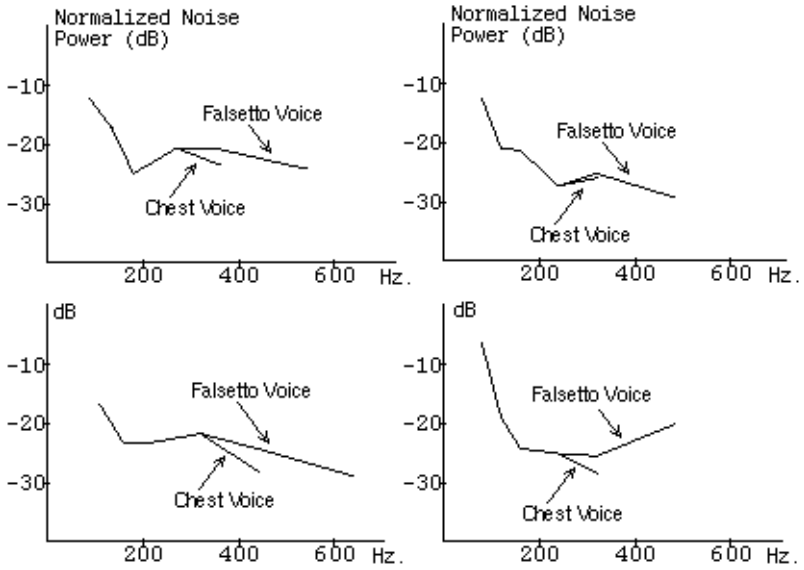

Figure 2.31: Average Normalized Noise Power (NNP) of male singers as a function of frequency. The curves corresponding to chest register and falsetto register are shown.

production in the glottis is that of a time-varying process which is dependent on flow and the area of the aperture, so it is likely that any increase in flow is being offset by changes in the time-varying area function. In the falsetto register there is a direct relationship between phonation frequency and flow [171], so there is a likelihood of higher noise power for increasing frequency in this range. All of the male falsetto test subjects showed an increase in noise power when entering the falsetto register, and some of the falsetto data exhibited an increase in noise power with increasing frequency.

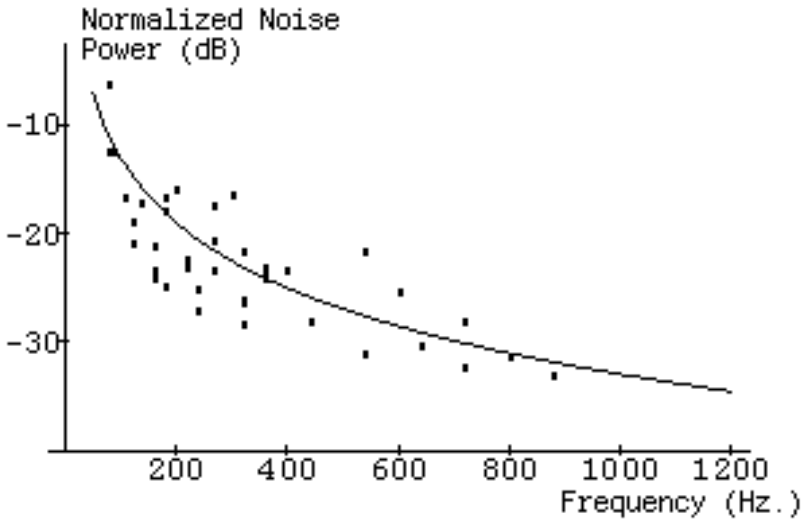

Figure 2.32: Average Normalized Noise Power (NNP) of eight additional singers as a function of frequency. The smooth line is a plot of $log\frac{k}{frequency^{1.2}}$.

2.10.2 Pulsed Noise in Singer Voices

A time domain analysis of pulsed noise was performed using the extracted noise residual periods from the four singer subjects. The period residuals were divided into six segments, with the first segment being that which contains the glottal closure epoch. The average noise power for each of the six sub-segments was computed across 200 periods. Figures 2.33 and 2.34 show the average noise power at six period positions for the four singers. The three superimposed graphs represent the three dynamic levels, and the five sets of curves represent the five notes sung. Since no clear relationship existed between time-domain noise power behavior and dynamic level, the three curves are not labeled separately. The waveform shown below the graphs is a typical waveform from the particular singer, aligned to show the point at which each of the average powers were computed within a typical period.

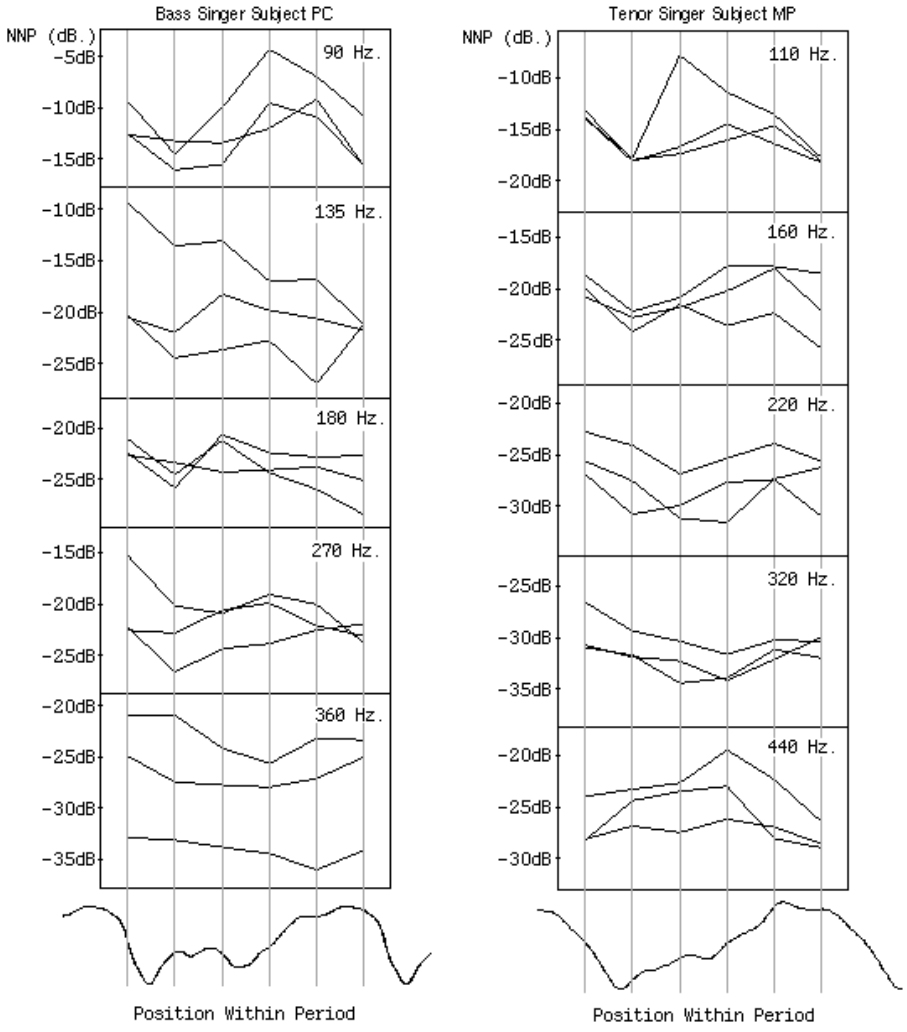

Figure 2.33: Average Normalized Noise Power (NNP) of bass singer PC and tenor singer MP as a function of the position within a typical period.

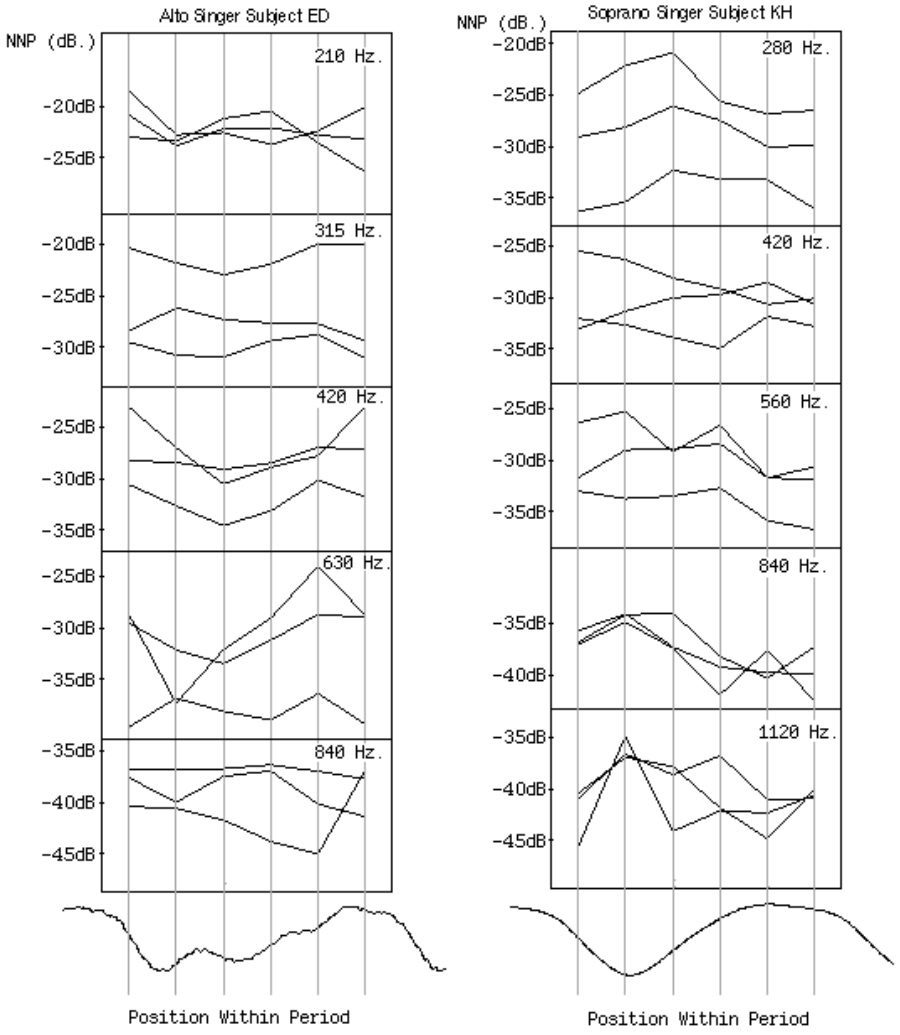

Figure 2.34: Average Normalized Noise Power (NNP) of alto singer ED and soprano singer KH as a function of the position within a typical period.

For low frequency phonation in the bass and tenor subjects, the curves show a dual-pulse nature. The primary pulse occurs at the initial glottal opening phase and the secondary pulse occurs at the glottal closure epoch. This is demonstrated in Figure 2.33 in the bass singer subject PC curves at 90, 180, and 270 Hz. The tenor subject MP exhibited dual-pulse activity in the 110, 160, and 220 Hz. curves of Figure 2.33. The alto subject ED exhibited dual-pulse activity in the 210 Hz. curves of Figure 2.34. The soprano singer subject KH exhibited a small modulation dual-pulse noise signal on two of the 420 Hz. curves of Figure 2.34.

As frequency increases, the time domain behavior shifts to a single-pulse nature, with one broad pulse of noise centered at the glottal open phase. This is shown in the 360 Hz. curves of the bass subject PC in Figure 2.33. The tenor subject MP demonstrated single-pulse activity in the 320 Hz. curves of Figure 2.33. The alto subject ED demonstrated single-pulse activity in the 315, 420, and 630 Hz. curves of 2.34.

Yet a third regime of pulsed noise generation was observed in the 430 Hz. curves of tenor subject MP in Figure 2.33. This same activity is seen in the 280, 560, 840, and 1120 Hz. curves of soprano subject KH in Figure 2.34. These curves show a single broad noise pulse centered at the glottal opening event. This is consistent with the fact that the glottal folds do not completely close in female vocal fold oscillation, and in male falsetto oscillation [21].

The NDR's averaged 5.35 dB, with a variance of 5.9 dB. The large variance reflects the wide range of NDR's encountered, with the maximum being 13.4 dB and the minimum being 1.4 dB. There was a weak inverse relationship of NDR upon pitch in the male singers, specifically a high NDR in the lowest notes of phonation. This is consistent with the predictions of turbulent behavior in the frequency regions above and below 200 Hz. from Section 2.5.

2.11 Subharmonics in the Singing Voice

Subharmonics are periodicities which occur at time intervals which are longer than the intended or perceived period of a quasi-periodic waveform. These can be thought of as undertones, similar in definition to musical overtones. The numbering system for overtones specifies that the first overtone is the fundamental, then successive integer multiples of the fundamental are indexed by the integer multiplication factor (the third overtone of 100 Hz.

is 300 Hz.). Using the overtone nomenclature, the third subharmonic (undertone) of 100 Hz. is 33.3 Hz. Subharmonics exhibit their own overtone series, thus causing sinusoidal peaks to appear between the 'actual' harmonics of the spectrum of a quasi-periodic signal.

In studies of bowed-string instruments, pulsed noise has been shown to be important in bow-string slip phase initiation, and plays a part in the generation of sub-harmonics [118][139][117]. The voice also exhibits measurable and audible sub-harmonics, and one plausible cause is the interaction of the glottal folds with reflected noise pulses. Diplophonia is a disorder of the voice which is characterized by the generation of subharmonics of such extreme amplitude that the pitch of phonation is obscured. The name comes from the fact that many diplophonic waveforms exhibit an extremely strong second subharmonic, thus yielding a pitch period twice the length of the intended pitch. Waveforms of this kind are of the class of pathological waveforms which confound most machine pitch detection schemes. A study of diplophonic patients [121] theorized that diplophonia was a beat phonomenon caused by the two vocal folds vibrating independently at different frequencies. Simulations of such independent fold vibration were performed which produced waveforms which resembled PhotoGlottoGraphic (PGG) waveforms obtained from the test subjects. This theory indicates that the assumptions of symmetry in most physical models of the glottis are not valid.

The generation of subharmonics, however, is not necessarily a pathological condition of the voice. In fact it is quite common in the trained 'resonant' voices of singers and actors. Figure 2.35 shows a clear subharmonic and its overtones in the time and frequency domain plots of the sung tone of a professional baritone soloist. The power of the subharmonic signal is 7 dB above the noise floor, and 20 dB below the 'periodic' component of the signal.

One common method used to detect diplophonia and subharmonics is to form an autocorrelation signal as defined in Equation 2.1. If the autocorrelation signal component corresponding to two periods of lag is larger than the component corresponding to one period of lag, the signal contains significant 2nd subharmonic components. Other methods of detecting subharmonics can be implemented in the frequency domain.

For this study, the residual from periodic prediction of Equation 2.35 was used to study subharmonics in normal singer voices. Noise was extracted using the periodic prediction method, first with the prediction period equal to the period intended by the singer, then

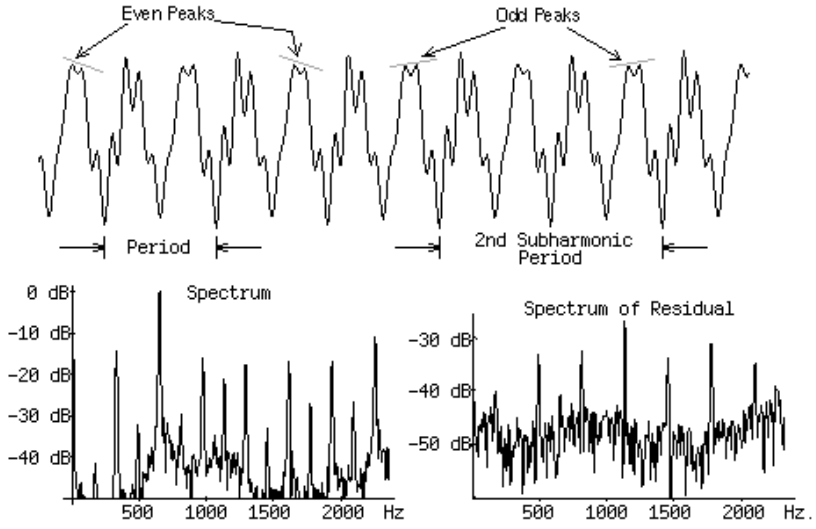

Figure 2.35: A bass singer voice waveform displaying clear subharmonics. The plot at the lower left is the frequency spectrum of the waveform. The plot at the lower right is the frequency spectrum of the residual from periodic prediction showing the subharmonic components only.

with a prediction period twice that of the intended period. In a perfectly periodic signal corrupted by a white noise process, any delay length which is an integer multiple of the actual period yields the same prediction results. For a quasi-periodic signal with no subharmonics, a two-period predictor will usually perform worse than the single-period predictor, because in such a signal pairs of adjacent periods are usually more similar than periods which are separated by more than one period. In a signal which contains second subharmonic components, the periodic predictor with doubled period accurately predicts any multiple of the 2nd sub-harmonic, and thus yields an error signal free of this subharmonic component. If significant second sub-harmonic components are present in a signal, the NNP should be less than that yielded by prediction of the signal using the actual fundamental period.

Four singers were recorded at three volume levels and five pitches. Eight singers were recorded at medium volume and five pitches. Periodic prediction was performed at one period and two periods of delay. Normalized Noise Power was computed for the residual signals. If the NNP was greater for the single lag predictor residual than for the double lag predictor residual, the tone was judged to contain a significant second subharmonic

component.

All of the 12 singers exhibited subharmonics in one or more sung tones, and 28 of the total 100 vocal samples showed a measurable component of the second subharmonic. The largest detectible subharmonic referenced to the residual noise was 8.9 dB above the glottal noise signal. The largest detectible subharmonic referenced to the periodic component of the glottal wave was 17.9 dB below the periodic glottal component. The male subjects were more likely to produce subharmonics than the female subjects, with the males displaying 2nd subharmonic components in 32 percent and the females in 24 percent of the tones. Soft phonation (musical pianissimo) exhibited subharmonics only in one subject at high frequencies. Medium and loud phonation exhibited subharmonics in all subjects at higher phonation frequencies. Table 2.2 summarizes the results of this analysis. A zero indicates that there was no detectible subharmonic, and a number represents the subharmonic power, referenced to the residual signal with the subharmonic removed.

To model subharmonics using wavetable synthesis methods, the wavetable is filled with the sub-harmonic period of the glottal wave. To represent a second subharmonic component, the wavetable length is doubled and two 'periods' (one period of the second subharmonic) are stored.

2.12 Use of Noise Residual for Vocal Tract Filter Identification

One common problem with LPC or other source/filter analysis methods is presented when attempting to analyze female vocal tones [22]. It is common to encounter sung tones in which the fundamental lies above the location of the first formant. Since LPC is a least squares minimization technique, the filter spectrum is fit to the harmonic peaks, ignoring any spectral information lying between the harmonics. Methods of identifying the underlying formant envelope using the trajectories of the harmonics were proposed previously [97] [69]. Assuming the residual component is generated near the glottal source, clues to the underlying vocal tract resonance curve are contained in the residual spectrum which is ignored by LPC. If the harmonics of the periodic component can be extracted from the spectrum of the sung tone without disturbing the residual spectrum, LPC can be applied

Bass Subject PC				*Tenor Subject MP*			
Freq.	pianissimo	mezzo forte	forte	Freq.	pianissimo	mezzo forte	forte
90	0	0	0	110	0	0	0
133	0	1.6	0	165	0	0	0
180	0	2.5	3.0	220	0	1.2	0
270	0	0.2	0	330	8.9	0	0
360	0	0.9	0.2	440	2.6	0	5.8

Alto Subject ED				*Soprano Subject KH*			
Freq.	pianissimo	mezzo forte	forte	Freq.	pianissimo	mezzo forte	forte
220	0	0	0	290	0	0	0
315	0	0.7	0	440	0	0	0
440	0	0	0	580	0	0	0
630	0	0	0	880	0	0	0
880	0	0	0.1	1160	0	1.4	0

Bass WR		*Bass WB*		*Ten. RC*		*Ten. AB*	
Freq.	mf	Freq.	mf	Freq.	mf	Freq.	mf
80	0	82	0	90	0	110	0
120	0	122	0	135	0	165	0
160	0	165	0	180	0	220	0.1
240	0	245	0	270	1.0	330	1.4
320	4.9	330	1.0	360	0	440	0.6

Alto LU		*Alto AD*		*Sop. KB*		*Sop. CC*	
Freq.	mf	Freq.	mf	Freq.	mf	Freq.	mf
180	2.4	180	0	200	0	220	0
26	3.2	275	0.9	300	0.8	330	0
360	1.0	360	0	400	0	440	0
525	0.4	550	0	600	0.8	660	1.4
720	0	720	0	800	0	880	1.9

Table 2.2: Data from detection of 2nd subharmonic in 12 singer voices. A zero indicates that no subharmonic component was found.

to the residual. This is similar to the whisper method of vocal tract transfer function identification described in Section 1.7.4.

Figure 2.36 shows a synthetic vocal tract spectrum, the spectrum of a vocal tone with noise added to the glottal source, and residual spectra obtained by periodic prediction and period similarity processing. The smooth curves on the lower three plots are LPC spectra, and the formants are indicated to the right of each curve. The LPC spectra of the two residual signals both detected the formant near 400 Hz. All three LPC analyses missed the tightly grouped second and third formants, although these are visibly evident in the periodic prediction spectrum.

2.13 Pulsed Noise in Other Musical Systems

The pulsed noise extraction and analysis techniques were used in another study [119] to analyze musical instrument tones. In the case of bowed strings, the sliding of the bow against the string during the slip phase of oscillation causes friction, and thus noise is both radiated and introduced into the string [118]. Pulsed noise has been shown to be important in bow-string slip phase initiation, and to play a part in the generation of sub-harmonics in stringed instruments [139]. Noise extraction was performed on a cello tone of 150 Hz using the SANSY system [160]. Figure 2.37 is a time domain plot of the magnitude of the resynthesized residual, and clearly shows the pulsed noise bursts.

The case of wind-driven instruments is similar to that of the glottis. In the reed family, returning noise pulses interact with the generation of future noise pulses, causing correlation between successive periods of noise. In this case the period similarity method yields better results, as this method is less sensitive than periodic prediction to period-to-period correlations which decay with time. The interaction of noise pulses with the oscillator at each period is less clear in the case of the voice, where the oscillator is weakly loaded by the vocal tract and the tube length does not determine the frequency of oscillation. Noise extraction was performed on clarinet signals using the period prediction method. Two tones of approximately 200 Hz were analyzed; one played loudly with a soft reed, and the other played softly with a stiff reed. Figure 2.38 shows the power surfaces and spectral periods. The soft reed flexes greatly in oscillation, yielding a two pulse noise surface consistent with a large open aperture phase. The stiff reed barely interrupts the air flow in a softly blown

tone, and thus the noise power surface is flatter (breathier), and exhibits only a single ridge. This is consistent with a single noise pulse as the reed constricts the aperture. The NDR of the soft reed tone was 15.28 dB, and the NDR of the stiff reed tone was 2.47 dB.

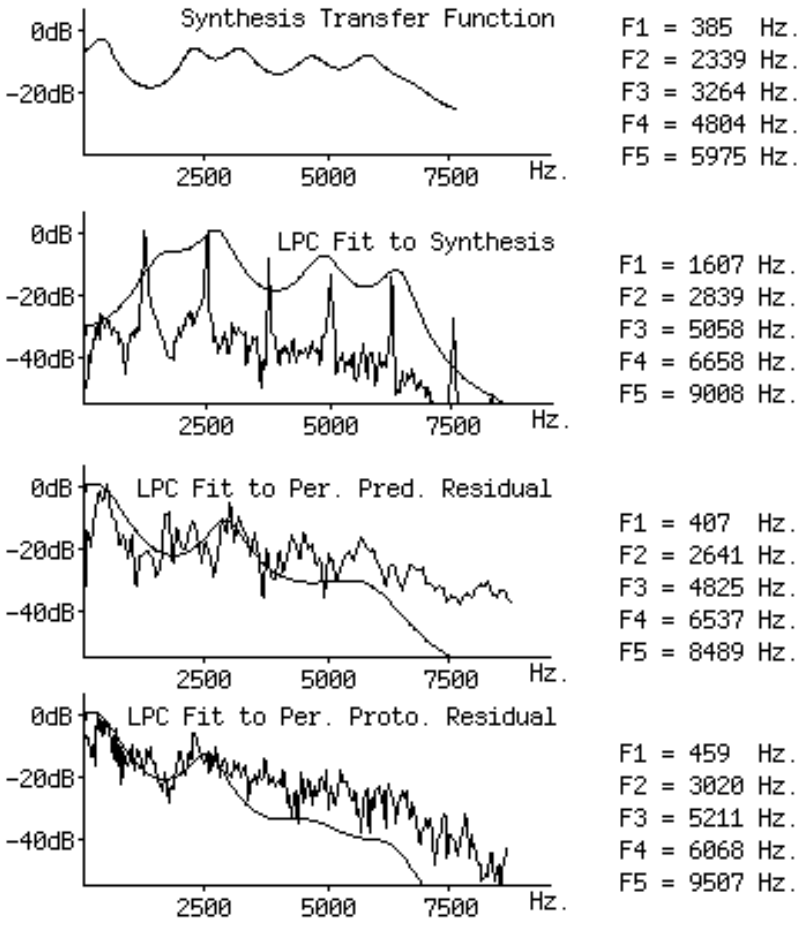

Figure 2.36: Top to Bottom: Vocal tract response used for synthesis, LPC spectral fit to synthesized tone, LPC spectral fit to residual obtained by periodic prediction, LPC spectral fit to residual obtained by period similarity processing. Formant frequencies are noted to the right of each spectrum.

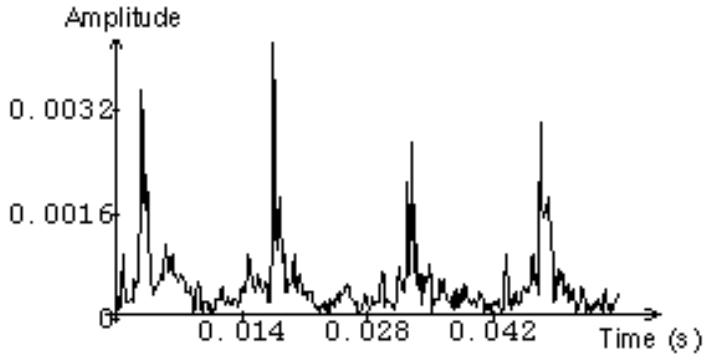

Figure 2.37: Magnitude of residual signal of bowed cello tone shows clear noise bursts.

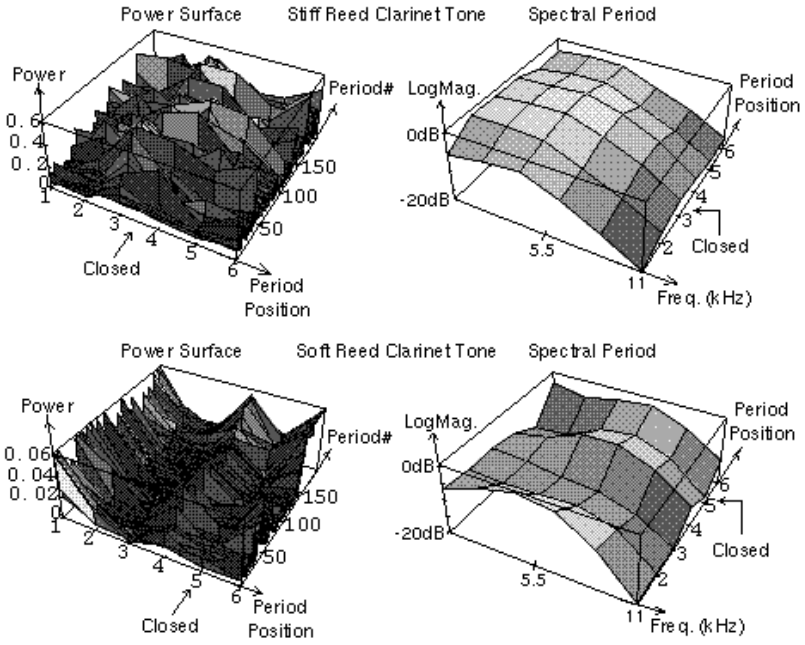

Figure 2.38: Power surfaces (left) and spectral periods (right) of clarinet tones generated with stiff reed (above) and soft reed (below).

Chapter 3

Software Systems for Singing Synthesis

Two voice synthesis systems were constructed using the waveguide multiple acoustic tube model of the human vocal tract. One system is a real-time Digital Signal Processor (DSP) interface program, which allows graphical interactive experimentation with the various control parameters. The other system is a text-driven software synthesis program. The vocal tract is modeled in both systems by a digital Waveguide Filter (WGF) network, controlled directly by shape parameters. A nasal tract WGF is coupled to the vocal tract at the velum bifurcation point. Glottal source pulses are stored and retrieved from multiple wavetables. A filtered pulsed noise component is added to the periodic glottal source, simulating the turbulence generated as air flows through the oscillating vocal folds. To simulate the turbulences of fricatives and other consonants, a filtered noise source can be made arbitrarily resonant at two frequencies and placed at any point within the vocal tract.

The real-time DSP program is called SPASM (Singing Physical Articulatory Synthesis Model). The vocal tract shape is graphically displayed by a cross section of a human head. Sliders on an editor window control the radius of each vocal tract segment, the size of the velum opening into the nasal tract, and the radius of each nasal tract segment. A Formant Editor Window displays the log-magnitude frequency response of the tract. The glottal pulse shape is edited in the time domain, and the spectrum is edited in the frequency

domain. Other real-time controls allow experimentation with pitch and vibrato. The system can interactively record a consonant and design a matching filter for use in resynthesis. Similarly, the user can record a vowel, and the periodic glottal waveform and noise parameters are identified by the system and used for resynthesis. All control parameters can be saved as disk files.

The software synthesis system is called "singer," and takes as input a file of C function calls specifying the events to be synthesized. These function calls are a time-ordered event list for controlling the singer model. An event specification includes a transition time, shape and glottal files as created by the SPASM system, noise and glottal volumes, glottal frequency (either in Hz. or as a musical note name), and vibrato amount. The system synthesizes a sound file, smoothly interpolating from each set of parameters to the next over the times specified. All parameters of shape, glottal input, and noise filter control are interpolated on the single sample level. In this way smoothly varying connected singing performances are generated. All other parameters, such as random vibrato amount and periodic vibrato speed may be changed at any time but, for computational speed, are not interpolated.

3.1 The Synthesis Model

Figure 3.1 shows a block diagram of the model constructed for singing synthesis. A variety of sound sources are injected into the WGF acoustic tube model of the vocal tract. All waveform oscillators may be loaded with arbitrary waveforms.

Two glottal wavetables are provided to allow slow variations in the source under explicit control, or vibrato-synchronous variations. The glottal noise source consists of four-pole filtered white noise, multiplied by an arbitrary time domain waveshape synchronized to the glottal oscillators. This allows pulsed noise to be simulated and mixed with the periodic glottal source. Vibrato is simulated by a wavetable oscillator (sine default), mixed with four-pole filtered white noise.

Four-pole filtered white noise is injected into the oropharyngeal WGF by mixing with the forward-going wave component. The noise can be injected into any number of sections, as controlled by independent gain controls.

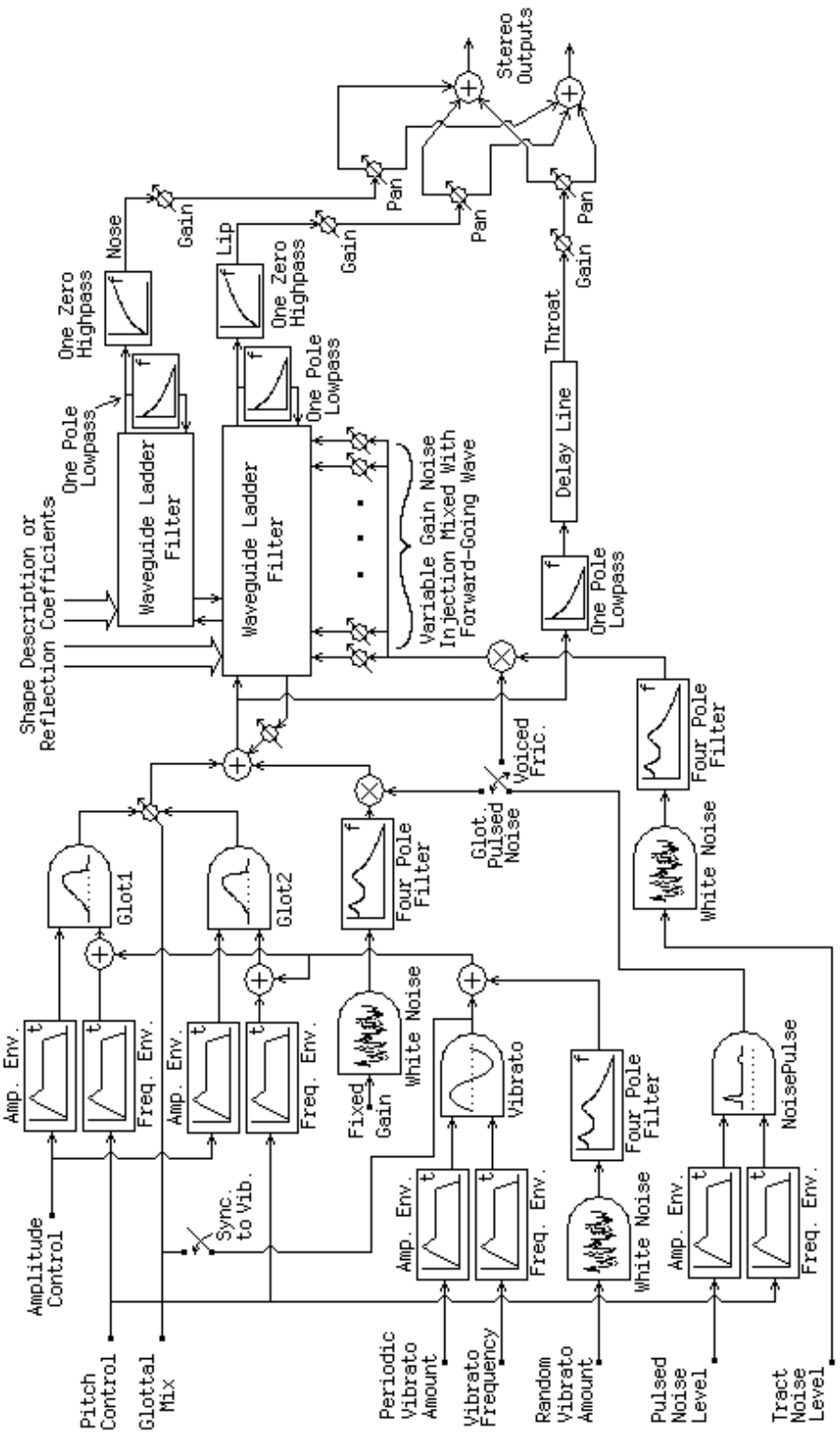

Figure 3.1: Block diagram of the model used for singing voice synthesis.

The mixed glottal source signal is injected into the vocal tract filter. The glottal reflection is modeled by a simple reflection coefficient, and the lip and nostril conditions are modeled by a one-pole low-pass filter for the reflection characteristic, and a one-zero filter for the transmission characteristic. The transcutaneous radiation component is modeled by a one-pole low-pass filter and a delay line. All outputs have independent gain and stereo pan controls.

3.2 The SPASM System

3.2.1 Design Goals

The SPASM (Singing Physical Articulatory Synthesis Model) system was developed to place the waveguide vocal tract model into a graphical interactive environment for experimentation, synthesis, and library construction. The following design goals were used:

1. The primary goal of the system is to produce high quality musical vocal synthesis.

2. Some knowledge of music and musical acoustics on the part of the target user is assumed, but no knowledge of digital filter design or other engineering topics should be necessary to use the system for synthesis experiments.

3. Technically trained or experienced users should be able to access advanced features.

4. Selection of model characteristics should be made so that, wherever possible, control parameters are physically meaningful. Guidelines 1. and 2. take precedence over this requirement.

3.2.2 The System Screen

Figure 3.2 shows the main system screen. The various windows allow the user to modify the parameters controlling the model. Displays show shape, time, and spectral descriptions of the model and signals. The windows visible when the program is first run are those required for a beginner to do initial synthesis experiments. Figure 3.2a is the Vocal Tract

shape Editor window, which controls and displays the shape of the vocal tract. Figure 3.2b is the Glottal Excitation Editor window, allowing time and frequency domain control of the glottal source, and saving glottal description files to disk. Figure 3.2c is the Noise (turbulence) Generator controller, which controls placement of the noise source within the tract, the gain of the injected noise, and provides access to a more elaborate editor for the noise source. Figure 3.2d is the Phoneme Synthesis and Library window, where shapes are tested with short synthesis examples, and vocal tract description files are saved to disk. Figure 3.2e is the Performance Feature Editor window, allowing parameters affecting pitch to be controlled and saved to disk. Figure 3.2f is the Diphone Synthesis and Library window, which permits transitions between shape and glottal states to be specified, auditioned, and saved to disk. Hidden windows can be called up for more advanced control and analysis functions.

3.2.3 Vocal Tract Shape

The Vocal Tract Editor provides control over the shape of the acoustic tube (and thus the digital filter) which models the vocal tract. Shapes are saved to or loaded from disk files. Sliders in the graphical editor window control the radius of each segment of the tract. The path through the nasal airway is controlled by a velum position slider. A graphical cross-section of a human head provides immediate feedback to the user about the vocal tract shape. An additional text window showing the radii in centimeters allows the user to enter parameters with greater accuracy. Another window allows the editing of the nasal tract shape parameters, although these characteristics are usually not varied in connected speech and singing. Switches and sliders control and mix the lip, nose, and throat radiation outputs.

One other tract shape control window is the Shape Space Interpolator. This window allows the user to enter a number of shape filenames into text fields. Each point along the edge of the round shape space control area represents a region dominated by a particular library shape. The user may control the current vocal tract parameters by moving a cursor about the control area, thus determining the "mix" of shapes. This control is particularly useful in the real time DSP synthesis mode, discussed in Section 3.2.9. Figure 3.3 shows the vocal tract shape control windows.

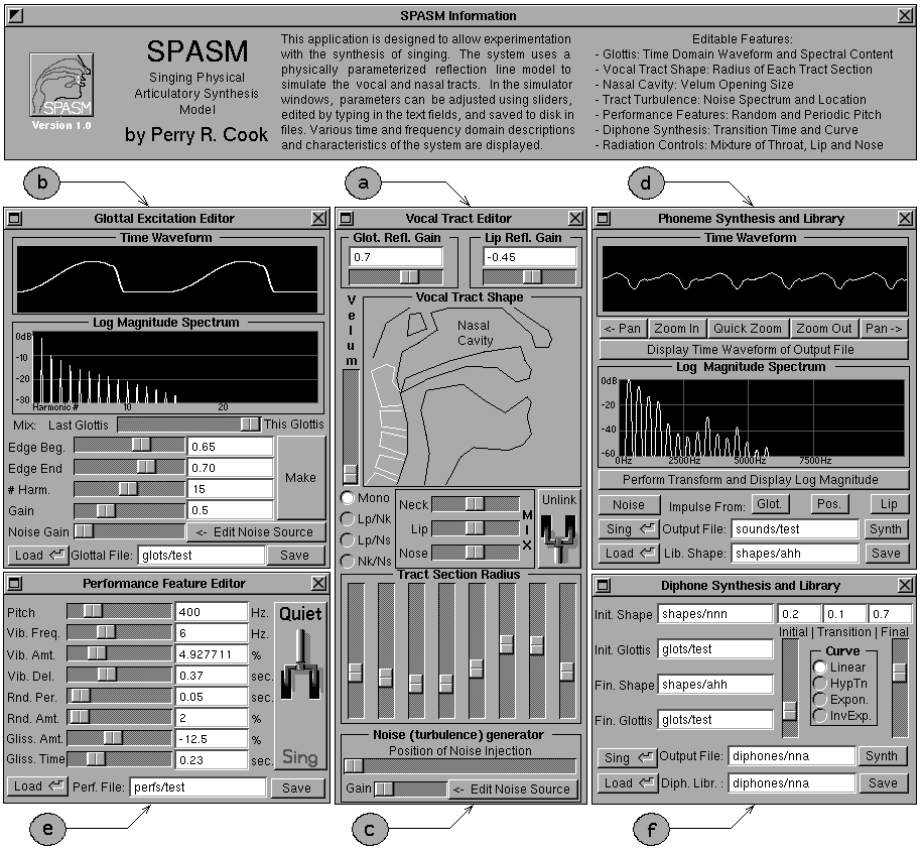

Figure 3.2: The initial SPASM screen, showing the windows which open upon running the program.

3.2.4 The Glottal Source

The glottal source used in this system copies the time-domain and spectral properties of the pressure, velocity, or power waveform of the human glottis. The glottal source waveform is additively synthesized from Fourier coefficients controlled by simple parameters entered in the editor, or from a library file of coefficients derived from analysis data. The simple parameter editor controls operate principally on the time-domain glottal waveform. Parameters include the number of harmonics used for synthesis (primarily to prevent aliasing), the overall amplitude, and the position and slope of the falling edge of the glottal pulse, which has been shown to be an important feature when describing vocal effort (Sundberg 1987;

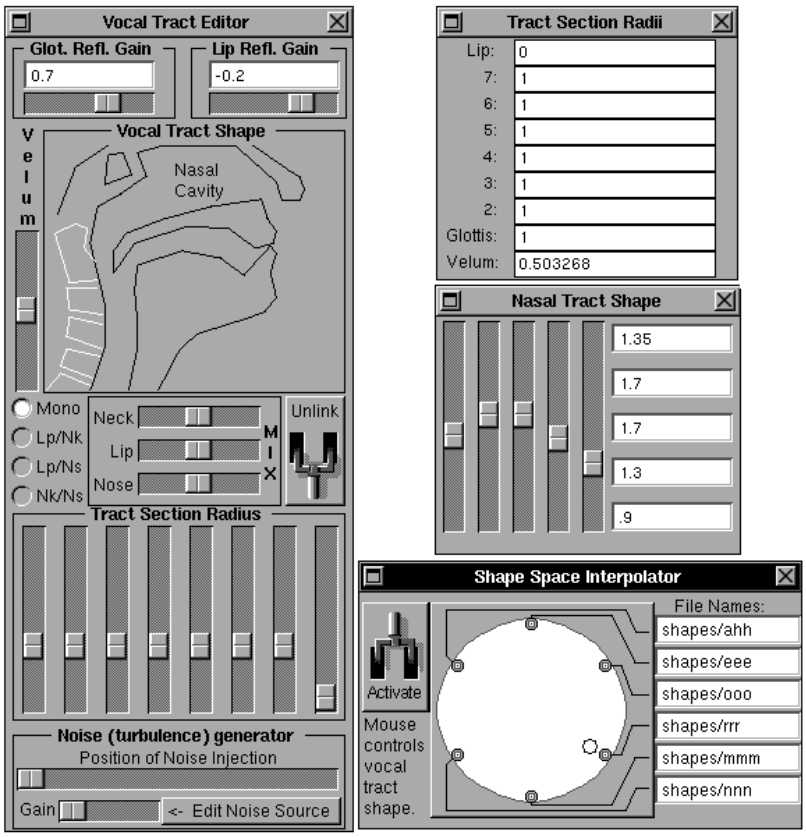

Figure 3.3: Windows for controlling vocal tract and nasal tract shape.

Cummings 1990; Rosenberg 1971). Graphical displays of the log-magnitude spectrum and the time-domain waveform are provided. Parameters may be saved to and loaded from disk files. The Glottal Noise Editor allows the user to specify a time-domain pulse shape for an additive noise source, simulating the pulsating noise generated as flow through the glottal folds is interrupted. A default set of natural amplitude and frequency control functions are available for synthesis, or the user can edit the performance features using sliders or text fields. Performance controls affect frequency features and are located in the Performance Parameters window. Figure 3.4 shows the glottal source editing and analysis windows, and performance controller window.

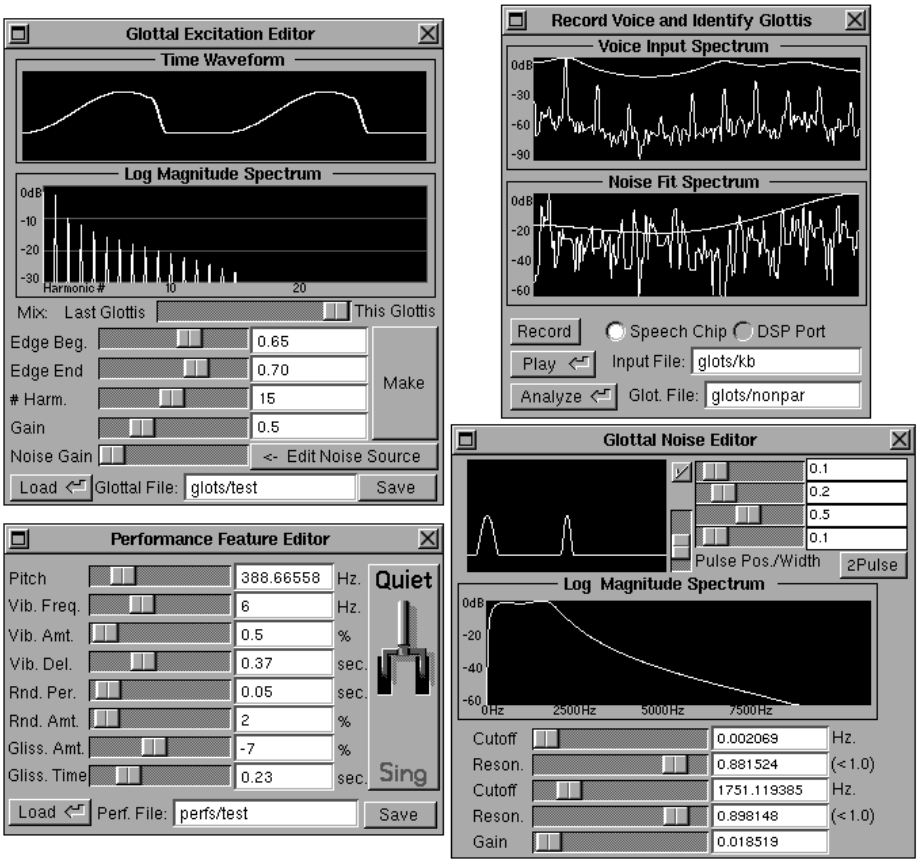

Figure 3.4: Windows for controlling and identifying the glottal pulse source.

3.2.5 The Noise Source

To simulate the turbulences of fricatives and other consonants, a noise source can be placed at any point in the oropharyngeal path of the vocal tract. The output of the noise source can be made arbitrarily resonant at two frequencies by a four-pole filter. This allows for injection of a tuned source of local turbulent noise at a point of constriction. Noise source parameters are saved as part of shape files. Figure 3.5 shows the noise controller windows.

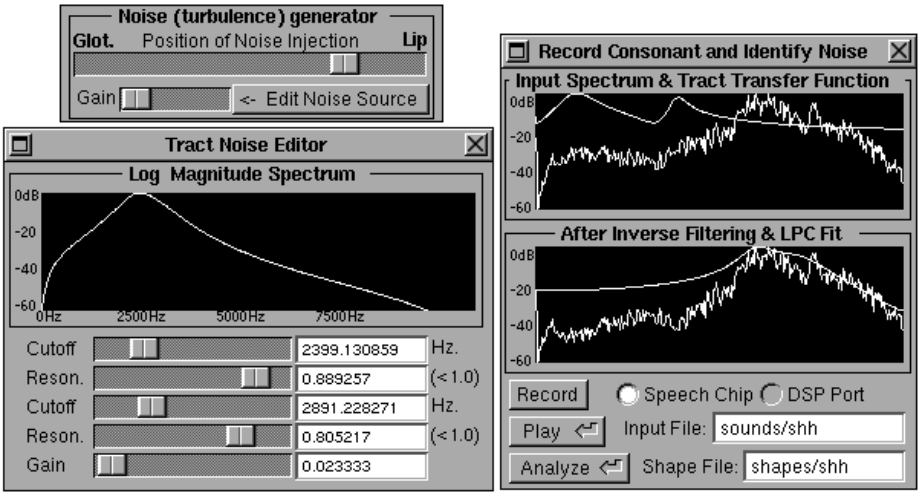

Figure 3.5: Windows for controlling and identifying the turbulent noise source.

3.2.6 Glottal Pulse and Noise Source Identification

To aid the user in obtaining natural source parameters quickly, and to provide frameworks for both rapid detailed study of glottal and noise sources, there are various identification/analysis functions built into the system. The Record Consonant and Identify Noise window allows the user to record a fricative consonant, or specify a prerecorded sound file. Using LPC, the system then designs a four-pole filter which matches the spectrum of the recorded sound. Similarly, the Record Voice and Identify Glottis (Figure 3.4) window allows the user to inverse filter a recorded voice, using the current SPASM vocal tract configuration as the prototype. The inverse filtering process yields an estimate of a glottal pulse function.

An interactive inverse filtering window is available for more accurate identification of vocal tract transfer function and glottal input function. The user can fit an LPC filter to a sound, and multiple representations of the filter are available. The coefficients of the filter are displayed, or alternatively, the center frequencies and radius locations of the poles are displayed. The poles of the filter are located and displayed on a Z plane view. Mapping of filter parameters onto vocal tract shape description quantities (radii or areas) is available. The user may edit all filter representations. Using this tool, interactive inverse filtering of the vocal tract transfer function is accomplished. Figure 3.6 shows the Interactive Filtering

Workspace.

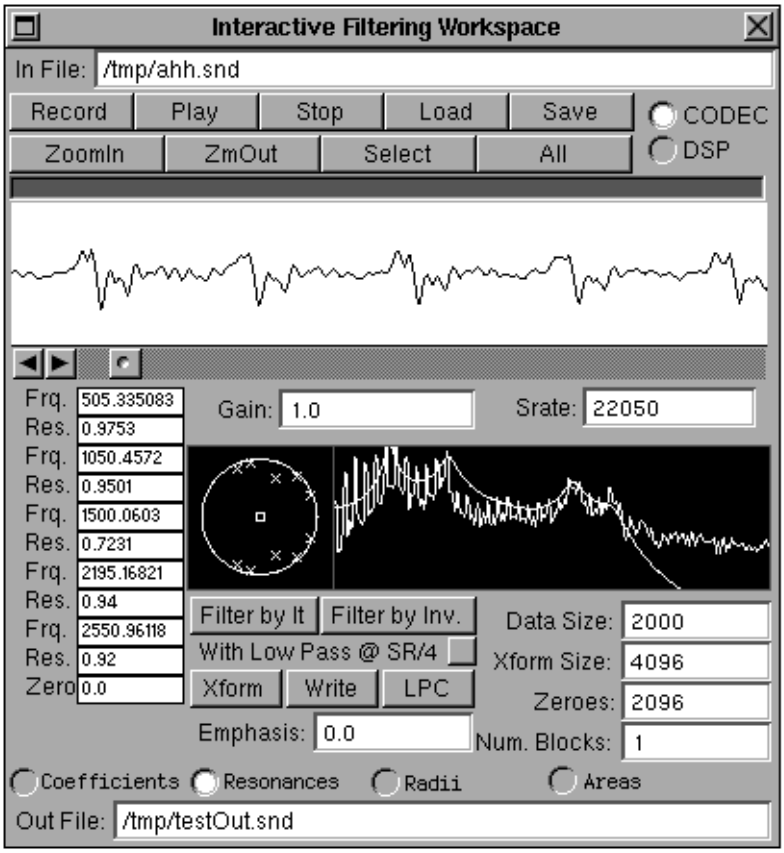

Figure 3.6: An interactive filter editor, with controls for fitting, editing, and applying filters to sounds.

3.2.7 Phoneme and Diphone Synthesis

Once the glottal source, the noise source, and the vocal tract shape are established, synthesis of a short musical 'performance' is accomplished by mouse clicking the synthesis button, located in the Phoneme Synthesis and Library window (Figure 3.7). Once the synthesis is complete, the result is heard via the computer's internal digital to analog converters. The file may be played back repeatedly by mouse clicking the Sing button. By typing any file name into the Output File field and pressing the carriage return, the file is played

back. Synthesis of different files allows speedy A/B comparison of sound examples. The time domain waveform and the log magnitude spectrum of the synthesized phoneme can be displayed.

Diphthongs are constructed by specifying initial and final sets of parameters, an interpolation curve, and the time in seconds of the initial and final steady state segments. The Diphone Synthesis and Library window (Figure 3.7) controls these synthesis parameters. Since the synthesis yields one second of sound, specifying the duration of the initial and final states also specifies the transition time. Curves available for interpolation include linear, hyperbolic tangent, and exponentials. In the case of the glottis, the interpolation is carried out between initial and final wave tables. For controlling the noise generators, the filter pole parameters are interpolated in the Z-plane as radius and angle quantities. The vocal tract and nasal tract scattering relationships are interpolated in the radius space.

3.2.8 Formant Editor and Display

The Formant Editor window (Figure 3.7) allows the user to display and edit features of the vocal tract filter in the formant domain (the one in which the ear perceives speech sounds). When this window is activated, the system impulse response is obtained. A log-magnitude transform is computed and displayed, and peaks are located and marked. Each of the first few (selectable) peaks is associated with a text-field/slider control, and the user may move the markers to new locations. By depressing the Doit button, the system adaptively moves the formant peaks to the desired locations, modifying the vocal tract in a least squares perturbation fashion. When the Link button is active, the formant display updates each time any change is made in vocal tract control parameters. This allows the connection between vocal tract controls and vocal tract filter spectrum to be interactively explored.

3.2.9 Real Time DSP Synthesis

By clicking the Sing switch in the Performance Features window (Figure 3.4), the system uses a Digital Signal Processing (DSP) chip to synthesize in real time. Performance features active in real time are pitch, vibrato speed and amount, and random vibrato amount. Tract section, velum, and noise parameter controls modify the model and sound in real time. New glottis wavetables may be synthesized and down-loaded to the DSP chip. A cross fader

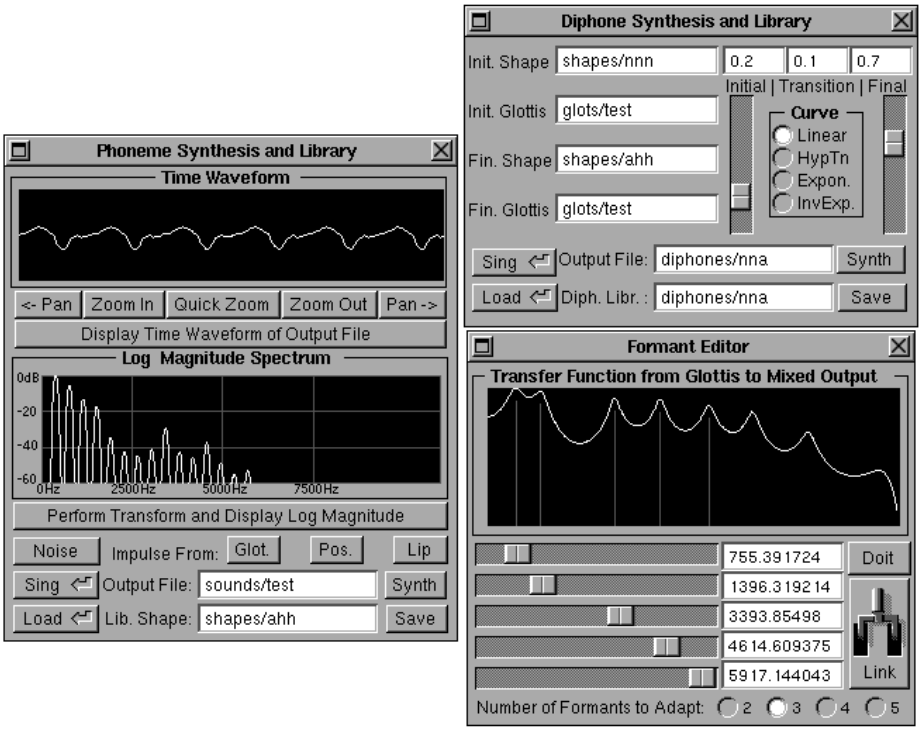

Figure 3.7: Phoneme and Diphone synthesis control windows, and the Formant Editor/display window.

controls the mix between the current glottal waveform and the last loaded glottal wave.

3.2.10 Object Oriented Programming Structure

The techniques of Object Oriented Programming Systems (OOPS) are becoming increasingly popular for the development of software [183][188]. Because of the notions of heirarchical inheritance [184] and abstract data types [186][190], program code generated in such systems is extendible and reusable [187][189]. The paradigm of object oriented programming languages is that of *Instances* of *Classes* (objects) passing *Messages* (function calls and data) to each other. The *Class Heirarchy* and *Class Definitions* define the behaviors of and relationships between different types of objects. The *Class Heirarchy* is the family tree followed to determine how a particular object behaves. Redundant coding is reduced

by the fact that objects *Inherit Instance Variables* (variables which are objects themselves, whose values are known only to the particular instance of an object) and *Methods* (pieces of executable code) from their ancestors. Only behaviors unique to the object need to be defined. All others are inherited from the ancestor. Any behavior significantly different from that of the ancestor may be *Overridden* by simple redefinition. OOPS techniques allow rapid prototyping of software and the construction of libraries which may then be shared among many developers of similar systems.

The SPASM system was developed in the Objective C [182] object-oriented programming extensions to the C programming language [185]. The important Class Definitions of the SPASM program are given in Appendix B. Figure 3.8 shows a block diagram of some the objects in the SPASM program, and the type of information that is passed between objects. Such an organization of the code allows for flexible modification of the synthesis model, controls, and user interface.

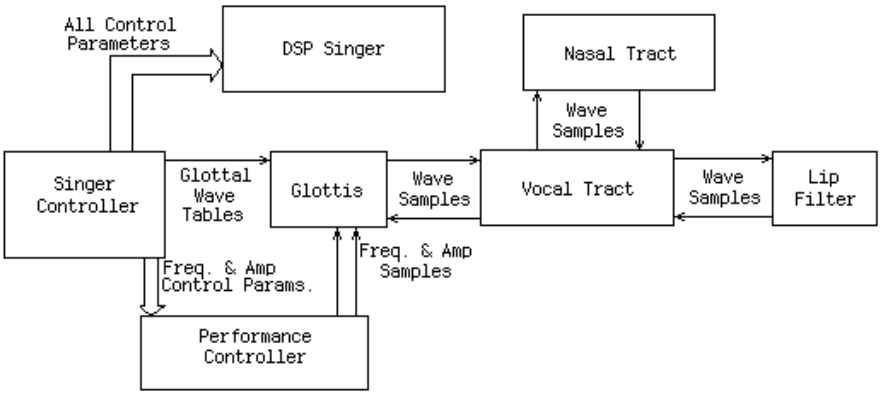

Figure 3.8: Some of the objects in the SPASM program, and the type of information that is passed between objects.

Design of the glottis is an example of the flexible programming features afforded by OOPS. In the current implementation of the model, the Glottis object is a waveform synthesizer which generates the glottal wave under the control of a Performance Controller object. The Performance Controller object generates frequency and amplitude information and passes it to the Glottis object. The Glottis object generates a sample by combining some of the wave variable coming into the superglottal area from the Vocal Tract object with the value

from a wavetable. The wavetable is an instance variable of the Glottis object. This sample is passed to the Vocal Tract object as its current input. If a physical model of the glottis were constructed, the only code that would be edited is that of the Glottis object, replacing wavetable lookup with mass-spring oscillator calculations. The glottis would derive the necessary local control variables of mass values and spring constants from the frequency and amplitude values being passed from the Performance Controller object. Such a design paradigm allows pieces of the voice system model to be isolated and refined individually, or perhaps assigned to different programmers for refinement. By keeping the interface to all glottal models the same, various models of the glottis could be compared rapidly, without requiring any changes to objects which interact with the glottis.

3.3 The Singer Software Synthesis System

For many applications, accurate control of synthesis at the single sample level is desired. Repeatable synthesis is important for the generation of sounds for psychoacoustic testing and other investigations. Toward these ends, a software synthesis system based on the WGF model of the vocal tract was constructed. The program is written entirely in the ANSII C programming language [185], simplifying porting to any computer. The program takes as input a text file containing a series of C function calls. The C functions specify target values for the model parameters, and times for the transitions to take place. Some arguments to the functions are the names of shape and glottal files as created using the SPASM system. Others are floating point numbers specifying pitch, vibrato, and other important performance controls. For speed of synthesis, linear interpolation is performed on all parameters.

The *synthesize* function is the heart of the singer program. This function interpolates the singer model from the last parameter values to a new set of parameter arguments, synthesizing and writing the samples to a sound file.

The arguments are:

time	is the time over which the current transition will take place.
FShape	is a vocal tract shape file as created by the SPASM system.
FGlot	is a glottal file as created by the SPASM system.

FPitch	is the average pitch in Hz of the glottal source.
	Note names (af4 instead of 415.3 Hz.) may be used.
FGlotAmp	is the amplitude of the glottal source component.
FNoiseAmp	is the amplitude of the fricative noise component.
FVibrAmt	is the amount of periodic vibrato of the glottal source.
	1.0 is 100% vibrato.
fd	is a file descriptor of an open sound file.

One other synthesis function synthesizes silence.

silence(time,fd)

Some non-interpolated parameters may be changed instantly by function calls.

setPerfVibrFreq(float afreq); Vibrato frequency in Hz.

setPerfRndAmt(float r); Random vibrato amount.

setGlotNoiseGain(float gain); Glottal noise gain.

setPulseShape(float p1pos,float p1width,float p2pos,
float p2width,float p2height,float pfloor); Glottal noise pulse shape.

setGlotNoiseFilter(float gnfreq1,float gnrad1,
float gnfreq2,float gnrad2); Glottal noise filter characteristics.

Figure 3.9 shows a typical Singer control file. The shape "eeesh" is the shape "eee" with the noise generator placed as it would be in the case of shape "shh". This prevents the noise source from moving during this transition.

```
singer(fd)
     int fd;
{//                  This example sings the name "Shiela" with a crescendo at the end

     set_file_path("/local/prc/Library/SPASM/");          // file path to shape
                                                          //       and glot directories
     init();                                              // initialize state variables
     setup("shh","soft",400.0,0.0,0.0,0.0,0.0);           // Initial setup for
                                                          //       performance

//                   time     shape     glot.    freq.    glotAmp   noiseAmp    %vibr.
     synthesize(  0.3,      "shh",    "soft",   400.0,     0.0,       0.3,        0.00,     fd);
     synthesize(  0.1,     "eeesh",   "soft",   430.0,     0.2,       0.3,        0.04,     fd);
     synthesize(  0.7,     "eeesh",   "soft",   a4    ,    0.2,       0.0,        0.07,     fd);
     synthesize(  0.2,      "lll",    "soft",   440.0,     0.4,       0.0,        0.04,     fd);
     synthesize(  0.2,      "ahh",    "soft",   400.0,     0.3,       0.0,        0.00,     fd);
     synthesize(  0.2,      "ahh",    "soft",   400.0,     0.3,       0.0,        0.00,     fd);
     synthesize(  1.5,      "ahh",    "loud",   400.0,     1.0,       0.0,        0.08,     fd);
     synthesize(  0.1,      "ahh",    "soft",   400.0,     0.0,       0.0,        0.08,     fd);

     silence(0.5,fd);                                     // Write some silence
     return;
}
```

Figure 3.9: Singer command file to synthesize a sung performance of the name "Shiela".

Chapter 4

Conclusions and Suggestions for Future Research

4.1 Conclusions

The approach in the research presented in this dissertation has been to view the human vocal mechanism as a time varying linear system. A simple linear model was developed, and investigations were conducted to determine which controls provide the greatest flexibility, which features are the most important perceptually, and how the important features and controls can be added to the simple linear model. Taking the viewpoint that the vocal system is a time varying linear system, rather than a non-linear system, allows standard and proven techniques of linear system analysis to be employed in obtaining values for the control parameters of the model.

A new algorithm for tracking speech which directly drives the articulatory vocal tract model was presented. This algorithm was investigated, and benefits of its use were discussed.

The pitch deviation component in the vocal signal is an extremely important perceptual feature, without which vocal synthesis sounds machine-like. A new algorithm for tracking pitch was presented, tested, and used in a study of singer pitch deviation. Rules for low and high-frequency pitch deviation components were derived from the experimental data.

The phenomenon of noise generation near the glottal source was investigated. New methods

of extracting, analyzing, and visualizing the non-periodic component of the vocal signal were presented, and a study of source noise in singer voices was conducted. It was discovered that there is a significant time-domain structure to the noise present in the glottal source, consistent with a hypothesis of pulsed turbulence derived from a fluid-dynamics analysis of glottal source behavior. Rules for additive synthesis of this pulsed noise component were derived from the experimental data.

A number of examples of singing synthesis have been done using the SPASM and singer programs. The combination of real-time DSP control and software synthesis allows the user to quickly experiment with the model, yet produce repeatable results. Many synthesis attempts yield extremely natural sounds on the first attempt. Since descriptions of unnatural sounding synthesized sounds often rely on physical references ("She sounds like her jaw is open too far", or "His tongue sounds fat"), the physical parameters indicate what to do to the model controls if the synthesis does not sound correct. The programs have been made available to composers for use in musical compositions, and to psychologists for use in the generation of stimuli for psychoacoustic testing.

4.2 Suggestions for Future Research

Of the topics which were investigated in this dissertation, the two greatest areas for future research are those of articulatory speech tracking, and noise in the glottal source. Section 1.7.3 discusses various areas for future research in articulatory tracking. These topics will be briefly listed here:

- Investigation of the norms used for vocal tract adaptation and identification. Norms other than least-squares should be investigated, as well as other schemes which attach penalties and weightings based on physical constraints of the human vocal tract.

- Vector quantization of vocal tract shapes. This reduces search complexity and memory usage.

- Library construction by selection and ordering of shapes which best fit the perceptual boundaries of phonemes and diphones.

- Use of general and specialized hardware, and optimization to bring the system to real-time capability.

The topic of noise generation in the vocal mechanism is not new, but seems to be entering a new era. Until recently, most research into vocal noise has been conducted in the areas of vocal pathology, concentrating on abnormal voices. Realizations that the noise component in normal voices is an important perceptual component have caused much new research in this area, with a concentration on the study of normal voices and using the results for more natural synthesis. Areas worthy of investigation are:

- Distributed noise generation in the vocal tract. Disturbances which are formed at one point then propagate downstream causing noise at locations within the vocal tract.

- Use of the noise component for vocal tract parameter identification, and for identification of individual speakers/singers. Features of the noise signal might indicate physiological differences between individuals.

The greatest area for improvement of the entire vocal model lies in modeling of the glottal source. Physical models based on mass/spring systems or finite element simulations of non-homogeneous material are currently too complex to allow high-quality real-time sound synthesis. As computing power increases, however, these models hold the greatest promise of true improvement in natural sounding vocal synthesis, controlled by intuitive and physically meaningful parameters.

Appendix A

Fourier and Hartley Transforms

This appendix will define the Fourier and Hartley Transforms and present theorems which are relevant to the calculations performed in the dissertation. The Fourier Transform [146] in discrete time and frequency is called the Discrete Fourier Transform (DFT), and is defined by:

$$X(m) = DFT\{x(n)\} = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi nm}{N}} \tag{A.1}$$

where

$$e^{-j\theta} = cos(\theta) - jsin(\theta) \tag{A.2}$$

The frequency in Hz. of a transform sample can be determined from the index m and the sampling rate F_S by the relationship:

$$\text{Frequency} = \frac{mF_S}{N} \tag{A.3}$$

Inverse transformation is defined by:

$$x(n) = \frac{1}{N} \sum_{n=0}^{N-1} X(m) e^{\frac{-j2\pi nm}{N}} \tag{A.4}$$

Various operations are simplified by transformation into the frequency domain. The convolution of two signals, defined by:

$$x * y(n) = \sum_{m=-\infty}^{+\infty} x(m)y(n-m) \tag{A.5}$$

can be transformed into the frequency domain by using the relationship:

$$DFT\{x * y(n)\} = X(m)Y(m) \tag{A.6}$$

Thus, given restrictions on the time extent of the two signals, the convolution operation is changed into a simple multiplication operation. Deconvolution can be performed in the frequency domain by a division operation, provided that the frequency transform contains no zero components.

The autocorrelation operation, defined as:

$$x \otimes x(n) = \sum_{i=q}^{q+N-1} x(i)x(i+m) \tag{A.7}$$

can be transformed into the frequency domain by using the relationship:

$$DFT\{x \otimes x(n)\} = X(m) * X(-m) \tag{A.8}$$

where $-m$ corresponds to $N-m$ in the DFT formulation. Real time signals exhibit Hermitian symmetry,

$$X(-m) = X^*(m) \tag{A.9}$$

so the autocorrelation relationship in the frequency domain simplifies to:

$$DFT\{x \otimes x(n)\} = |X(m)|^2 \tag{A.10}$$

To obtain the autocorrelation function in the time domain, inverse transformation is performed. The symmetry properties of the frequency domain autocorrelation function allow inverse transformation by the cosine transform:

$$x \otimes x(n) = \frac{2}{N} \sum_{n=0}^{\frac{N}{2}-1} |X(m)|^2 cos(\frac{2\pi nm}{N}) \tag{A.11}$$

Calculation of signal power is accomplished the same way in the time and frequency domains, as given by:

$$\sum_{n=0}^{N-1} |x(n)|^2 = \sum_{n=0}^{N-1} |X(m)|^2 \tag{A.12}$$

The Discrete Hartley Transform and its inverse are given by:

$$X_H(m) = DHT\{x(n)\} = \sum_{n=0}^{N-1} x(n) cas(\frac{2\pi nm}{N}) \tag{A.13}$$

$$x(n) = \frac{1}{N} \sum_{n=0}^{N-1} X(m) cas(\frac{2\pi nm}{N}) \tag{A.14}$$

where

$$cas(\theta) = cos(\theta) + sin(\theta) \tag{A.15}$$

The Hartley Transform [147] operates on real data and yields real data, so transformation of real data is somewhat simplified by the use of the Hartley Transform. The Fourier Transform calculations can be optimized to accept real data, yielding the same computational complexity as the Hartley Transform. Properties of the Fourier Transform and the sine and cosine functions, specifically those of symmetry, evenness, and oddness, allow simple relationships between the Fourier and Hartley Transforms to be derived:

$$X_H(m) = X_{real}(m) - X_{imag}(m) \tag{A.16}$$

$$X(m) = X_H even(m) - jX_H odd(m) \tag{A.17}$$

From these relationships, theorems such as the autocorrelation and power relationships can be derived for the Hartley Transform.

Appendix B

Object-Oriented Class Descriptions

The classes used in the SPASM software/DSP singing synthesis program are described in this appendix. The form of a class description is:

Class: SuperClass

where Class inherits instance variables and methods from SuperClass.

Methods are specified by:

- (type) methodName: (type) argument1 optional: (type) arg2 . . .

where the data types are C data types for all methods defined in this Appendix. The type before the methodName describes the data type of the returned quantity (default is an object id).

Classes and Methods Used in SPASM singing synthesis system

DiphController: Object

Diphone Controller - Controls transitions between shapes and glottal states during diphone synthesis. Synthesizes short soundfiles from transition parameters. Saves and

loads parameter files to/from disk.

DSPSinger: SynthPatch

NeXT Musickit Motorola 56001 DSP Chip custom Synthpatch.
Uses various NeXT standard and custom unit generators.

FloatView: View

Displays floating point data arrays with normalization and notation of minimum and maximum values.

FormantEditor: Object

Locates and displays formants in a spectrum. Controls vocal tract to move least-squares to match a given set of formants.

GlotAnalyzer: Object

Records and plays soundfiles for analysis. Inverse filters input spectrum by vocal tract spectrum in the frequency-domain to yield estimated glottal spectrum. Saves resultant glottal file to disk.

Glottis: Object

Models glottis as wavetable synthesizer. Models reflection characteristic of incoming wave value from vocal tract as reflection coefficient. Principle sample generation method is:

- (float) next: (float) ampl with: (float) tractSamp;

which takes an amplitude and incoming vocal tract sample and yields a sample of output. Frequency in Hz. is set with the method:

- setFreq: (float) freq;

and reflection coefficient is set by:

- setYourGlotReflGain: (float) value;

Last output sample can be retrieved by:

- (float) lastOut: sender;

Loads and saves glottal parameters and Fourier coefficients to/from disk.

LipFilter: Object

Models lip reflection/transmission filter as simple low-pass/high-pass filters. Principle sample generation method is:

- (float) next: (float) input;

which takes an input sample and yields an output sample. Reflection gain value is set with:

- setYourLipReflGain: (float) value;

and state variables are cleared with:

- clearOut: sender;

Last output and reflection samples can be retrieved by:

- (float) lastOut: sender;
- (float) lastRefl: sender;

NasalTract: Object

Models nasopharynx as WaveGuide Digital Filter (WGF). Reflection and transmission characteristics of nostrils are included in the object. The principal sample generation method is:

- (float) next: (float) plusSamp with: (float) minusSamp;

which accepts a sample from the vocal tract glottal side (plusSamp) and a sample from vocal tract lip side (minusSamp) and yields a sample for injection into vocal tract glottis side. Sample for injection into vocal tract lip side is retrieved by:

- (float) lastPlusRefl: sender;

Sample for injection into vocal tract glottis side is retreived by:

- (float) lastMinusRefl: sender;

Velum opening size is set by:

- setYourVelumPosition: (float) value;

Scattering relations are set by:

- setShape: (float) leftRadius with: (float) rightRadius;

where leftRadius is the vocal tract radius to the glottal side of the velum, and rightRadius is the vocal tract radius to the lip side of the velum. Last wave sample output from nostrils can be retreived by:

- (float) lastOutput: sender;

NoiseAnalyzer: Object

Records and plays soundfiles for analysis. Inverse filters input spectrum by vocal tract spectrum in the frequency-domain to yield estimated noise spectrum. Fits LPC filter to spectrum, and passes resultant filter parameters to its NoiseController.

NoiseController: Object

Synthesizes noise with random number generator and four-pole filter. The principal sample generation method is:

- (float) next: (float) ampl;

which accepts an amplitude value and returns a sample. Amplitude value is multiplied by internal gain value. Filter parameters can be set by:

- setYourNoiseGain: (float) value;
- setYourNoiseAngle: (float) value;
- setYourNoiseRadius: (float) value;
- setYourNoiseAngle2: (float) value;
- setYourNoiseRadius2: (float) value;

where radii and angles are positions in the Z plane. Filter state variables can be cleared by:

- clear: sender;

PerfController: Object

Synthesizes pitch and amplitude control signals. The principal sample generation method is:

- (float) next;

which increments the object's internal time and returns an amplitude sample. Last frequency sample can be retreived by:

- (float) frequency;

and last amplitude sample can be retreived by:

- (float) amplitude;

parameters are accessed by:

- setPerfVibrFreq: (float) aFreq;
- setPerfPitch: (float) aPiitch;
- setPerfVibrAmt: (float) aVibratoAmt;
- setPerfRndAmt: (float) aRandomAmt;
- setPerfRndPeriod: (float) aRandomPeriod;

Saves and loads parameters to/from disk.

PhonController: Object

Loads, plays, and displays soundfiles and frequency transforms. Aquires impulse responses of vocal tract for analysis. Synthesizes short performances. Loads and saves vocal tract shape files to/from disk.

ShapeInterpolater: View

Uses mouse position within a region to interpolate between a number of vocal tract shapes.

SignalProcessor: Object

Applies windows to signals. Performs frequency transforms, computes magnitude and log magnitude spectra. Inverts matrices, computes LPC coefficients, and finds complex roots of polynomials.

SingerController: Application

Controls DSPSinger in realtime. Provides main interface to mouse-controlled events. This object is the SPASM application itself.

SpectrumView: View

Displays spectra with optional markers for gain, frequency, peaks, etc.

TractView: View

Displays vocal tract shape. Displays position and gain of fricative noise source. Shape parameters are set by:

- updateRadii: (float *) r;
- updateVelum: (float) v;
- updateBoth: (float *) r vel: (float) v;

VocalTract: Object

Models oropharynx as WaveGuide Digital Filter (WGF). The principal sample generation method is:

- (float) next: (float) glotSamp with: (float) lipSamp;

which accepts a sample from the vocal tract glottal side and a sample from vocal tract lip side and yields an output sample at the lip end. Output sample at glottis end is retrieved by:

- (float) lastMinus: sender;

Last ouput sample from lip end is retrieved by:

- (float) lastPlus: sender;

Tract shape or scattering coefficients can be set by:

- setYourRadius: (int) location to: (float) value;
- setCoeff: (int) location to: (float) value;

A signal can be mixed with the wave variable in the vocal tract by:

- addValue: (float) value at: (int) impulsePosition;

The tract will automatically add noise obtained from its noise generator at the current noise position by using:

- addNoise: (float) ampl;

All state variables are reset by:

- clearOut:sender;

ZPlaneView: View

Displays poles and zeroes on the complex Z plane. Methods are:

- setBackGround: (float) gray;
- setDraw: (float) gray;
- drawPole: (float) radius angle: (float) angle;
- drawZero: (float) radius angle: (float) angle;
- drawUnitCircle;
- clear;

Appendix C

Sound Examples

This appendix lists and describes the sound examples which accompany this dissertation. The sound examples are available on various tape formats from: Center for Computer Research in Music and Acoustics, Department of Music, Stanford University, Stanford, CA. 94305.

All sound examples are played twice.

1. Vowel synthesis examples.
2. Diphthong transition synthesis examples.
3. Nasal synthesis examples.
4. Nasal to vowel transition synthesis examples.
5. Voiced plosive synthesis examples.
6. Crescendo synthesis example.
7. Synthesis of sung "Shiela".
8. Synthesis of singer exercise.
9. Original utterance of "OooEeeAhh".
10. FAST resynthesis of "OooEeeAhh".

11. Original male utterance of "Easy".

12. FAST female resynthesis of "Easy".

13. Vowel synthesis examples without pitch deviation.

14. Vowel synthesis examples with pitch deviation.

15. Male vocal tone.

16. Periodic part of male vocal tone.

17. Residual part of male vocal tone extracted by period similarity processing.

18. Male vocal tone with subharmonic.

19. Extracted subharmonic component.

20. Four octave arpeggio: no pitch deviation or noise.

21. Four octave arpeggio: fixed rate and amount of periodic vibrato only.

22. Four octave arpeggio: random pitch deviation and noise.

23. Four octave arpeggio: rule-based random and periodic vibrato, no noise.

24. Four octave arpeggio: rule-based random and periodic vibrato, fixed glottal noise.

25. Four octave arpeggio: rule-based random and periodic vibrato, rule based glottal noise.

Bibliography

Singing, Speech, and Acoustics

[1] D. R. Appelman, *The Science of Vocal Pedagogy.* Bloomington IN: Indiana University Press, 1967.

[2] J. Backus, *The Acoustical Foundations of Music.* New York: W. W. Norton, 1969.

[3] G. Bennett, "Singing Synthesis in Electronic Music," in *Research Aspects on Singing*, pp. 34-50, Stockholm: Royal Swedish Academy of Music, 1981.

[4] G. Bloothooft and R. Plomp, "Spectral Analysis of Sung Vowels. III. Characteristics of Singers and Modes of Singing," Journal of the Acoustical Society of America, vol. 79, no. 3, pp. 852-864, 1986.

[5] G. Bloothooft and R. Plomp, "The Sound Level of the Singer's Formant in Professional Singing," Journal of the Acoustical Society of America, vol. 79, no. 6, pp. 2028-2033, 1986.

[6] T. Chiba and M. Kajiyama, *The Vowel - Its Nature and Structure.* Tokyo: Phonetic Society of Japan, 1941.

[7] *CRC Handbook of Chemistry and Physics.* Boca Raton, FL: CRC Press, 1984-1985.

[8] G. Fant, *Acoustic Theory of Speech Production.* The Hague, The Netherlands: Mouton, 1960.

[9] J. L. Flanagan, *Speech Analysis, Synthesis, and Perception.* 2nd. Ed., Berlin: Springer Verlag, 1972.

[10] J. L. Flanagan, *Speech Synthesis.* 2nd. Ed., Stroudsburg, PA.: Hutchinson and Ross, 1973.

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

[12] R. Miller, *The Structure of Singing: System and Art in Vocal Technique.* New York: Schirmer Books, Macmillan, 1986.

[13] P. M. Morse, *Vibration and Sound.* Acoustical Society of America, 1976.

[14] P. M. Morse and K. U. Ingard, *Theoretical Acoustics.* Princeton, NJ: Princeton University Press, 1968.

[15] D. O'Shaughnessy, *Speech Communication, Human and Machine.* Addison-Wesley, Reading, MA, 1987.

[16] M. Rothenberg, "The Voice Source In Singing," in *Research Aspects on Singing*, pp. 15-33, Stockholm: Royal Swedish Academy of Music, 1981.

[17] R. W. Schafer and J. D. Markel eds., *Speech Analysis.* New York: IEEE Press, 1979.

[18] N. Scotto di Carlo and D. Autesserre, "Movements of the Velum in Singing," Journal of Research in Singing and Applied Vocal Pedagogy, vol. 11, no. 1 pp. 3-13, 1987.

[19] D. Stanley, *The Science of Voice.* New York: Carl Fischer, 1958.

[20] J. Sundberg, "Articulatory interpretation of the "singing formant"," Journal of the Acoustical Society of America, vol. 55, no. 4, pp. 838-844, 1974.

[21] J. Sundberg, *The Science of The Singing Voice.* Dekalb Il.: Northern Illinois University Press, 1987.

[22] J. Sundberg, "Vibrato and Vowel Identification," Archives of Acoustics, vol. 2, pp. 257-266, 1977.

[23] G. J. Troup, G. Welch, M. Volo, A. Tronconi, F. Ferrero, and E. Farnetani, "On Velum Opening in Singing," Journal of Research in Singing and Applied Vocal Pedagogy, vol. 13, no. 1 pp. 35-39, 1988.

[24] W. Vennard, *Singing, the Mechanism and the Technic.* New York: Carl Fischer, 1967.

[25] B. D. Wyke, "Laryngeal Neuromuscular Control Systems in Singing," Folia Phoniatrica, vol. 26, pp. 295-306, 1974.

Spectral and Formant-Based Voice Models

[26] J. B. Allen, "Short Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 25, no. 3, pp. 235-238, 1977.

[27] J. B. Allen and L. R. Rabiner, "A Unified Approach to Short-Time Fourier Analysis and Synthesis," Proceedings of the IEEE, vol. 65, no. 11, pp. 1558-1564, 1977.

[28] J. M. Chowning, "Frequency Modulation Synthesis of the Singing Voice," in *Some Current Directions in Computer Music Research.*, Cambridge MA, MIT Press, pp. 57-63, 1989.

[29] J. M. Chowning, "The Synthesis of Complex Audio Spectra by Means of Frequency Modulation," Journal of the Audio Engineering Society, vol. 21, no. 7, Sept. 1973.

[30] M. Dolson, "The Phase Vocoder: A Tutorial," Computer Music Journal, vol. 10, no. 4, pp. 14-27, 1986.

[31] J. L. Flanagen and R. M. Golden, "Phase Vocoder," Bell Systems Technical Journal, vol. 45, pp. 1493-1509, 1966.

[32] W. Kaegi and S. Tempelaars, "VOSIM - A New Sound Synthesis System," Journal of the Audio Engineering Society, vol. 26, no. 6, pp. 418-425, 1978.

[33] D. H. Klatt, "Software for a Cascade/Parallel Formant Synthesizer," Journal of the Acoustical Society of America, vol. 67, no. 3, pp. 971-995, 1980.

[34] R. J. McAulay and T. F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 34, no. 4, pp. 744-754, 1986.

[35] J. A. Moorer, "The Use of the Phase Vocoder in Computer Music Applications," Journal of the Audio Engineering Society, vol. 26, no. 1/2, pp. 42-45, 1978.

[36] N. B. Pinto, D. G. Childers and A. L. Lalwani, "Formant Speech Synthesis: Improving Production Quality," IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 37, no. 12, pp. 1870-1887, 1989.

[37] T. F. Quatieri and R. J. McAulay, "Speech Transformations Based on a Sinusoidal Representation," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 34, no. 6, pp. 1449-1464, 1986.

[38] L. R. Rabiner, "Digital-Formant Synthesizer," Journal of the Acoustical Society of America, vol. 43, no. 4, pp. 822-828, 1968.

[39] L. R. Rabiner, "Speech Synthesis by Rule: An Acoustic Domain Approach," Bell System Technical Journal, vol. 47, pp. 17-37, 1968.

[40] X. Rodet, "Time-Domain Formant-Wave-Function Synthesis," Computer Music Journal, vol. 8, no. 3, pp. 9-14, 1984.

[41] X. Rodet, Y. Potard, and J. B. Barriere, "The CHANT Project: From the Synthesis of the Singing Voice to Synthesis in General," Computer Music Journal, vol. 8, no. 3, pp. 15-31, 1984.

Source/Filter and Articulatory Vocal Tract Models

[42] B. S. Atal and S. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave" Journal of the Acoustical Society of America, vol. 50, pp. 637-655, 1971.

[43] B. S. Atal and J. R. Remde, "A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates," Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing, pp. 614-617, 1982.

[44] B. S. Atal and M. R. Schroeder, "Adaptive Predictive Coding of Speech Signals," The Bell System Technical Journal, pp. 1973-1986, 1970.

[45] B. S. Atal and M. R. Schroeder, "Predictive Coding of Speech Signals and Subjective Error Criteria," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 27, no. 3, pp. 247-254, 1979.

[46] L. J. Bonder, "The n-Tube Formula and Some of its Consequences," Acustica, vol. 52, pp. 216-226, 1983.

[47] C. H. Coker, "A Model of Articulatory Dynamics and Control," Proc. of the IEEE, vol. 64, no. 4, pp. 542-560, 1976.

[48] P. R. Cook, "SPASM: a Real-Time Vocal Tract Physical Model Editor/Controller and Singer: the Companion Software Synthesis System," Colloque les Modèles Physiques Dans L'Analyse, la Production et la Création Sonore, ACROE, Grenoble, 1990, publication expected 1991.

[49] P. R. Cook, "Synthesis of the Singing Voice Using a Physically Parameterized Model of the Human Vocal Tract," Proc. of the International Computer Music Conference, pp. 69-72, Columbus, OH, 1989

[50] G. Fant, *Acoustic Theory of Speech Production with Calculations Based on X-Ray Studies of Russian Articulations.* 2nd Ed., The Hague, The Netherlands: Mouton, 1970.

[51] G. Fant, Q. Lin and P. Badin, "Speech Production Models: Constraints and Control Strategies," The Second Joint Meeting of the ASA and ASJ, NN6, 1988.

[52] Y. Kakita and O. Fujimura, "Computational Model of the Tongue," Journal of the Acoustical Society of America, vol. 62, S15A.

[53] Y. Kakita and K. Honda, "Stability of Vowel Formants Based on a Simple Acoustic Tube Model and a Tongue Model," The Second Joint Meeting of the ASA and ASJ, 1988.

[54] J. L. Kelly and C. C. Lochbaum, "Speech Synthesis," Proc. Fourth Intern. Congr. Acoust., paper G42, pp. 1-4, 1962.

[55] J. Liljencrants, "Speech Synthesis With a Reflection-Type Line Analog," DS Dissertation, Department of Speech Communication and Music Acoustics, Royal Institute of Technology, Stockholm, Sweden, 1985.

[56] J. Makhoul, "Linear Prediction: A Tutorial Review," Proc. of the IEEE, vol. 63, pp. 561-580, 1975.

[57] S. Maeda, "Generation and Propagation of Sounds Inside the Vocal Tract: A Time-Domain Simulation," Colloque les Modèles Physiques Dans L'Analyse, la Production et la Création Sonore, ACROE, Grenoble, 1990, publication expected 1991.

[58] P. Mermelstein, "Articulatory Model for the Study of Speech Production," Journal of the Acoustical Society of America, vol. 53, no. 4, pp. 1070-1082, 1973.

[59] X. Rodet, P. Depalle, "High Quality Synthesis-by-Rule of Consonants," Proceedings of the International Computer Music Conference, pp. 91-96, 1985.

[60] X. Rodet, P. Depalle and G. Poirot, "Diphone Sound Synthesis Based on Spectral Envelopes and Harmonic/Noise Excitation Functions," Proceedings of the International Computer Music Conference, pp. 313-321, 1988.

[61] S. Singhal and B. S. Atal, "Improving Performance of Multi-Pulse LPC Coders at Low Bit Rates," Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing, session 1.3, pp. 1-4, 1984.

[62] J. O. Smith, "Efficient Yet Accurate Models for Strings and Air Columns Using Sparse Lumping of Distributed Losses and Phase Dispersion," Colloque les Modèles Physiques Dans L'Analyse, la Production et la Création Sonore, ACROE, Grenoble, 1990, publication expected 1991.

[63] J. O. Smith, "Musical Applications of Digital Waveguides," Stanford University Depart of Music Report, STAN-M-39, 1987.

[64] J. O. Smith, "Waveguide Filter Tutorial," Proc. of the International Computer Music Conference, pp. 9-16, Urbana, Il, 1987.

[65] H. Suzuki and T. Nakai, "Speech Production by a Vocal Cords - Vocal Tract - Vocal Tract Wall Vibration Model," The Second Joint Meeting of the ASA and ASJ, NN8, 1988.

Glottal Source Models and Identification, Vocal Tract Shape Identification and Inverse Filtering

[66] B. S. Atal, J. J. Chang, M. V. Mathews and J. W. Tukey, "Inversion of Articulatory-To-Acoustic Transformation in the Vocal Tract by a Computer-Sorting Technique," Journal of the Acoustical Society of America, vol. 63, no. 5, pp. 1535-1555, 1978.

[67] B. Cranen, L. Boves, "On the Measurement of Glottal Flow," Journal of the Acoustical Society of America, vol. 84, no. 3, pp. 888-900, 1988.

[68] K. E. Cummings and M. A. Clements, "Analysis of Glottal Waveforms Across Stress Styles," Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, S6b.6, pp. 369-372, Albuquerque, NM, 1990.

[69] T. Galas and X. Rodet, "An Improved Cepstral Method for Deconvolution of Source-Filter Systems With Discrete Spectra: Application to Musical Sound Signals," Proc. of the International Computer Music Conference, pp. 82-84, Glasgow, Scotland, 1990.

[70] J. Gauffin and J. Sundberg, "Data on the Glottal Voice Source Behavior in Vowel Production," Quarterly Report of the Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, no. 2, pp. 61-70, 1980.

[71] B. Gopinath and M. M. Sondhi, "Determination of the Shape of the Human Vocal Tract From Acoustical Measurements," The Bell System Technical Journal, pp. 1195-1214, July-August 1970.

[72] J. N. Holmes, "An Investigation of the Volume Velocity Waveform at the Larynx During Speech by Means of an Inverse Filter," Fourth International Congress on Acoustics, Copenhagen, August 1962.

[73] K. Ishizaka and J. L. Flanagan, "Synthesis of Voiced Sounds From a Two-Mass Model of the Vocal Cords," Bell System Technical Journal, vol. 51, pp. 1233-1268, 1972.

[74] J. Kacprowski, "Physical Models of the Larynx Source," Archives of Acoustics, vol. 2, no. 1, pp. 47-70, 1977.

[75] F. L. E. Lecluse, M. P. Brocaar and J. Verschuure, "The Electroglottography and its Relation to Glottal Activity," Folia Phoniatrica, vol. 27, pp. 215-224, 1975.

[76] J. D. Markel, "Digital Inverse Filtering - A New Tool for Formant Trajectory Estimation," IEEE Trans. on Audio and Electroacoustics, AU-20, pp. 129-137, 1972.

[77] P. Mermelstein, "Determination of the Vocal-Tract Shape from Measured Formant Frequencies," Journal of the Acoustical Society of America, vol. 41, no. 5, pp. 1283-1294, 1967.

[78] J. E. Miller and M. V. Mathews, "Investigation of the Glottal Waveshape by Automatic Inverse Filtering," Journal of the Acoustical Society of America, vol. 35, p. 1836, 1963.

[79] A. E. Rosenberg, "Effect of Glottal Pulse Shape on the Quality of Natural Vowels," Journal of the Acoustical Society of America, vol. 49, no. 2.2, pp. 583-590, 1971.

[80] M. Rothenberg, "A New Inverse-Filtering Technique for Deriving the Glottal Air Flow Waveform During Voicing," Journal of the Acoustical Society of America, vol. 53, no. 6, pp. 1632-1645, 1973.

[81] M. R. Schroeder, "Determination of the Geometry of the Human Vocal Tract From Acoustic Measurements," Journal of the Acoustical Society of America, vol. 41, no. 4, pp. 1002-1010, 1967.

[82] M. M. Sondhi and B. Gopinath, "Determination of the Vocal-Tract Shape from Impulse Response at the Lips," Journal of the Acoustical Society of America, vol. 49, no. 6, pp. 1867-1873, 1971.

[83] J. Sundberg and J. Gauffin, "Spectral Correlates of Glottal Voice Source Waveform Characteristics," Journal of Speech and Hearing Research, vol. 32, pp. 556-565, 1989.

[84] I. R. Titze, "The Human vocal Cords: A Mathematical Model. Part 1", Phonetica, vol. 28, pp. 129-170, 1973.

[85] I. R. Titze, "The Human vocal Cords: A Mathematical Model. Part 2", Phonetica, vol. 29, pp. 1-21, 1974.

[86] I. R. Titze and D. T. Talkin, "A Theoretical Study of the Effects of Various Laryngeal Configurations on the Acoustics of Phonation," Journal of the Acoustical Society of America, vol. 66, no. 1, pp. 60-74, 1979.

[87] I. R. Titze and D. T. Talkin, "The Physics of Small-Amplitude Oscillation of the Vocal Folds," Journal of the Acoustical Society of America, vol. 83, pp. 1536-1552, 1988.

[88] H. Wakita, "Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms," IEEE Trans. on Audio and Electroacoustics, AU-21, pp. 417-427, 1973.

[89] H. Wakita, "Estimation of Vocal-Tract Shapes from Acoustical Analysis of the Speech Wave: the State of the Art," IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 27, pp. 281-285, 1979.

Pitch Detection and Singer Vibrato

[90] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch Extraction from Linear Prediction Residual for Identification of Closed Glottis Interval," IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 27, no. 4, pp. 309-319, 1979.

[91] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch Extraction of Voiced Speech," IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 23, no. 6, pp. 562-570, 1975.

[92] Y. M. Cheng and D. O'Shaughnessy, "Automatic and Reliable Estimation of Glottal Closure Instant and Period," IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 37, no. 12, pp. 1805-1815, 1989.

[93] J. F. Deem, W. H. Manning, J. V. Knack and J. S. Matesich, "The Automatic Extraction of Pitch Perturbation Using Microcomputers: Some Methodological Considerations," Journal of Speech and Hearing Research, vol. 32, pp. 689-697, 1989.

[94] T. Haji, S. Horiguchi, T. Baer and W. Gould, "Frequency and Amplitude Perturbation Analysis of Electroglottograph During Sustained Phonation," Journal of the Acoustical Society of America, vol. 80, no. 1, pp. 58-62, 1986.

[95] W. Hess, *Pitch Determination of Speech Signals.* Berlin: Springer Verlag, 1983.

[96] T. Kobayashi and H. Sekine, "Statistical Properties of Fluctuation of Pitch Intervals and its Modeling for Natural Synthetic Speech," Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, S6a.8, pp. 321-324, Albuquerque, NM, 1990.

[97] R. Maher and J. Beauchamp, "An Investigation of Vocal Vibrato for Synthesis," Applied Acoustics, vol. 30, pp. 219-245, 1990.

[98] C. A. McGonegal, L. R. Rabiner and A. E. Rosenberg, "A SemiAutomatic Pitch Detector (SAPD)," IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 23, no. 6, pp. 570-574, 1975.

[99] R. L. Miller and E. S. Weibel, "Measurements of the Fundamental Period of Speech Using a Delay Line," Journal of the Acoustical Society of America, vol. 28, Abstract, 1956.

[100] J. A. Moorer, "The Optimum Comb Method of Pitch Period Analysis of Continuous Digitized Speech," IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 22, no. 5, pp. 330-338, 1974.

[101] D. Morrill and P. R. Cook, "Hardware, Software, and Compositional Tools for a Real-Time Improvised Solo Trumpet Work," Proc. of the International Computer Music Conference, pp. 211-214, Columbus, OH, 1989.

[102] H. Ney, "A Time Warping Approach to Fundamental Period Estimation," IEEE Trans., SMC-12, pp. 383-388 [8.3.3], 1982.

[103] R. R. Orikoff and R. J. Baken, "Fundamental Frequency Modulation of the Human Voice by the Heartbeat: Preliminary Results and Possible Mechanisms," Journal of the Acoustical Society of America, vol. 85, no. 2, pp. 888-893, 1989.

[104] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg and C. A. McGonegal, "A Comparative Performance Study of Several Pitch Detection Algorithms," IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 24, no. 5, pp. 399-418, 1976.

[105] C. L. Reid, "The Nature of Vibrato," Journal of Research in Singing, vol. 12, no. 2, pp. 39-61, 1989.

[106] D. C. Rife, "Digital Tone Parameter Estimation in the Presence of Goussian Noise," PhD dissertation, Polytechnical Institute of Brooklyn, Brooklyn, NY, 1973.

[107] D. C. Rife and R. R. Boorstyn, "Single-Tone Parameter Estimation from Discrete-Time Observations," IEEE Trans. on Information Theory, vol. 20, no. 5, pp. 591-598, 1974.

[108] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg and H. J. Manley, "Average Magnitude Difference Function Pitch Extractor," IEEE Trans. on Acoustics, Speech and Signal Processing, vol. 22, no. 5, pp. 353-362, 1974.

[109] M. M. Sondhi, "New Methods of Pitch Extraction," IEEE Trans. on Audio and ElectroAcoustics, vol. 16, no. 2, pp. 262-266, 1968.

[110] H. W. Strube, "Determination of the Instant of Glottal Closure From the Speech Wave," Journal of the Acoustical Society of America, vol. 56, no. 5, pp. 1625-1629, 1974.

[111] S. Ternström and A. Friberg, "Analysis and Simulation of Small Variations in the Fundamental Frequency of Sustained Vowels," Quarterly Report of the Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, no. 3, pp. 1-14, 1989.

Airflow and Noise in Vocal Signals

[112] P. Badin, "Acoustics of Voiceless Fricatives: Production Theory and Data," Quarterly Report of the Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, no. 3, pp. 33-55, 1989.

[113] W. K. Blake, *Mechanics of Flow-Induced Sound and Vibration, Volume 1: General Concepts and Elementary Sources.* New York: Academic Press, 1986.

[114] W. K. Blake, *Mechanics of Flow-Induced Sound and Vibration, Volume 2: Complex Flow-Structure Interactions.* New York: Academic Press, 1986.

[115] F. T. Brown, D. L. Margolis and R. P. Shah, "Small-Amplitude Frequency Behavior of Fluid Lines With Turbulent Flow," Transactions of the ASME, vol. 91D, pp. 678-693, 1969.

[116] G. A. Cavagna and R. Margaria, "Airflow Rates and Efficiency Changes During Phonation," Annals of the New York Academy of Sciences, 155, Article 1, Special Issue - Sound Production in Man, pp. 152-164, Nov. 1968.

[117] C. D. Chafe, "Pulsed Noise and Microtransients in Physical Models of Musical Instruments," Colloque les Modèles Physiques Dans L'Analyse, la Production et la Création Sonore, ACROE, Grenoble, 1990, publication expected 1991.

[118] C. D. Chafe, "Pulsed Noise in Self-Sustained Oscillations of Musical Instruments," Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, A1.17, pp. 1157-1160, Albuquerque, NM, 1990.

[119] P. R. Cook, C. D. Chafe and J. O. Smith, "Pulsed Noise in Musical Systems, Techniques for Extraction, Analysis, and Visualization," Proc. of the International Computer Music Conference, pp. 63-65, Glasgow, Scotland, 1990.

[120] N. B. Cox, M. R. Ito and M. D. Morrison, "Technical Considerations in Computation of Spectral Harmonics-to-Noise Ratios For Sustained Vowels," Journal of Speech and Hearing Research, vol. 32, pp. 203-218, 1989.

[121] B. R. Gerratt, K. Precoda, D. G. Hanson and G. S. Berke, "Source Characteristics of Diplophonia," Journal of the Acoustical Society of America, vol. 83, S66, 1988.

[122] J. M. Heinz and K. N. Stevens, "On the Properties of Voiceless Fricative Consonants," Journal of the Acoustical Society of America, vol. 33, pp. 589-596, 1961.

[123] J. Hillenbrand, "A Methodological Study of Perturbation and Additive Noise in Synthetically Generated Voice Signals," Journal of Speech and Hearing Research, vol. 30, pp. 448-461, 1987.

[124] A. Hirschberg, "Some Fluid Dynamics Aspects of Speech," Fourth Colloquium Signaalanalyse en Spraak, COLSAS, Instituut voor Perceptie Onderzoek, Eindhoven, 1990.

[125] A. Hirschberg, R. W. A. van de Laar, J. P. Marrou-Maurières, A. P. J. Wijnands, H. J. Dane, S. G. Kruijswijk, and A. J. M. Houtsma, "A Quasi-Stationary Model of Air Flow in the Reed Channel of Single-Reed Woodwind Instruments," Acustica, vol. 70, pp. 146-154, 1990.

[126] E. B. Holmberg, R. E. Hillman and J. S. Perkell, "Glottal Airflow and Transglottal Air Pressure Measurements for Male and Female Speakers in Soft, Normal, and Loud Voice," Journal of the Acoustical Society of America, vol. 84, no. 2, pp. 511-529, 1988.

[127] H. Iijima, N. Miki, and N. Nagai, "Viscous Flow Analyses of the Glottal Model Using a Finite Element Method," The Second Joint Meeting of the ASA and ASJ, NN10, 1988.

[128] J. F. Kaiser, "Some Observations on Vocal Tract Operation from a Fluid Flow Point of View," Conference on Physiology and Biophysics of Voice, Iowa City, IA, 1983.

[129] H. Kasuya, S. Ogawa, K. Mashima and S. Ebihara, "Normalized Noise Energy as an Acoustic Measure to Evaluate Pathologic Voice," Journal of the Acoustical Society of America, vol. 80, no. 5, pp. 1329-1334, 1986.

[130] G. C. Kingston, "Experimental and Theoretical Studies of Pulsating Turbulent Flow," PhD Dissertation, Clarkson College of Technology, 1975.

[131] Y. Kioke and M. Hirano, "Glottal-Area Time Function and Subglottal-Pressure Variation," Journal of the Acoustical Society of America, vol. 54, no. 6, pp. 1618-1627, 1973.

[132] H. Muta, T. Baer, K. Wagatsuma, T. Muraoka and H. Fukuda, "A Pitch-Synchronous Analysis of Hoarseness in Running Speech," Journal of the Acoustical Society of America, vol. 84, no. 4, pp. 1292-1301, 1988.

[133] N. B. Pinto and I. R. Titze, "Unification of Perturbation Measures in Speech Signals," Journal of the Acoustical Society of America, vol. 87, no. 3, pp. 1278-1289, 1990.

[134] H. J. Rubin, M. LeCover and W. Vennard, "Vocal Intensity, Subglottic Pressure and Air Flow Relationships in Singers," Folia Phoniatrica, vol. 19, pp. 393-413, 1967.

[135] N. P. Solomon, G. N. McCall, M. W. Trosset and W. C. Gray, "Laryngeal Configuration and Constriction During Two Types of Whispering," Journal of Speech and Hearing Research, vol. 32, pp. 161-174, 1989.

[136] V. L. Streeter, *Fluid Dynamics.* New York: Mcgraw-Hill, 1948.

[137] H. M. Teager, "Some Observations on Oral Air Flow During Phonation," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 5, pp. 599-601, 1980.

[138] H. M. Teager and S. M. Teager, "A Phenomenological Model for Vowel Production in the Vocal Tract," in R. G. Daniloff Ed., *Speech Science - Recent Advances*, San Diego, CA: College-Hill Press, 1985.

[139] R. T. Schumacher and C. D. Chafe, "Characterization of Aperiodicity in Nearly Periodic Signals," Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, A1.18, pp. 1161-1164, Albuquerque, NM, 1990.

[140] C. H. Shadle, "The Acoustics of Fricative Consonants," PhD dissertation, Massachusetts Institute of Technology Research Laboratory of Electronics, Cambridge, MA, 1985.

[141] E. Yumoto, W. Gould and T. Baer, "Harmonics-to-Noise Ratio as an Index of the Degree of Hoarseness," Journal of the Acoustical Society of America, vol. 71, no. 6, pp. 1544-1550, 1982.

[142] E. Yumoto, Y. Sasaki and H. Okamura, "Harmonics-to-Noise Ratio and Psychophysical Measurement of the Degree of Hoarseness," Journal of Speech and Hearing Research, vol. 27, pp. 2-6, 1984.

Spectrum Analysis and Digital Signal Processing

[143] N. Ahmed, D. R. Hummels, M. Uhl and D. L. Soldan, "A Sequential Regression Algorithm for Recursive Filters," Electronics Letters, p. 266, Apr. 27, 1978.

[144] I. Bowler, P. Manning, A. Purvis and N. Bailey, "On Mapping N Articulation Onto M Synthesizer-Control Parameters," Proc. of the International Computer Music Conference, pp. 181-184, Glasgow, Scotland, 1990.

[145] A. M. Bruckstein and T. Kailath, "Inverse Scattering for Discrete Transmission Line Models," SIAM Review, vol. 29, no. 3, pp. 359-389, 1987.

[146] R. N. Bracewell, *The Fourier Transform and Its Applications.* New York: McGraw-Hill, 1986.

[147] R. N. Bracewell, *The Hartley Transform.* New York: Oxford University Press, 1986.

[148] K. P. Bube and R. Burridge, "The One-Dimensional Inverse Problem of Reflection Seismology," SIAM Review, vol. 25, no. 4, pp. 497-559, 1983.

[149] H. Chamberlin, *Musical Applications of Microprocessors.* New Jersey: Hayden Book Company, 1980.

[150] J. M. Cioffi, "Limited-Precision Effects in Adaptive Filtering," IEEE Transactions on Circuits and Systems, vol. 34, no. 7, pp. 821-833, 1987.

[151] J. Durbin, "The Fitting of Time-Series Models," Rev. Inst. Int. Stat., vol. 28, no. 3, pp. 233-243, 1960.

[152] A. H. Gray, "Passive Cascaded Lattice Digital Filters," IEEE Transactions on Circuits and Systems, vol. 27, no. 5, pp. 339-344, 1980.

[153] F. J. Harris, "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform," Proceedings of the IEEE, vol. 66, no. 1, pp. 51-84, 1978.

[154] L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification.* Cambridge, MA: MIT Press, 1987.

[155] F. R. Moore, "Table Lookup Noise for Sinusoidal Digital Oscillators," in C. Roads and J. Strawn eds. *Foundations of Computer Music.* pp. 326-334, MIT Press, 1985.

[156] A. H. Nuttall, "Some Windows With Very Good Sidelobe Behavior," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 29, no. 1, pp. 84-91, 1981.

[157] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing.* New Jersey, Prentice-Hall, 1975.

[158] L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing.* New Jersey: Prentice-Hall, 1975.

[159] R. W. Schafer and L. R. Rabiner, "A Digital Signal Processing Approach to Interpolation," Proceedings of the IEEE, vol. 61, pp. 692-702, 1973.

[160] X. J. Serra, "A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic Plus Stochastic Decomposition," Ph.D. Dissertation. Dept. of Music Rep. STAN-M-58, Stanford University, 1989.

[161] J. O. Smith, "Techniques for Digital Filter Design and System Identification with Application to the Violin," PhD Dissertation, Stanford University Department of Electrical Engineering, 1985.

[162] J. O. Smith and P. Gossett, "A Flexible Sampling-Rate Conversion Method," Proc. of the IEEE Conference on Acoustics, Speech, and Signal Processing, San Diego, CA, March, 1984.

[163] B. Widrow and M. Hoff Jr., "Adaptive Switching Circuits," IRE Wescon Conv. Rec., Pt. 4, pp. 96-104, 1960.

[164] B. Widrow and S. D. Stearns, *Adaptive Signal Processing.* New Jersey: Prentice-Hall, 1985.

[165] A. E. Yagle, "Fast Algorithms for Estimation and Signal Processing: An Inverse Scattering Framework," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 6, pp. 957-959, 1989.

Neurological and Perceptual Aspects of Music, Speech, and Singing

[166] W. M. Campbell and J. F. Michel, "The Effects of Auditory Masking on Vocal Vibrato," in V. Lawrence and B. Weinberg eds., *Transcripts of the Eighth Symposium: Care of the Professional Voice, Part I: Physical Factors in Voice, Vibrato, Registers.* pp. 50-56, New York: The Voice Foundation, 1980.

[167] P. Delattre, A. M. Liberman, F. S. Cooper and L. J. Gerstman, "An Experimental Study of the Acoustic Determinants of Vowel Color: Observations on One- and Two-Formant Vowels Synthesized From Spectrographic Patterns," Word, pp. 195-210, 1952.

[168] L. Elliot and A. Niemoeller, "The Role of Hearing in Controlling Voice Fundamental Frequency," International Audiology, vol. 9, pp. 47-52, 1970.

[169] H. Gardner, *Frames of Mind: the Theory of Multiple Intelligences.* New York: Basic Books, 1983.

[170] J. Hillenbrand, "Perception of Aperiodicities in Synthetically Generated Voices," Journal of the Acoustical Society of America, vol. 83, no. 6, pp. 2361-2371, 1988.

[171] N. Isshiki, "Regulatory Mechanism of Voice Intensity Variation," Journal of Speech and Hearing Research, vol. 7, pp. 17-29, 1965.

[172] W. Klein, R. Plomp and L. Pols, "Vowel Spectra, Vowel Spaces, and Vowel Identification," Journal of the Acoustical Society of America, vol. 48, pp. 999-1009, 1970.

[173] A. M. Liberman, "On Finding that Speech Is Special," American Psychologist, vol. 37, no. 2, pp. 148-167, 1982.

[174] S. McAdams, "Spectral Fusion, Spectral Parsing and the Formation of Auditory Images," PhD Dissertation, Stanford University Department of Hearing and Speech Sciences, 1984.

[175] B. C. J. Moore, *An Introduction to the Psychology of Hearing.* London: Academic Press, 1982.

[176] G. E. Peterson and H. L. Barney, "Control Methods Used in a Study of the Vowels," Journal of the Acoustical Society of America, vol. 24, no. 2, pp. 175-184, 1952.

[177] J. R. Pierce, R. Lipes, and C. Cheetham, "Uncertainty Concerning the Direct Use of Time Information in Hearing: Place Clues in White-Spectra Stimuli," Journal of the Acoustical Society of America, vol. 61, no. 6, pp. 1609-1621, 1977.

[178] T. Shipp, J. Sundberg and E. T. Doherty, "The Effect of Delayed Auditory Feedback on Vocal Vibrato," Journal of Voice, vol. 1, no. 2, pp. 123-141, 1988.

[179] J. I. Shonle, and K. E. Horan, "The Pitch of Vibrato Tones," Journal of the Acoustical Society of America, vol. 67, no. 1, pp. 246-252, 1980.

[180] J. Sundberg, "To Perceive One's Own Voice and Another Person's Voice," in Research Aspects on Singing, pp. 80-96, Stockholm, Royal Swedish Academy of Music, 1981.

[181] D. Ward and E. Burns, "Singing Without Auditory Feedback," Journal of Research in Singing and Applied Vocal Pedagogy, vol. 1, no. 2 pp. 24-44, 1978.

Programming Languages and Object Oriented Programming

[182] *Objective-C 4.0 User Reference Manual.* The Stepstone Corporation, 1988.

[183] B. J. Cox, *Object-Oriented Programming: An Evolutionary Approach.* Addison-Wesley, 1986.

[184] D. C. Halbert and P. D. O'Brien, "Using Types and Inheritance in Object-Oriented Languages," Digital Equipment Corporation, 1986.

[185] B. W. Kernighan and D. M. Ritchie, *The C Programming Language.* Second edition, Prentice-Hall, 1988.

[186] B. Liskov, A. Snyder, R. Atkinson and C. Schaffert, "Abstraction Mechanisms in CLU," Communications of the ACM, vol. 20, no. 8, pp. 564-576, 1977.

[187] B. Meyer, "Eiffel: Programming for Reusability and Extendibility," SIGPLAN Notices, vol. 22, no. 2, pp. 85-94, 1987.

[188] B. Meyer, *Object-Oriented Software Construction.* Prentice Hall International Series in Computer Science, 1988.

[189] B. Meyer, "Reusability: The Case for Object-Oriented Design," IEEE Software Magazine, pp. 50-64, March, 1987.

[190] B. Stroustrup, "Data Abstraction in C," Bell Systems Technical Journal, vol. 63, no. 8, pp. 1701-1732, 1984.