UNIVERSITY OF CALIFORNIA

Los Angeles

Analysis and Synthesis of Pathological Vowels

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Electrical Engineering

by

Brian Charles Gabelman

2003

The dissertation of Brian Charles Gabelman is approved.

_____

Rajeev Jain

_____

Kung Yao

_____

David Rennels

_____

Abeer Alwan, Committee Chair

University of California, Los Angeles

2003

DEDICATION

To Anna, Lesley, and Raphael

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I wish to acknowledge the following for assistance in generation of this project:

1. Professor Abeer Alwan for providing an introduction to the art of scientific publication, guidance for project management, and time for reviewing results and documentation.

2. Drs. Bruce Gerratt and Jody Kreiman for demonstrating the cutting-edge challenges of the study of pathological voices and for providing support and venues for research at the Veterans Administration Voicelab and later at the UCLA Voicelab (a.k.a. “Bureau of Glottal Affairs”).

3. Professors Alan Wilson, Henry Samueli, Rajeev Jain, and Dave Rennels for providing excellent background coursework in digital signal processing, systems architecture, and real-time computing, all of which were applied in this project.

4. NIH grants for funding, including NIH/NIDCD grant DC01797.

VITA

1971	B.S., Physics Iowa State University Ames, Iowa
1971 – 1982	Member of the Technical Staff Aerospace Control Systems Engineering Rockwell International and TRW Los Angeles, CA
1982	M.S., Electrical Engineering Loyola Marymount University Los Angeles, CA
1982 – 1993	Microprocessor Engineer Mattel Electronics, EOIS, and Gerratt Automotive Torrance, CA
1991	United States Patent #5,050,376 Control System for Diesel Particulate Trap Regeneration System
1994-2003	Research Assistant UCLA Voicelab Los Angeles, CA
1994	Engineer Degree, Electrical Engineering UCLA Los Angeles, CA

PUBLICATIONS AND PRESENTATIONS

Crother, C. A,. Abramson, R., Spryer, E. N, Gabelman, B. "Analysis of Relaxed Static Stability and Maneuver Load Control Applications to a Large Bomber." Air Force Flight Dynamics Laboratory Technical Report AFFDL-TR-72-7, February 01, 1972

Crother, C. A., Gabelman, B., Langton, D. "Structural Mode Effects on Flying Qualities in Turbulence," Air Force Flight Dynamics Laboratory Technical Report AFFDL-TR-73-88, August, 1973.

Stiglic, P., Hardy, J., Gabelman, B. "Control Considerations for an On-Line, Active Regeneration System for Diesel Particulate Traps, Transactions of the ASME, Journal of Engineering for Gas Turbines and Power, Vol. 111, 404-409, July, 1989.

Gabelman, B., Kreiman, J., Gerratt, B., Alwan, A. "Optimization for Source Waveform Synthesis of Pathological Voices," 131st Meeting of the Acoustical Society of America. Unpublished poster 4aC18. May, 1996, Indianapolis, IN.

Gabelman, B., Kreiman, J., Gerratt, B., Antonanzas-Barroso, N., Alwan, A. "LF source model adequacy for pathological voices." 134th Meeting of the Acoustical Society Meeting of America, Unpublished poster 5aSC17. November, 1997, San Diego, CA.

Gabelman, B., Kreiman, J., Gerratt, B., Antonanzas-Barroso, N. "Perceptually motivated modeling of noise in pathological voices." Proceedings of the 16 International Congress on Acoustics and 135th Meeting of the Acoustical Society of America. P. 1293 and unpublished Poster 2pSC30. June, 1998. Seattle, WA.

Gabelman, B., Kreiman, J., Gerratt, B., Alwan, A. "Synthesis of Nonperiodic Features of Pathological Voices" 141st Meeting of the Acoustical Society of America. Unpublished poster 3aSC25. May, 2001, Chicago, IL.

Gabelman, B., and Alwan, A. "Analysis by synthesis of FM modulation and aspiration noise components in pathological voices." In ICASSP Conference Proceedings (pp. 449-452). May, 2002, Orlando, FL,

Gabelman, B., and Alwan, A. "Analysis and Synthesis of AM Components of Pathological Voices." In IEEE 2002 Workshop on Speech Synthesis. (Paper

#20154). September 11, 2002, Santa Monica, CA. IEEE Catalog Number 02EX555. ISBN. 0-7803-7396-0.

ABSTRACT OF THE DISSERTATION

Analysis and Synthesis of Pathological Vowels

by

Brian Charles Gabelman

Doctor of Philosophy in Electrical Engineering

University of California, Los Angelses, 2003

Professor Abeer Alwan, Chair

Objective methods for evaluation of pathological voices have long been sought for both clinical use and basic research; classification of types and severity of voice defects has been largely accomplished by trained clinicians using subjective rating systems. Automatic modeling and parameterization of pathological vowel samples could provide a useful advance in this effort. Towards this goal, this work develops approaches to automatically analyze and parameterize voices with methods specialized to pathological vowels. The nonperiodic components of sustained vowels, which are prominent features of pathological voices, are modeled as three phenomena: frequency modulation (FM), amplitude modulation (AM), and aspiration noise. Improved modeling accuracy was obtained by AM and FM demodulation of the original voice prior to aspiration noise analysis. Pathological voices were re-synthesized incorporating the models and measured levels of nonperiodic components. The resulting waveforms sounded more natural than the synthetic voices that did not model AM/FM and/or aspiration noise. Major results of

the work include a real-time vowel synthesizer allowing instantaneous adjustment of parameters, a flexible software vowel synthesizer, and application of an external stimulation technique to aid in the identification of the vocal tract transfer function in the case of voices with spectrally deficient glottal driving waveforms.

Chapter 1

Introduction

A method for accurately describing pathologies in the human voice in acoustic terms has long been sought. Rating scales of “roughness” or “breathiness” have been applied, but are heavily rater-dependent. Ideally, pathological voices would be sampled and automatically analyzed in terms of model parameters, which could provide ratings that are more objective. This work mounted an extended effort to apply principles of electrical engineering and signal processing to the study of pathological human voices. Pathological vowels were analyzed via the source-filter model, producing objective parameters defining the voice and allowing accurate re-synthesis. Model parameters in particular included nonperiodic components, which are most prominent and defining in pathological voices. Nonperiodic components were expressed in terms of nonperiodic frequency modulation (FM), consisting of both high frequency period variation (HFPV) and low frequency (tremor), nonperiodic amplitude modulation (AM) consisting of both high frequency shimmer and low frequency power variation, and aspiration noise. The

integration of this three component model with AM and FM demodulation techniques yielded a novel approach to the analysis and synthesis of pathological vowels. This work addressed the experimental question: "Can modeling nonperiodic components with AM, FM, and aspiration noise improve the accuracy of analysis and fidelity of synthesis of pathological vowels?" This question was addressed both by re-analysis of the synthetic signals and by subjective analysis-by-synthesis (SABS) experiments in which a listener adjusts the model parameters of a synthesizer to produce a synthetic voice sample which matches the original voice as closely as possible.

In brief, the general process of voice analysis and modeling is displayed in Fig. 1.1, and consists of the following steps:

1. Pathological voice samples are recorded from patients.
2. The voice samples are analyzed into descriptive parameters that provide sufficient information to reconstruct them.
3. Parameter extraction is validated manually and modified where necessary.
4. Synthetic versions of the original voices are computed.
5. The synthetic versions are compared to the original in perceptual experiments.

In this effort, the pathological voices used were selected from a range of disorders including vocal nodules, cancer, and lack of neural control. Pathological voices may result from a large variety of conditions. Examples include cleft palate, deaf talkers, and dysarthria [21]. In this study of sustained pathological vowels, a large source of difference

from the normal voice lies in the mechanism of generation of the source driving function in the source-filter model of speech production (Fig. 1.2). In this model, the source is the time variation in airflow from the lungs provided by the vibrations of the vocal folds. In normal voices the glottal vibrations are rhythmic and produce abrupt closures of the vocal folds, which generate a steady fundamental frequency and excites the higher frequency resonances of the vocal tract; this generates vowels of a pleasing perceptual quality, which are deemed "normal." In pathological voices, the physical structures of the glottis and their neural control mechanisms may be disrupted, producing irregular vibrations and slow or incomplete closure; this may result in voices that are perceived as abnormal. Terms such as "rough," "breathy," "creaky," "gargled," "hoarse," or "raspy" may be applied to these.

1.1 Motivation

Research in modeling and synthesizing pathological vowels is motivated by at least two goals:

1. Objective analysis and parameterization of pathology in voices. Previous efforts [23] have expounded on the need for non-subjective measures of pathology in voices. For example, different clinicians rate the same voice differently on subjective scales of "breathiness" or "roughness." Such ratings acquire importance as they are used in the evaluation of costly medical procedures sometimes applied to improve voice quality.

Objective measures provide a much-needed standard against which to measure results of such efforts. In the ideal scenario, a fully automatic voice analysis system samples the pathological voice and establishes objective measures for voice acoustic measurements such as jitter, shimmer, tremor, volume variation, fundamental frequency variation, formant modulation and other parameters. Validity of this automatic analysis would be confirmed by subjective analysis by synthesis experiments (SABS) experiments in which a listener adjusts the model parameters of a synthesizer to produce a synthetic voice sample which matches the original voice as closely as possible, thus validating the automatically determined measures. The set of measures would then constitute a standard against which different voices or the same voice at different times could be compared.

2. Generation of high fidelity synthetic voice samples for use in perceptual studies. A second goal is to generate synthetic vowel samples matching the original as closely as possible. Once accurate synthetic versions of voices have been established, their parameters of analysis and synthesis can be varied in SABS experiments to establish their perceptual effects. Accurate synthetic samples provide the starting point for several types of studies, including: [17]

- Evaluation of the perceptual importance of each parameter of synthesis.
- Measurement of variations in listener perception.
- Measurement of minimum perceivable changes in parameters (“difference lymens”).

1.2 Background and Related Work

In order to place the current project into perspective with existing similar work, some of the related efforts are outlined. In regard to complete analysis/synthesis approaches for pathological voice study, the following are related:

Childers and Lee [5] describe a study in which voices containing vocal fry, falsetto, and breathiness are analyzed and synthesized. The two step LP analysis procedure for formant determination and inverse filtering (adopted by Norma Antonanzas for the current study as described in Section 2.1) is used. Fundamental frequency determination is aided via EGG. The least-squares fitted (Lijencrants-Fant) LF [12, 31] source model (Section 2.2) is used, and the study focuses on LF parameter effects expressed as OQ (open quotient) and SQ (speed quotient). Aspiration noise is measured directly from spectra as inter-harmonic energy and is simulated via high pass filtered random noise.

Bangayan, Long, Alwan, Kreiman, Gerratt [2] report a semi-automated approach to analysis/synthesis of pathological voices. This study used existing software (Sensimetrics Sensyn 1.1 Klatt synthsizer, Sensimetrics SpeechStation 3.1, and Entropic Research Laboratory WAVES). The source-filter model of speech production is used. Automated analysis consisted of LP determination of formants and fundamental frequency tracking. Manual adjustments of source waveform, formants, and pole-zero pairs were used to match original and synthetic spectra. Fundamental frequency tracking was performed

manually for tracking lock failures. Nonperiodic features (AM and FM) were modeled either using available Klatt parameters (sinusoidal or random low frequency fundamental frequency variation) or manually varying fundamental frequency and amplitude over frame periods to match the original. Determination of aspiration noise level was by SABS (subjective analysis by synthesis). Distinction via cutoff frequency of high and low frequency AM and FM was not made; jitter and shimmer are not explicitly addressed. Good synthetic matches to about half of 24 pathological voices are reported.

Endo and Kasuya [10] describe a system that analyzes pathological sustained /a/ into parametric time sequences which characterize fundamental frequency, harmonic amplitude, harmonic spectra, and aspiration noise. Each of these is apparently evaluated over a short period of several fundamental frequency periods, and is expressed as a mean value, trend, and perturbation. Re-synthesis using these parameters is claimed to closely approach the original. This approach differs in several ways from the current project. Endo and Kasuya describe spectral properties only of the harmonic component, and this is expressed in time varying Fourier coefficients, rather than time-constant physical formants (resonances of the vocal tract) as is done here. Aspiration noise is evaluated, but without spectral features. FM variations are not specifically measured, but are accounted for in the perturbation of the fundamental period. The glottal flow derivative waveform is not evaluated at all, as the source-filter model is not specifically used; the effects of source waveform are accounted for in the time varying Fourier coefficients. In general,

the system of Endo and Kasuya seems less ambitious in encapsulating speech quality into physically understandable parameters, but may be successful at reproduction.

In regard to the measurement of nonperiodic qualities in speech, several previous studies relate to this work:

An early study by Hillenbrand [18] investigated measurement of noise levels in synthetic vowels containing jitter (high frequency FM), shimmer (high frequency AM), and aspiration noise. The technique of noise measurement used was a time domain approach of Yumoto [34] in which averaged fundamental frequency periods of the original voice time series are averaged over a short time frame to obtain a noise free representation, which is then used as a reference to subtract from each fundamental frequency cycle in the frame; the difference signal is attributed to noise and is averaged to calculate signal to noise ratios. Fundamental frequency periods were determined via a combination of manual fundamental frequency cycle marking and automatic zero crossing determination. Levels of jitter and shimmer were then automatically measured based on the fundamental frequency cycle markings. A systematic relationship between measured HNR (harmonic to noise ratio) and the levels of added jitter and shimmer was observed.

In regard to the application of ES (external source) methods for improved formant identification, the following previous works are noteworthy:

An early (1942) effort employing artificial sources is described by Tarnoczy [29,30]. A spark gap driven by a 500 volt relaxation oscillator circuit is used to generate an

impulsive noise source to excite the vocal tract; a probe containing the spark gap is shown inserted deep into the oral cavity. Oscillographs were used to record the resulting response from the vocal tract. Using simple measurements from the oscillographs, estimates are made for the frequency and “decay pattern” (bandwidth) of the first two formants of several vowels. Even with the primitive equipment available, it was possible to distinguish the effects on bandwidths of open versus closed glottis.

In 1958, House and Stevens [19] describe a very similar experiment employing the improved technology of that period. Their system again employed a spark gap, but the vocal tract transient response was magnetically recorded and played back repetitively for display on an oscilloscope to be photographed for analysis. Each formant was individually analyzed by passing the recorded voice signal through a narrow band filter tuned to the formant frequency; an exponential decay curve was fitted to the resulting photograph to determine formant bandwidth. The objective of the study was formant bandwidth determination and comparison to mathematically modeled values. The effect of open and closed glottis on bandwidths was noted: open glottis was observed to increase bandwidth.

An improvement in formant identification was made in 1971 by Fujimura and Lindqvist [13]. In this (still computer-less) setup, a chirp (sinusoidal sweep) replaces the impulse as the artificial source. The external source signal is applied by a transducer affixed directly to the outside of the throat directly above the glottis. The response is recorded by a microphone at the lips. Considerable effort was made to acoustically shield

the source from the microphone to minimize the direct path signal component. A mechanical linkage arrangement coupled the oscillator frequency dial to a pen recorder to directly produce frequency response curves. The oscillator was driven through an 8.5 second sweep, while the subject maintained the vowel configuration of the vocal tract. Despite this long period, good frequency response curves were obtained for 250 subjects. Formant bandwidths and nasalization were studied in detail.

The method of external source identification realized gains in speed and accuracy with the application of microcomputers. A 1991 effort by Djeradi, Guerin, Badin, and Perrier [9] employed pseudo-random noise as an excitation. As in [13], the stimulus was applied to the vocal tract by direct contact: a small loudspeaker was pressed directly to the throat. A baffle was inserted above the source to minimize direct path signal to the microphone at the lips. The vocal tract transfer function was calculated as the Fourier transform of the cross correlation of the (known) pseudo-random excitation signal and the resulting microphone signal. By adjustment of the length of the pseudo-random sequence, identification could occur in as little as 100ms. The technique was also shown to remain functional (although with additional noise) in the presence of vocalization. Higher power (at least 4 times the voice signal) was recommended to obtain good identification with simultaneous voicing.

In recent work by Epps, Smith, and Wolfe [11], a system (RAVE) has been developed that is capable of displays of formants in real time. This system uses an external source supplied acoustically to the lips via a loudspeaker and acoustic ducting. A microphone,

positioned next to the lips, recorded both vocalizations and response to the external source. The external source signal consisted of a broadband sum of harmonic components controlled to yield about 5 Hz spacing (far superior to the roughly 100 Hz spacing of the natural voice). Two computers with associated A/D's, D/A, amplifiers, and filters were used to generate the stimulus and record the response. Transfer function determination was made directly from the spectra of the response; a correction for background effects of transducers, acoustics of face and room, etc., was generated by measuring the baseline response with the subject's mouth closed. (This approach was used very effectively in the demonstration described in Section 3.3). A frame time of about 0.2 seconds allowed fairly precise identification uncorrupted by articulator movement. Results with simultaneous voicing are obtained by removing the voice signal from the spectrum by measuring fundamental frequency and creating a comb filter to notch out the voice harmonics. The results of resonance identification are compared with LP analysis, and the RAVE results are shown to be far more accurate than LP. The authors illustrate the use of the technique to display formant information on a CRT in real time for speech analysis and language training.

The current project seeks to achieve higher levels of automation and objectivity in the analysis and synthesis of pathological vowels. Previous efforts have been limited in objectivity and productivity because they relied on manual and perceptual methods of voice analysis. This work is important because by replacing manual operations with automatic processing both increased productivity and reproducibility may be achieved. By

improving in the accuracy and robustness of analytic approaches, more of the process of analysis and synthesis may be made hands-off.

1.3 Dissertation Outline

The rest of this document is organized as follows:

Chapter 2 discusses the collection and analysis of pathological voices. Chapter 3 discusses the improvement of formant estimation via the application of external source system identification. Chapter 4 discusses two synthesizers created to study pathological vowels: a real-time hardware implementation and a software implementation. Chapter 5 discusses details of the algorithms of synthesis of pathological voices. Chapter 6 summarizes the dissertation.

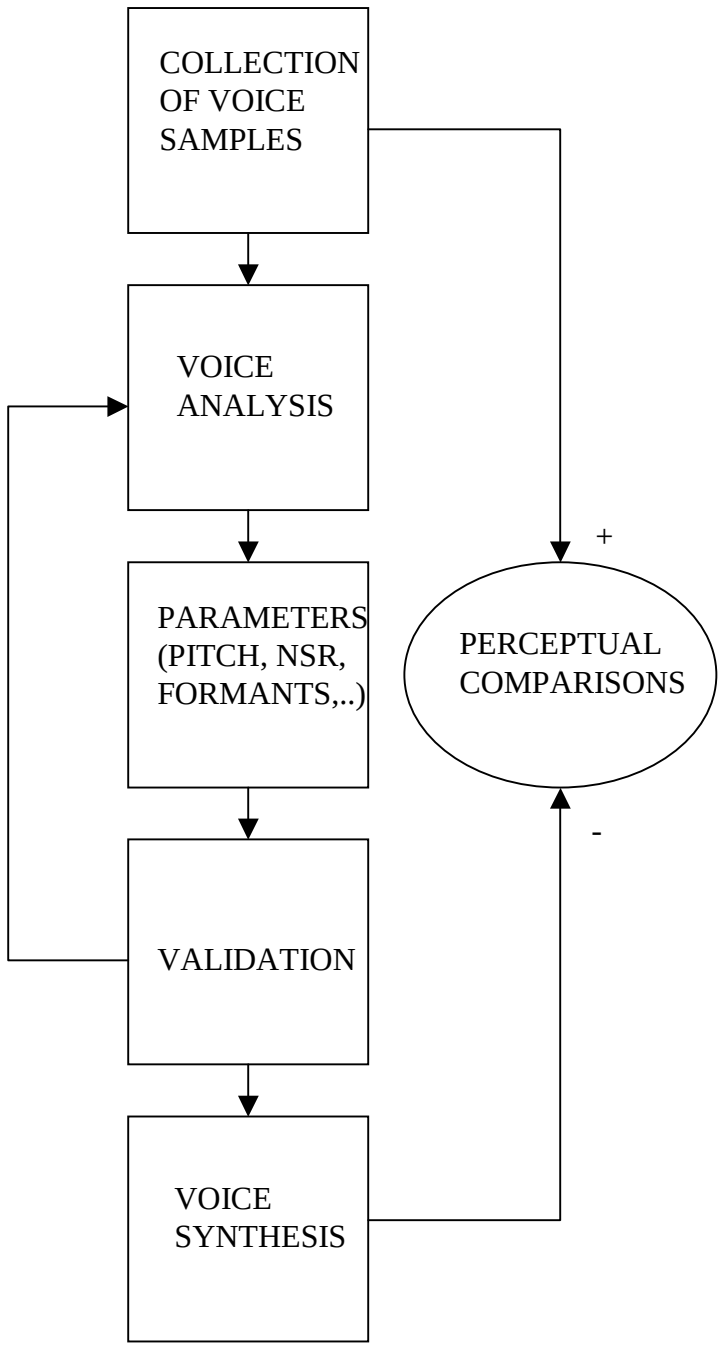


Figure 1.1. Overview of pathological voice analysis. Processing begins with digital sampling of pathological vowels, which are then analyzed into model parameters. The parameters are validated and recalculated if necessary. Using the parameters, a synthetic version of the voice is created and compared to the original recording.

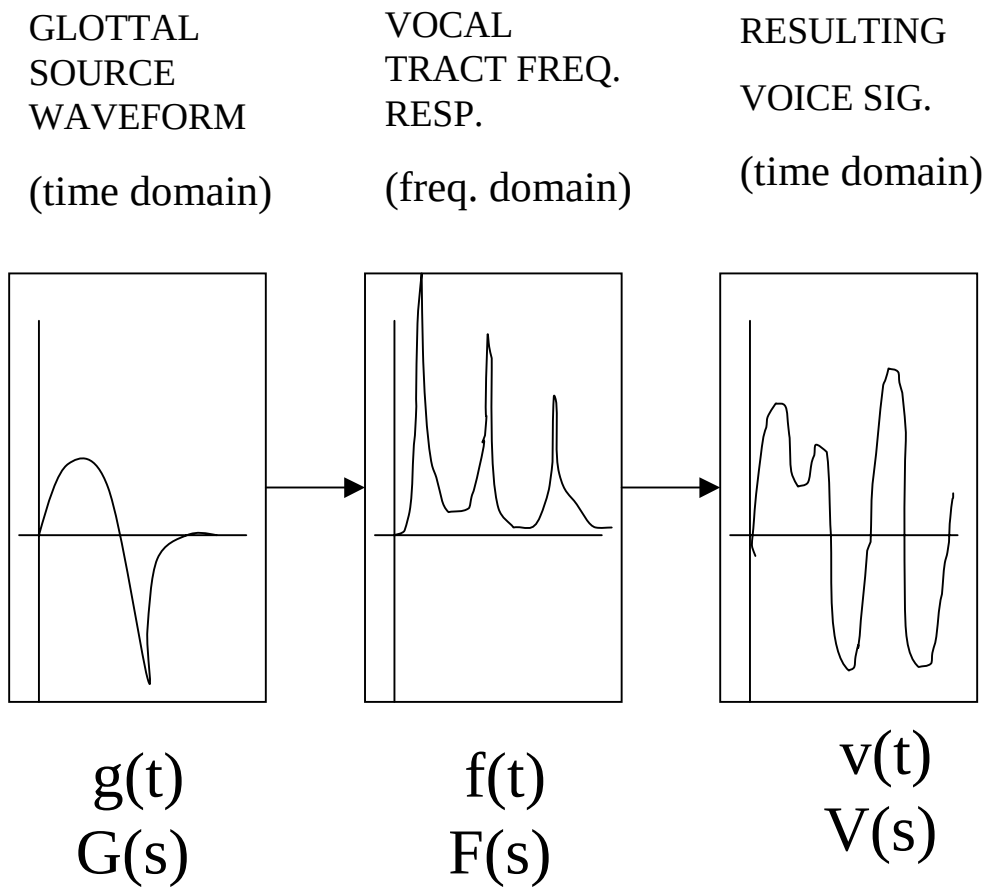


Figure 1.2. Source-filter model for speech synthesis. The glottal waveform is filtered by the vocal tract frequency response to produce the voice signal. In the time domain, the glottal waveform g(t) is convolved with the vocal tract impulse response f(t) to obtain the voice signal v(t). Equivalently, in the frequency domain the transforms of these signals are multiplied to obtain the response. Mathematically, the source and vocal tract are indistinguishable, and changes in one can compensate for changes in the other and still produce an identical voice.

Chapter 2

Analysis of Pathological Voices

The study of pathological voices begins with the digitization and analysis of sample tokens. In this work, the analysis was performed on subjects vocalizing the sustained vowel /a/. Data used in this dissertation include a set of thirty-one pathological voices (randomly selected) representing a range of disorders, age, and gender. Additional voice data of normal voices simulating pathological conditions were also collected. Actual or simulated pathological voice data were then analyzed into the parameters described in Sections 2.1 – 2.5, which were found upon synthesis to provide in most cases a good basis for reproduction of the original sound. All analysis was performed offline using custom software specifically for the analysis of pathological voices and written in C and MATLAB (a commercial software package distributed by MathWorks TM). The steps of data analysis and synthesis are summarized in Fig. 2.1 and discussed in detail in the sections that follow. Briefly, the voice time series is first analyzed for formants (vocal tract resonances), which are then used to inverse filter the time series to obtain the source

glottal flow derivative time series (driving waveform in the source-filter model). The source waveform is then fitted to a predefined waveshape: the LF model. Finally, the nonperiodic features of the time series (nonperiodic FM variations, nonperiodic AM variations, and aspiration noise) are quantified.

Processing begins with collection of voice data. The voices covered a range of disorders and age of both sexes, but excluded bicyclic cases. The pathological voice samples used in this study were collected at 20 kHz by Dr. B. Gerratt at UCLA using a one-inch B&K (Bruel & Kjaer) condenser microphone with a cathode follower preamplifier and antialiasing filter. This setup yields a flat frequency response from about 10 Hz to 20 kHz (the microphone/preamplifier proved not responsive to DC); it was used to record 1 second samples of patients vocalizing the sustained vowel /a/. These samples were then low-pass filtered offline with a FIR filter and decimated to 10 kHz before analysis. Additional normal and simulated pathological voice data were collected by the author for the study of vocal tract identification (Section 3.3) using a professional 0.8 cm condenser microphone element. These samples were digitized at 40 kHz using a sigma-delta A/D (which incorporates the antialiasing function).

The current five-component voice model provides the basis for analysis and synthesis of pathological voices. It includes the following components:

1. Formants. The basic source/filter model is used, which assumes an independent glottal source signal that is filtered with an all-pole resonator model of the vocal tract

2. Source waveform. The glottal source signal, which is obtained in analysis via inverse filtering, is fitted to the LF form [31].

3. Source nonperiodic FM (frequency modulation). Precision fundamental frequency tracking is used to establish both high and low frequency FM effects, which are then included in the synthetic version.

4. Source nonperiodic AM (amplitude modulation). Fundamental frequency pulse energy analysis establishes AM effects, which are then included in the synthetic version.

5. Source aspiration noise. Spectrally-shaped Gaussian white noise is measured in the original voice via cepstral filtering [24] and is added to the synthetic source time series.

In the following (Sections 2.1 – 2.6) the analysis of each of these components is discussed in detail.

2.1 Formant Analysis: LP Analysis and Inverse Filtering

The first step in the analysis of a pathological voice is the determination of formants via the established technique of LP (linear prediction analysis [28]) and application via inverse filtering to determine an estimate of the raw glottal flow derivative time series. This process attempts to perform system identification of both the source time series signal and the vocal tract model system in the source-filter model of speech production by merely using the final output time series of speech; the source/filter model is demonstrated in Fig. 1.2. Mathematically, the source signal and vocal tract system are

just two transfer functions, and once they have been multiplied together to form the output signal, there is no way to segregate them again without additional knowledge. The process of identifying both glottal source time series and vocal tract system is ambitious, and it involves considerable intervention of apriori knowledge and manipulations beyond basic LP. The process has been fairly successful for normal voices, but encounters additional difficulties in the case of pathological voices. In some pathological voices, the method proved ineffective in establishing proper vowel quality, as shown by re-synthesis and perceptual testing, necessitating (sometimes unsuccessful) manual attempts to adjust formants.

The following briefly describes the basic process of LP. The results of LP/inverse filtering are then illustrated for idealized impulsive sources; ideal results are then contrasted with the results of application to more realistic LF (Section 2.2.2) sources. Finally, the apriori strategies and techniques attempting to improve the system identification are outlined. These techniques were collected or established by N. Antonanzas-Barroso and encapsulated by her in the computer program “invf” for source/vocal tract identification [1] at the UCLA Voicelab. This program was used to identify formants and raw glottal flow derivative in the first processing steps in this study (Fig. 2.1).

2.1.1. Basic LP/Inverse Filtering

Briefly, LP performs as follows; see [28] for details. Refer to Fig.2.2; in the figure and the following description, the notation of [28] is used. The basic process of LP attempts to form a P-th order linear FIR (finite impulse response) predictor that takes as its input the previous N samples of system output. That is, given the N previous samples s(n) of a system H, it attempts to predict the next sample. As shown in Fig.2.2, the predictor output s'(n) is subtracted from the system output s(n) to form an error signal e(n). The predictor (alpha's) is derived by minimizing the summed error squared of e(n) over the selected analysis interval; this least squares minimization yields a set of P simultaneous linear equations in the P alpha's which is easily solved. If the unknown system is assumed to consist of an all-pole IIR (infinite impulse response) system, as shown in Fig. 2.2, and the a's of the unknown correspond to the alpha's of the predictor, then the system formed by the predictor and subtractor form a new system A(z) that estimates the inverse of H(z). Ideally, then, it is possible to determine the predictor using LP, form the inverse system A(z) (which contains the formant estimates), and apply it to the original time series s(n) to inverse filter the voice and generate an estimate of the driving function u(n) which corresponds in our case to the glottal flow derivative (compare Fig. 2.2 to Fig.1.2).

Two versions of LP are applied in practice: autocorrelation and covariance. The versions differ only in the assumption made about samples outside the analysis window (range of samples used to calculate the alpha's). The autocorrelation method assumes samples outside the window are zero; it gives rise to a Toeplitz system [28] which is

efficiently solved via Durbin's algorithm. The covariance method does not assume zero values outside the analysis window, and gives rise to a more computationally complex system [28]. In practice, the autocorrelation method is applied to a range of several fundamental frequency periods of voice signal, and a first estimate of A(z) is obtained. The window of analysis is then restricted to what is believed to be the "closed phase" (assumed quiet period of the source), and a more accurate estimate of A(z) is attempted via the covariance method; the assumption here is that the interfering activity of the source dynamics is removed, which permits a better vocal tract estimate. In practice, however, the less impulse-like the source (the greater the portion of the fundamental frequency period that the source has nonzero activity), the less successful LP is in vocal tract identification. The technique is unable to segregate source from vocal tract system using only their convolved result, the output time series s(n). (See also the discussion in Section 3.1).

2.1.2 Idealized LP: Impulsive Sources

LP in both its variants accurately identifies the vocal tract if the source time series is a simple impulse train and the vocal tract transfer function is all-pole. That is, as expected, if the source has no dynamics (z or Laplace transform is unity), LP has no difficulty in assigning all detected dynamics to the only remaining part of the system: the vocal tract. That is, in the case of impulsive sources, there is no ambiguity between the source signal and the vocal tract system, and LP correctly identifies the vocal tract system. The performance of LP on impulsive sources is illustrated in Fig. 2.3 to Fig. 2.5.

A simple vocal tract with four resonators for the vowel /a/ is shown in Fig. 2.3. The impulse train source time series is shown at the top, the pole locations (roots of the resonators) in the middle plot, and the resulting voice output time series is shown at the bottom. For this ideal system, applying the inverse filter, which is simply the FIR filter formed from the reciprocal of H(z) vocal tract, successfully reproduces the impulse train input exactly.

Fig. 2.4 illustrates the application of autocorrelation LP to the impulse train system of Fig. 2.3. Using 10 fundamental frequency periods of the output of this system windowed with a Hamming window yields the pole position shown in the top plot; there is very good agreement with the true poles. Also, the prediction FIR filter, when applied to the output time series s(n), yields an accurate match s'(n) to the original signal, as shown in the middle plot. When the resulting inverse filter A(z) is applied to the output s(n), an accurate approximation to the input impulse train is obtained, as shown in the bottom plot. The application of the covariance method to the impulse train system also produces good results, as shown in Fig. 2.5. Accurate poles, prediction, and inverse filtering are obtained. Here a window length of one fundamental frequency period was used.

2.1.3 Non-impulsive Source Functions

When more realistic glottal source waveforms are employed, however, the results change drastically. The LP now has the impossible task of trying to separate the

convolved source signal and vocal tract system. To illustrate, the system of Fig. 2.6 is constructed. This system is analogous to Fig. 2.3 except the impulse train input is now replaced with a realistic LF waveform (Section 2.2.2); the same vocal tract is used, and the output appears somewhat similar. The results of application of autocorrelation LP are shown in Fig. 2.7; the same sample window and position are used as in Fig. 2.4. As may be seen, however, the pole position now contains significant errors: the higher frequency poles become progressively less accurate, and a pair of real poles replaces one of the complex pairs. The inverse filtered signal now contains large ripples due to incomplete pole-zero cancellation. Only the prediction signal remains accurate.

Application of covariance LP alone results in little improvement. The results of covariance analysis are summarized in Fig. 2.8. In an effort to provide time separation of signal and system (see discussion in Section 3.1), the analysis window is shortened to 40 samples, which is about the length of time during which the source signal is zero (closed phase). The analysis window is then swept through all 100 possible positions and the locus, resulting from plotting the poles at each window position, is shown. As may be seen, the true pole positions are never achieved, and error is more pronounced at the higher frequencies; a pair of real roots often replaces a complex pair. The results of covariance LP using one of the more optimal window positions is shown in Fig.2.9 (compare to Fig.2.7). Again, poor pole accuracy results in formant ripple in the inverse filtered signal, which vaguely resembles the actual LF input; at least two of the uncanceled pole frequencies are clearly visible in the error signal. Prediction remains accurate.

2.1.4 Apriori Strategies

Several techniques have proven useful to improve the performance of LP, thus enabling more accurate separation of the glottal source signal. These methods allow the user to add known characteristics of vocal system to the simple least squares algorithm of LP.

1. Real roots are discarded. The vocal tract is assumed to consist of an all pole resonator model with no real poles.

2. Complex poles resulting from the covariance method occurring outside the unit circle are reflected back inside the unit circle by taking their reciprocal. It is known that the covariance method is not guaranteed to produce stable roots [25], while the vocal tract is a stable, passive system. The success of this strategy is shown in Fig. 2.10, which repeats the system of Fig.2.9. In the top plot the poles for a window position in the closed phase are shown; they all lie outside the unit circle. Reflecting them inside the unit circle places them into almost exact agreement with the true poles. The resulting inverse filtered signal (bottom) agrees well with the LF input, and prediction (after re-scaling) remains fairly good.

3. The length of the covariance analysis window can be varied in an effort to match a closed phase. Many pathological voices contain few if any quiet periods, however. The location of the closed phase (or closest approximation to a closed phase in the case of some pathological vowels) is chosen at the maxima of the residual.

4. The two poles per kilohertz rule is used to start analysis, but order of the system can then be optimized.

5. Provision for manual pole placement is made in the analysis software. Sometimes formant ripple in the inverse filtered signal can be reduced by manually tweaking pole positions. However, because of the high number of degrees of freedom (2 dimensions for each pole), this process is difficult to converge. Knowledge of the spectral tilts of typical voice signals can also give clues to improved pole placement.

6. A final resort is attempting to place formants by SABS (subjective analysis by synthesis). One of the synthesizers (Chapter 4) is used to compare the vowel quality of original and synthetic voices. However, experience shows this process is even more difficult than #3 above.

Formant analysis for pathological voices provides added problems over normal voices, invalidating assumptions made in normal LP. The source waveform may not be impulsive and lack quiet periods, making it impossible to apply covariance to separate signal and system. The source may be highly irregular from pulse to pulse, with the period and shape of each source pulse varying. For these and other reasons, application of LP to pathological voices is sometimes very difficult. External stimulation of the vocal tract (Chapter 3) may provide a means to faster and more accurate formant identification.

2.2. Fitting of Inverse Filtered Source

Having generated the raw inverse filtered source waveform, the next step in voice analysis is fitting the source to a mathematical model with adjustable parameters. The parameter values provide multiple benefits: They provide a means of comparing and rating voices and studying linguistic and voice quality effects. Furthermore, the parameter values are later used as input control values to both the hardware and software synthesizers. Several types of source models were investigated: filtered pulses, parabolic, and the LF model. The LF model had a greater ability to model the wider range of wave shapes found in pathological vowels. The hardware synthesizer (Section 4.1) was used to generate synthetic tokens of the vowel /a/ using each of these source models. Ultimately, a modified version of the LF model proved to be the most useful for analysis and synthesis of pathological vowels and was implemented in the software synthesizer (Section 4.2).

2.2.1 Simple Models

A variety of mathematical models have been used to fit the inverse filtered glottal flow derivative pulse [22,31] to match the wide range of waveforms found in pathological voices. These include simple impulses, impulses filtered by a variety of lead-lag filters, and polynomial fits. Simpler models have been shown to have shortcomings. For example, a two pole filtered impulse (which incorporates just two degrees of freedom) is shown by Deller [7] to have limitations in achieving the OQ (open quotient) vs SQ (speed quotient) relationship found in the source waveforms of even normal voices. Chasaide and Gobl [4] also show filtered impulses to have limitations: Their shape is time-reversed

with respect to the normal pulse. That is, the large rate of change is on the rising edge rather than the falling edge.

One of the more useful simple models used in the early synthesis efforts with the hardware real-time synthesizer (Section 4.1) is the KGLOTT88 parabolic model [22]. The equations describing this model are shown in Fig. 2.11, and an example plot of a KGLOTT88 glottal flow pulse and its derivative are shown in Fig. 2.12. This model significantly improved perceptual fidelity over simpler filtered impulses.

2.2.2 LF Model Characteristics

Our earlier studies with a real-time synthesizer employing a simple parabolic source model showed considerable improvement in perceived fidelity when the simplified LF source model [31] was implemented [16]. The simplified LF model, including our modifications, is shown in Fig. 2.13. This model elaborates on the idea of using simple algebraic and trigonometric curves by employing an exponentially modified sinusoid for the first part (opening phase) of the glottal flow derivative pulse and a decaying exponential for the second part (closing phase). The result appears remarkably similar to actual flow derivative pulses of normal voices.

As a result of experience processing considerable amounts of voice data, we found it useful to use a set of physically observable parameters to define the pulse to be fitted rather than the actual LF parameters; the actual LF parameters can then be derived via the solution of simultaneous equations, in a manner similar to that used by Qi [31]. Our

selection of observable parameters, which completely define the modified LF curve, results in a set of four values with an optional constraint.

tp = time of zero crossing

te = time of negative maxima

Ee = value of negative maxima

t2 = time of 50% on closing phase

m = linear coefficient of closing phase (OPTIONAL CONSTRAINT)

The first four values can be determined automatically from the time series. The optional fifth value *m* is calculated when the constraint is added that the glottal flow derivative pulse must return to zero. The original LF model contains the parameter *ta*, which is a measure of the return phase decay rate; it is the time interval from *te* to the intersection with the time axis of the tangent at *te*. The *ta* parameter in the original LF model is functionally replaced here with *t2*.

This model seems to allow considerable flexibility to adjust to the wider variety of pulse forms found in pathological voices. For example, for breathy voices, the glottal flow derivatives we found had almost sinusoidal form; by extending *tp* (time of zero crossing) to larger values the model waveshape becomes more sinusoidal.

2.2.3 Modified LF Model Calculation

In operation, calculation of the LF model proceeds as follows:

1. A typical glottal flow derivative pulse is selected from the raw inverse filtered time series (Fig. 2.14). For this study, the waveform of the glottal flow derivative is assumed to be fixed for the entire one-second voice sample. Since time varying source pulses are not currently modeled, the selected pulse is chosen to be a representative one if there is variation present.

2. The major features described in Section 2.2.2 (shown in Fig. 2.15) are automatically determined from the data. Numerical methods are applied to obtain maxima, zero crossings, etc from the raw inverse filtered glottal flow derivative time series. The equations shown in Fig. 2.16 are then used to obtain the initial set of LF parameter values. The LF curve obtained is plotted and visually checked against the original raw flow derivative pulse, as shown in Fig. 2.17. This step was almost invariably successful.

3. The major features determined in Section 2.2.2 are then used as the starting point for least squares optimization using the MATLAB function 'fmins' which uses the simplex search method [26]. Starting at a reasonable first approximation resulted in much quicker convergence and reduction in the chance of spurious solutions (local minima). The summed error squared between the flow derivative pulse and the fitted LF curve is minimized by varying the four major feature parameters *tp, Ee, te,* and *t2* to their optimal values. An optional constraint was provided to force the fitted flow derivative pulse to zero; this results in an added equation for the parameter *m* (see Fig.2.13). The optimized major feature parameters are then converted to the LF parameters using the equation set shown in Fig.2.18. The first three equations, unfortunately, form a simultaneous nonlinear

set. However, by iteratively solving them in the proper sequence shown, each equation is solved for its variable of greatest effect and the process rapidly and reliably converges. The steps of data collection through LF fitting complete the periodic analysis of the voice signal, and they are summarized in the first column of Fig. 2.1.

2.3 Analysis of Fundamental Frequency Variations

Fundamental frequency variation is one of the most perceptually important acoustic measures of voices. Unlike other measures such as aspiration noise, small variations in fundamental frequency are not ignored. Efforts to provide ever increasing levels of synthesizer fidelity led to a successful approach to modeling fundamental frequency variation in the pathological /a/ vowel samples collected.

2.3.1 Fundamental Frequency Analysis Approach

In order to parameterize fundamental frequency variation for accurate synthesis, fundamental frequency variations are tracked in the time domain via interpolation to the precision permitted by the data and then broken down into low frequency (tremor) and HFPV (high frequency fundamental frequency variations). Previous investigations have used a variety of measures of HFPV, usually some type of average [3]. The current approach refines HFPV analysis and synthesis by modeling the variation in the fundamental frequency period as a Gaussian distribution: the frequency deviation

excursions of HFPV are seen to be well-modeled by a Gaussian distribution in cases of successful tracking on short intervals of sustained /a/. In order to verify successful fundamental frequency tracking and aid later NSR (noise to signal ratio) calculations (Section 2.5), the original pathological voice is re-sampled to remove fundamental frequency variations. In summary, the following steps are performed:

1. Precision interpolating fundamental frequency tracking in the time domain.

2. Segregation of the fundamental frequency track into high and low frequency variations.

3. Gaussian modeling of the statistical properties of the high frequency fundamental frequency variations.

4. FM Demodulation of the original voice using measured fundamental frequency variations.

2.3.2 High Resolution Fundamental Frequency Tracking

Time domain fundamental frequency tracking is carried out over the entire original voice sample on a pulse by pulse basis, so that the short duration pulse period variations (HFPV) may be captured. Pulse period lengths are established by measuring the time interval between similar features in each pulse, such as maxima or minima points. For additional information, see Milenkovic [27] which describes application of correlation techniques to determine HFPV. In addition, Deem, Manning, Knack, and

Matesich [6] describe a process for determining period lengths in the time domain. Fundamental frequency tracking allows use of any of four signals for maxima/minima detection for cycle marking: original voice, glottal source, smoothed derivative of original voice, or smoothed differentiated glottal source. In difficult cases in which automatic loses tracking lock (fails to track accurately), the user may manually mark features on as many pulses as necessary to reestablish tracking lock. Interpolation between samples is used to determine fundamental frequency periods to less than one sample period (< 0.1 ms): a parabolic fit with a user-selected number of points is applied to the sample points surrounding the maxima/minima, and the "true" minima/minima is calculated from the fitted parabolic vertex. For some parts of the analysis, upsampling from 10 kHz to 40 kHz improves performance of the tracking algorithm. Upsampling is performed using the MATLAB (a commercial software package distributed by MathWorks TM) function “interp,” which uses symmetric filtering and minimizes the mean square error of the interpolated points [20]. Successful tracking is characterized by the absence of discontinuities in the resulting fundamental frequency versus time plot and Gaussian distribution in fundamental frequency period variations with standard deviation in the expected range of 0 to 1 percent. Fig. 2.19 illustrates automatic pulse feature selection for a typical signal; in this case, the selection of original voice minima yielded successful automatic tracking. Fig. 2.20 illustrates the resulting typical high-resolution fundamental frequency time series generated from the interpolating tracker.

2.3.3 Analysis into HFPV and Tremor

In order to analyze the voice sample for FM characteristics, the fundamental frequency time series is examined for both its frequency content and statistical features. Two fairly distinct types of fundamental frequency variation (FM modulation) are seen: low frequency (<10 Hz) changes associated with tremor, and high frequency (>10 Hz) cycle to cycle variation associated with HFPV. The selection of the 10 Hz cutoff is arbitrary, but seems to give satisfactory results when synthesized tokens are studied in perceptual tests. The tremor variations are perceptually associated with an unsteady voice, while HFPV gives rise to the perception of roughness. The high and low frequency components of the fundamental frequency track are respectively segregated into HFPV and tremor time histories using high and low pass filters with a cutoff frequency of 10 Hz respectively. High and low frequency components of the fundamental frequency track are separated in the frequency domain using the windowing technique on the Fourier transform, thus taking advantage of using the entire fundamental frequency time series offline (non-real-time). Fig. 2.21 displays the result for the fundamental frequency track of Fig. 2.20.

2.3.4 Gaussian Modeling of HFPV

Once tremor has been removed from the fundamental frequency track, the remaining higher frequency variations appear to be well modeled by a Gaussian distribution. To verify the success of fundamental frequency tracking, the statistical distribution of HFPV is displayed by histogramming the frequency deviation of the high

pass filtered fundamental frequency time series. Fig. 2.22 displays the histogram for the same voice signal as Fig.2.20. Successful tracking is characterized by a Gaussian distribution with one standard deviation of usually less than 1 or 2 percent. Bimodal distributions and large values of deviation are almost always associated with loss of lock in the fundamental frequency tracking algorithm. The measured standard deviation of fundamental frequency period, in units of percent, is also used to define the level of HFPV for later input to the synthesizer.

2.3.5 FM Demodulation

In order to investigate the effects of frequency modulation on measurement of aspiration noise (Section 2.5), the original voice time series is resampled to create versions of the original voice with (a) low frequency (tremor) removed and (b) all frequency variations removed. This is a refinement of the approach used in [14]. Using the fundamental frequency time series, a vector of unevenly spaced re-sampling times is created by making the instantaneous sample interval inversely proportional to the instantaneous fundamental frequency frequency. Simple linear interpolation of the original time series on this modified time vector creates a version of the original voice where all fundamental frequency periods are forced to be the same length; that is, fundamental frequency variations are removed. To remove only low frequency fundamental frequency variation, the low-pass filtered fundamental frequency time series is used to generate the modified time vector; to remove all fundamental frequency

variation, the unfiltered fundamental frequency time series is used. In order to verify the success of this FM demodulation, fundamental frequency tracking is re-performed on the re-sampled voices. Fig. 2.23 is an example of the effect of removing the tremor from the original voice. Note that the low frequency fundamental frequency variations in the re-sampled voice are greatly reduced. Fig. 2.24 shows the results of repeating fundamental frequency tracking on a successful re-sampling of the original voice to remove all fundamental frequency variation. The upper plot shows that fundamental frequency period variation has been reduced to less than 0.2 Hz, and the frequency is essentially constant at 266.4 Hz, as shown by the points clustered about this frequency. The lower plot shows that HFPV is less than 0.05%. Success of the process is also verified by listening to the re-sampled time series. Tremor removal creates a much steadier sounding token. Removal of all FM creates a token sounding almost synthetic in its stability. (Interestingly, however, removal of all fundamental frequency variation in some cases still leaves voice quality variations synchronous with the original tremor variations; these may be due to formant modulation or other effects.)

2.4 Analysis of Fundamental Frequency Pulse Power Variation

In a manner analogous to fundamental frequency variation, power variations in the fundamental frequency pulses of the vowel /a/ are analyzed. This approach uses a similar treatment of both the AM and FM variations.

2.4.1 Power Analysis Approach

Power analysis seeks to quantify the variations in amplitude, power, and energy within the pathological vowel sample. A rationale for measurement is constructed: the boundaries of fundamental frequency pulses are located with the aid of the results from fundamental frequency tracking (Section 2.3.2). Having defined the pulses, it is easy to calculate various measures of amplitude, energy, and power. Power is selected for subsequent analysis, since it is probably highly correlated with perception. As with fundamental frequency, power variations are segregated into high and low frequency phenomena. The high frequency variations (shimmer) are found to be Gaussian. Removal of AM variations from the original voice provides verification of processing and permits testing the effects of AM on NSR (Section 2.5.5).

2.4.2 Power Tracking

Analysis of the original pathological voice continues after fundamental frequency tracking with analysis of the amplitude, energy, and power of fundamental frequency pulses. The results of fundamental frequency tracking are used as a starting point for

estimating the maximum amplitude, sum of the absolute value of samples, energy, and power of each fundamental frequency pulse.

The first step of amplitude analysis is segregation of the fundamental frequency pulses of the original time waveform. The set of features within each cycle described in fundamental frequency tracking, such as pulse maxima, is assumed to exist between minima of each fundamental frequency pulse. The adjacent minima of the signal's amplitude envelope to either side of this center are used to define the boundaries of the fundamental frequency pulse. The set of features within each cycle described in Section 2.3.2, such as pulse maxima, are assumed to exist between minima of each fundamental frequency pulse. The adjacent minima of the signal's amplitude envelope to either side of the tracking feature are used to define the boundaries of the fundamental frequency pulse. The intent is to separate pulses by placing pulse boundaries so that the maxima of voice power in each pulse occurs between tracking features of the pulse and the minima of power occurs near the boundaries. This provides a natural separation of normal voice pulses and a reasonable approximation for pathological voices. Minor variations in the selection of pulse boundaries have little effect on the remainder of the power analysis.

Analysis proceeds via generation of the envelope of absolute value and power of the original voice. Starting at the features used for fundamental frequency tracking, the corresponding power envelope minima before and after the feature are located. For each fundamental frequency track feature there exists a power minima, so the resulting power minima instants (time values at the minima) are interspersed between the fundamental frequency tracking feature instants (time values at the features). The envelope minima

thus determined form a natural boundary for fundamental frequency pulses. Fig. 2.25 displays the absolute value of a short segment of a voice time series showing the fundamental frequency track maxima and the pulse boundaries (envelope minima) selected by the algorithm. In practice, power tracking proved to be a far easier task than fundamental frequency tracking; no manual interventions or alternative approaches were ever required

Having defined pulse boundaries, four measures of pulse strength are calculated: maximum amplitude, signal average, indicated by sum of absolute value, energy, and power. The maximum amplitude is the absolute value of the greatest extent (plus or minus) of the original voice samples within the pulse (between the pulse demarcations.) The sum of the absolute value is the addition of the absolute value of all the samples within the pulse. The energy is the sum of the squares of all samples in a pulse. The power is the energy divided by the number of samples within the pulse. Generally, all four measures track each other. Power was selected as the most useful, as it is probably most closely correlated with perceived signal strength. Fig.2.26 displays power for the same signal shown in Fig. 2.25.

2.4.3 Power Time Series Analysis

In a manner analogous to the fundamental frequency time series analysis, the original voice amplitude (AM) features are analyzed. Pulse power variations are now analyzed into a low frequency power time series track and a high frequency shimmer

measure. The power measure described in Section 2.4.2 is selected as the basis of this analysis over the other three measures, since it should be most closely related to the perceived signal level. A cutoff frequency is selected (usually 10 Hz), and a low pass FIR filter is constructed and applied to the power time series. The resulting low passed signal defines the low frequency power time series. The difference between the original power time series and the low frequency power time series defines the shimmer time series; the standard deviation of the shimmer time series estimates the amount of shimmer present. Fig. 2.27 illustrates the high and low frequency power variations. In the upper plot of Fig. 2.27, the low pass filtered power time series is shown as a dotted line; the original power time series, deviating above and below the low pass filtered signal, is plotted on top with a solid line. The original time series power closely follows the low pass filtered version, differing only by small high frequency variations. In the lower part of Fig. 2.27, the difference between the high and low pass filtered signals is plotted; it is essentially equal to the high pass filtered signal. Specific hallmark peaks in Fig. 2.27 illustrate this; the negative-going shimmer peaks in the top of Fig. 2.27 appearing at about 0.12 and 0.32 sec also appear directly below in the lower plot. A positive peak at about 0.34 sec also appears in both plots.

2.4.4 Gaussian Modeling of Shimmer

In the same manner as HFPV, the high frequency power variations are histogrammed, as shown in Fig.2.28. Again, a Gaussian distribution seems to model the high frequency power variations for our quasi-steady vowels. The standard deviation of

the distribution, measured in percentage power variation, provides a measure of shimmer. Power variations were usually larger than HFPV, typically a few percent. Unlike HFPV, bimodal or other unusual distributions signifying problems with power tracking were seldom seen, as power tracking was a much more robust process. Again, measured shimmer values were passed on to the synthesizer software for use in perceptual studies.

2.4.5 AM Demodulation

In a manner analogous to FM demodulation, analysis continues with AM demodulation. The AM variations may be minimized in the original voice by rescaling the samples in each fundamental frequency pulse in the time series such that all fundamental frequency pulses exhibit the same power level. Using the data collected in AM analysis, a scaling factor is calculated for each pulse to normalize its power level to the mean power level measured over the entire original voice one second sample. The resulting demodulated time series exhibits constant power over the 1-second sample. The modified time series is resubmitted to the AM analysis software to verify successful demodulation. Fig. 2.29 displays the result of a successful demodulation; note that the measured power level of the demodulated voice is constant. The other measures (sum, maxima, or energy) are not necessarily held constant, but are approximately constant. Informal listening by the authors of the AM demodulated waveform verifies that apparent changes in perceived volume level have been removed. The analysis steps of AM and FM

demodulation complete the first stage of the nonperiodic analysis, and are represented in the top center area of Fig. 2.1.

2.5 Analysis of Source Aspiration Noise

Having accounted for two known sources of nonperiodicity in the original voice, analysis continues with estimation of the ratio of nonperiodic energy in the original voice to periodic energy (NSR). Another source expected to account for the major part of the remaining nonperiodic energy in most pathological voices is aspiration noise generated by turbulent airflow through the glottis. (Possible additional remaining sources not yet measured are variation in the glottal waveform and variation in formants.) As with the flow derivative, this aspiration noise is spectrally shaped by the vocal tract. Both the ratio of nonperiodic to periodic energy (NSR) and the spectral shape the aspiration noise are approximated.

2.5.1 Noise Measurement Approach

Measurement of the aspiration noise component of a voice is performed using a modification of the cepstral notch filtering technique described in [24]. The original de Krom algorithm generated SNR (signal-to-noise ratio) in voices as a function of frequency; cepstral filtering and spectral peak envelope tracking were used to arrive at noise spectra and SNR functions The assumption in this work is that the nonperiodic energy measured and reflected in the NSR is due mainly to aspiration noise. A limitation

of this approach is the tradeoff between the inaccuracies due to the windowing effects of a short time sample and the instabilities (which include AM and FM) of the voice in a long time sample. Both of these effects tend to increase NSR to levels significantly above those set in SABS tests in which the listeners match synthetic voice samples to the original by varying the parameters of the synthesizer. The current approach removes AM and FM variations (Sections 2.4.5 and 2.3.5) before the cepstral noise measurement, thus allowing long samples to reduce windowing effects but maintaining stability of the voice over the longer sampling interval. This results in more accurate aspiration noise level estimates which are much closer to those set in the SABS tests. The effects of removal of FM and AM of both high and low frequency were tested individually and in combinations. Changes in measured NSR of as much as 12 dB were observed. The effect of removal of FM components was observed to be much greater than that of AM components. Effectively, the process of removal of AM and FM components from the natural voice helps to bridge the gap between natural and synthetic voices in a stepwise approach. Removal of each component renders a result that sounds progressively more like the synthetic version, and it aids in the identification of any unaccounted remaining differences between the original and synthetic voices.

As part of the cepstral noise analysis [24], the power spectrum of the glottal source aspiration noise component is estimated. After removal of the periodic signal from the voice sample, a spectrum of the non-periodic components is produced, which is then inverse-filtered to remove the vocal tract effects, yielding the source nonperiodic

spectrum alone. The shape of this spectrum is usually flat (without significant systematic variation of power with frequency) to within a few dB; a spectrum that is not flat would indicate aspiration noise with more complex frequency content. The spectrum is captured in a 25 point piecewise linear approximation.

2.5.2 Cepstral NSR Calculation

The relative amounts of nonperiodic and periodic energy in the original voice are estimated using the basic cepstral comb-liftering approach outlined in [24]. The following steps refer to Fig. 2.30, which summarize successive steps in the NSR analysis for one typical pathological voice (male, fundamental frequency variation from 118 Hz to 124 Hz, NSR at –24dB). First, the spectrum of the original voice is generated, as shown in Fig. 2.30a; the periodic components appear as the usual harmonic peaks. The spectrum is transformed to the cepstrum, as shown in Fig. 2.30b; now the periodic components from the spectrum are further condensed into the peaks of the cepstrum (eg., the large peak at 0.008 sec in Fig. 2.30c). Fig. 2.30c provides an expanded view of the area with highest periodic energy, at low cepstral time range. Using autocorrelation, the fundamental frequency is estimated and used to construct a periodic notch filter (comb-lifter) to apply to the cepstrum to remove the periodic components. Application of the comb-lifter to the cepstrum removes the periodic energy in the peaks, as shown in Fig. 2.30d, which shows its application to the cepstrum of Fig. 2.30c. (Note: the width of the cepstral notch seemed to have minimal effect on NSR measurement, so long as it is wide enough to

encompass the peak, since the spectral density of the noise component is much lower.) After removing the periodic components, transformation back to the spectrum yields the spectrum of the nonperiodic components alone. Fig. 2.30e re-displays the original spectrum and the resulting noise (nonperiodic) spectrum, which appears immediately below and partly overlapped with the original spectrum. Also shown in Fig. 2.30e is the magnitude plot of the vocal tract determined independently from the LPC analysis. (Successful analysis of stable voices is reflected in matching shapes in the spectra, as all three should contain the vocal tract shaping). At this point, Parseval's rule (which simply states that summed signal energy is the same in both the frequency and time domains) is applied to calculate NSR. The energy in the noise spectrum is simply summed and subtracted from the original spectrum energy to yield periodic energy, and hence the ratio of nonperiodic to periodic energy in the original time series.

2.5.3 Source Noise Spectrum

Finally, an estimate of the spectral shape of the source noise alone is calculated. In order to generate the spectrum of the source alone, the vocal tract log magnitude frequency response is generated from the all-pole filter used to model the vocal tract (generated in inverse filtering) and subtracted from the nonperiodic spectrum. The resulting source nonperiodic spectrum is smoothed in the frequency domain with a 100 point triangular window moving average filter. The process is completed by fitting the remaining spectral shape with a 25-point piecewise-linear model. In most cases, the

resulting spectrum is fairly flat. The final source spectrum and its fit are shown in Fig. 2.30f.

2.5.4 Effect of FM on NSR Measurement

The cepstral filtering noise analysis [24] implicitly assumes the signal being analyzed is stable in frequency. If there is variation in the fundamental frequency over the sample analyzed, the harmonic peaks of the spectrum and cepstrum will be broadened significantly compared to the width of the notch filter used to remove them. For example, if a 1ms cepstral notch width is used with a voice with fundamental frequency of 100 Hz, it only requires a 5% change in fundamental frequency to completely displace the rahmonic peak out of its intended notch, thus completely corrupting the NSR measurement. Smaller fundamental frequency changes would still significantly affect the NSR calculation.

The setting of the notch width of the comb lifter is a tradeoff. Wider notches remove “rahmonic peaks” that have been widened by both the variations in fundamental frequency over the sample and the windowing effect of short duration time samples. However, wider notches also remove a greater portion of the background nonperiodic energy underlying the notch, giving rise to lower than actual estimate of nonperiodic energy. Narrow notches fail to completely remove rahmonic peaks widened by the fundamental frequency variation and windowing, resulting in periodic energy being

included in the nonperiodic estimate and thus giving rise to a higher than actual estimate of nonperiodic energy.

An analogous tradeoff exists with sample duration. Short samples cause widened cepstral peaks. Longer samples alleviate the effect of windowing, but peaks may widen due to fundamental frequency variation (tremor and HFPV) over the duration of the sample. The ideal solution is a long, fundamental frequency stable sample, allowing narrow cepstral notch filtering and producing a more accurate NSR estimate. Thus, using the FM demodulated original voice (Section 2.3.5), it should be possible to improve the accuracy of the NSR estimate and hence the agreement between the SABS listener-set and the measured aspiration noise levels. This turns out to be the case. Fig. 2.31 shows the relative spectral shapes of original (top curve) and demodulated signals, and it illustrates the effect of resampling. The three curves are offset vertically by an arbitrary amount for clarity. The peaks associated with harmonic energy are narrowed by tremor removal (middle curve); removal of all FM narrows the peaks further (bottom curve). The resulting reduction in spillover from the cepstral notch filter reduces the NSR measurement from the original –8.5 dB, to –13.9 dB with tremor removed, and further to –18.3 dB with all FM removed. The improved NSR estimate more closely agrees with the listener-set aspiration noise level of –22dB for the synthetic version in the SABS experiments (Section 5.2).

2.5.5 Effect of AM on NSR Measurement

In order to test the effect of removal of AM variations on the NSR estimate, the AM demodulated original voice was also re-analyzed. Removal of AM variations does not seem to affect harmonic peak widths and NSR estimates nearly as much as removal of FM variations. In most cases, the effect is under 1 dB. (Section 2.5.6). In general, amplitude modulation with a random signal at the low levels characterized by volume and shimmer variation in a pathological voice would not be expected to vary harmonic peak width nearly as much as frequency modulation.

2.5.6 NSR Measurement Results

In order to test the affects of various combinations of tremor, HFPV, and AM demodulation, six versions of the original voice are created:

1. Unmodified original voice.
2. Original voice with AM demodulation
3. Original voice with FM tremor demodulation
4. Original voice with FM tremor and AM demodulation
5. Original voice with complete FM demodulation
6. Original voice with complete FM and AM demodulation

The NSR value, (here assumed to be entirely due to aspiration noise) is calculated for each version of 31 pathological voices. The result is plotted in Fig. 2.32. In the figure, the cepstral NSR values for each of the six versions are plotted versus case number, with the case numbers sorted in order of ascending NSR of the unmodified original voice. The

NSR values of each version are connected with line segments for ease of comparison. Several effects clearly emerge:

1. The process with the largest and most consistent effect is tremor removal, with cepstral NSR reductions in the 5 to 10 dB range. (Note the decrease from the top curve pair to the middle curve pair in Fig. 2.32).

2. Removal of HFPV provides an additional decrease in cepstral NSR of 1 dB to 5 dB, with an average of about 2dB. (Note the decrease from the middle curve pair to the bottom curve pair in Fig. 2.32).

2. The incremental decrease in cepstral NSR due to AM demodulation is very small (usually less than 1 dB). (Note the small difference between each of the two members within each of the three main curve pairs).

2.6 Summary

The steps of the analysis phase of voice processing have been detailed. Using the source-filter model of voice production as a basis, voices have been parameterized into formants, LF source waveform, fundamental frequency time history, amplitude time history, FM nonperiodic effects, AM nonperiodic effects, and aspiration noise. The limitations and practical aspects of LP analysis for inverse filtering and the determination of formants have been described. Source waveform fitting using the LF model has been

detailed. The nonperiodic features of pathological voices have been expressed in terms of AM and FM variations and aspiration noise. Gaussian distributions have been shown to model the small, high frequency effects in AM and FM variations. Aspiration noise has been found to be well modeled with spectrally shaped Gaussian noise.

The results of analysis form a set of parameters that may be used both for comparison of voices and for the generation of synthetic versions for perceptual testing. The following Chapters treat the subject of synthesis of pathological voices.

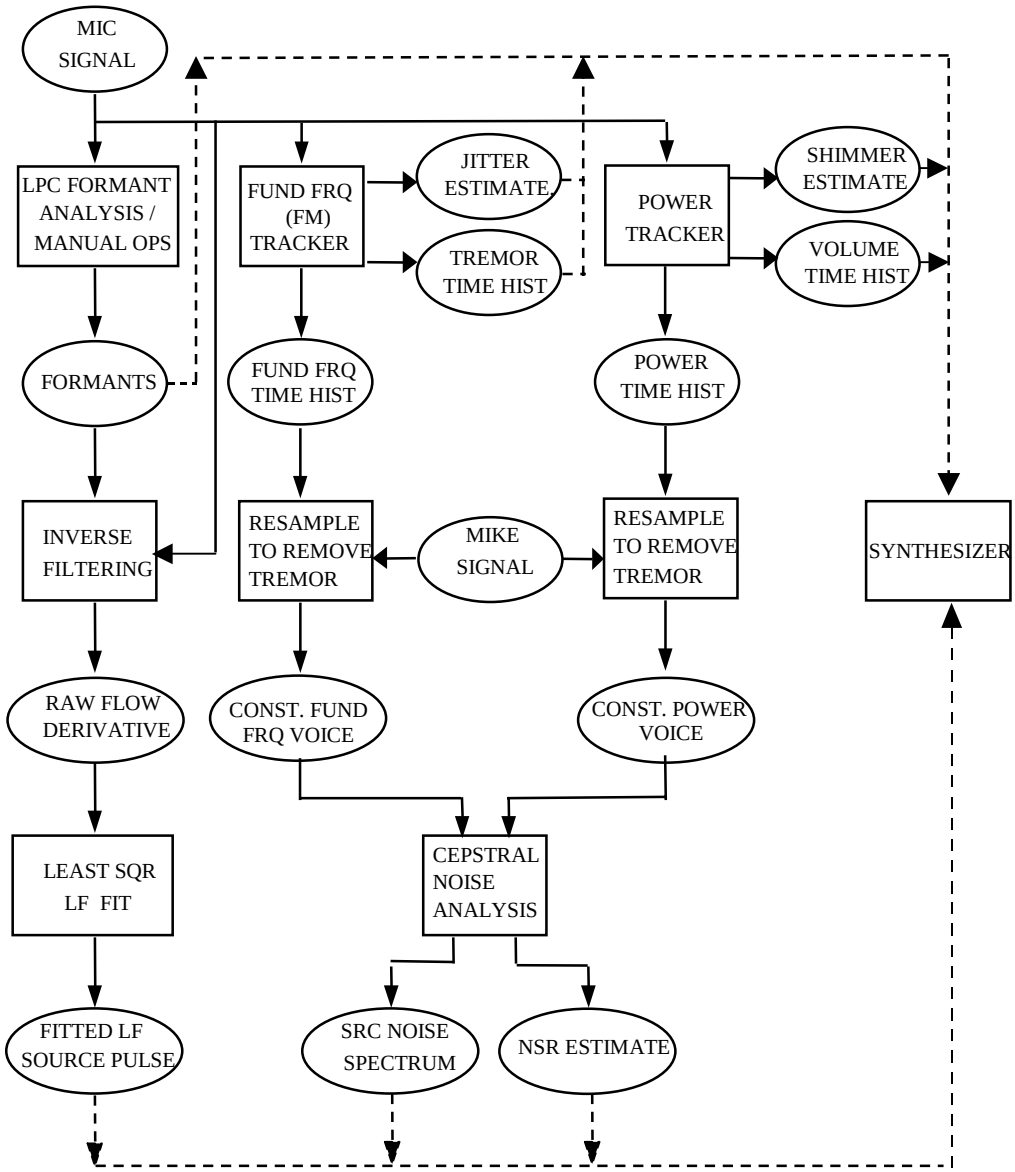


Figure 2.1. Overall voice analysis/synthesis steps. Details are omitted for clarity. Analysis (solid lines) steps include LPC formant analysis, inverse filtering, source fitting, fundamental frequency tracking, power tracking, jitter estimation, shimmer estimation, and source aspiration noise estimation. Results are input to the synthesizer (dashed lines).

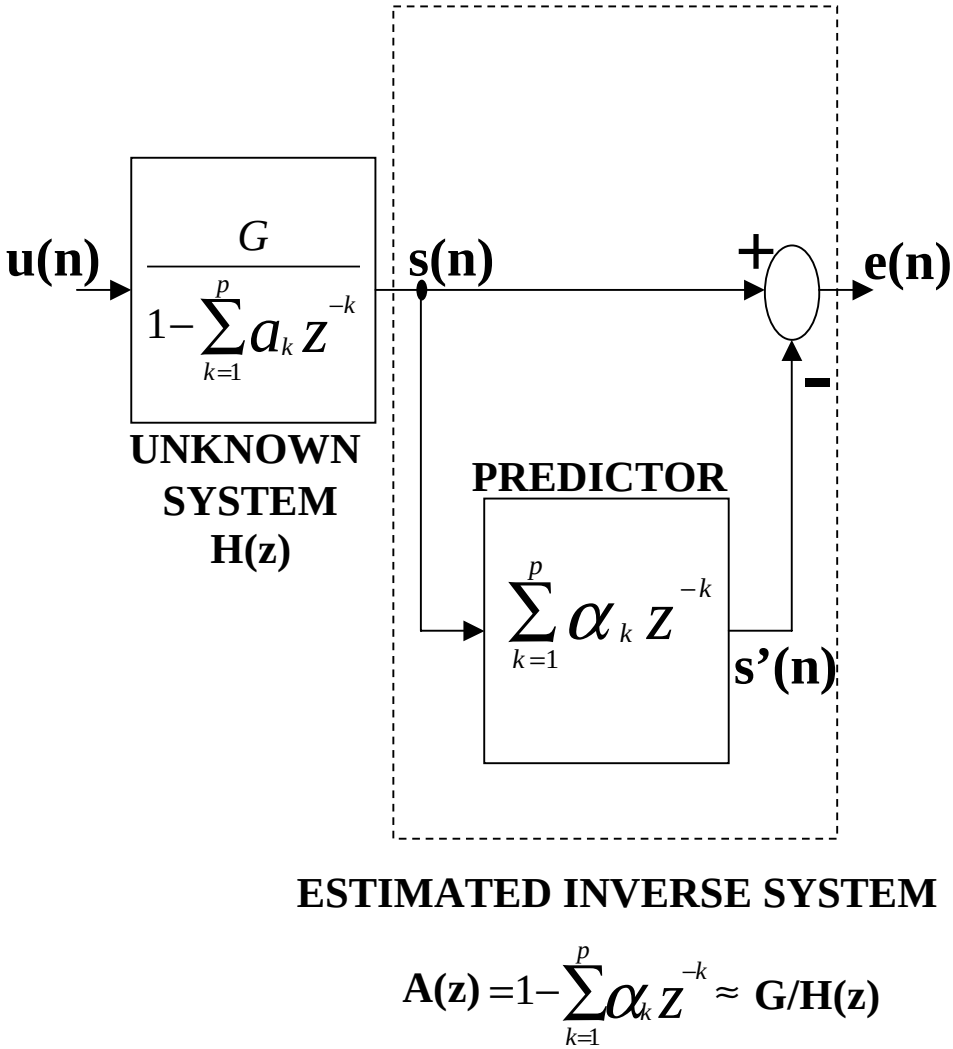


Figure 2.2. Block diagram of the process of LP. An unknown system H(z), assumed to be modeled by an all-pole resonator chain, is driven by a signal u(n). Only the output s(n) is observable. A predictor FIR filter is derived via least squares minimization between s(n) and the predictor output s'(n). The predictor is used to form A(z), which approximates an inverse system to H(z), and generates the error signal e(n). If u(n) is impulsive, A(z) models H(z) well. If, however, s(n) incorporates dynamics from both vocal tract and source signal, A(z) is unable to distinguish source from signal via LP alone.

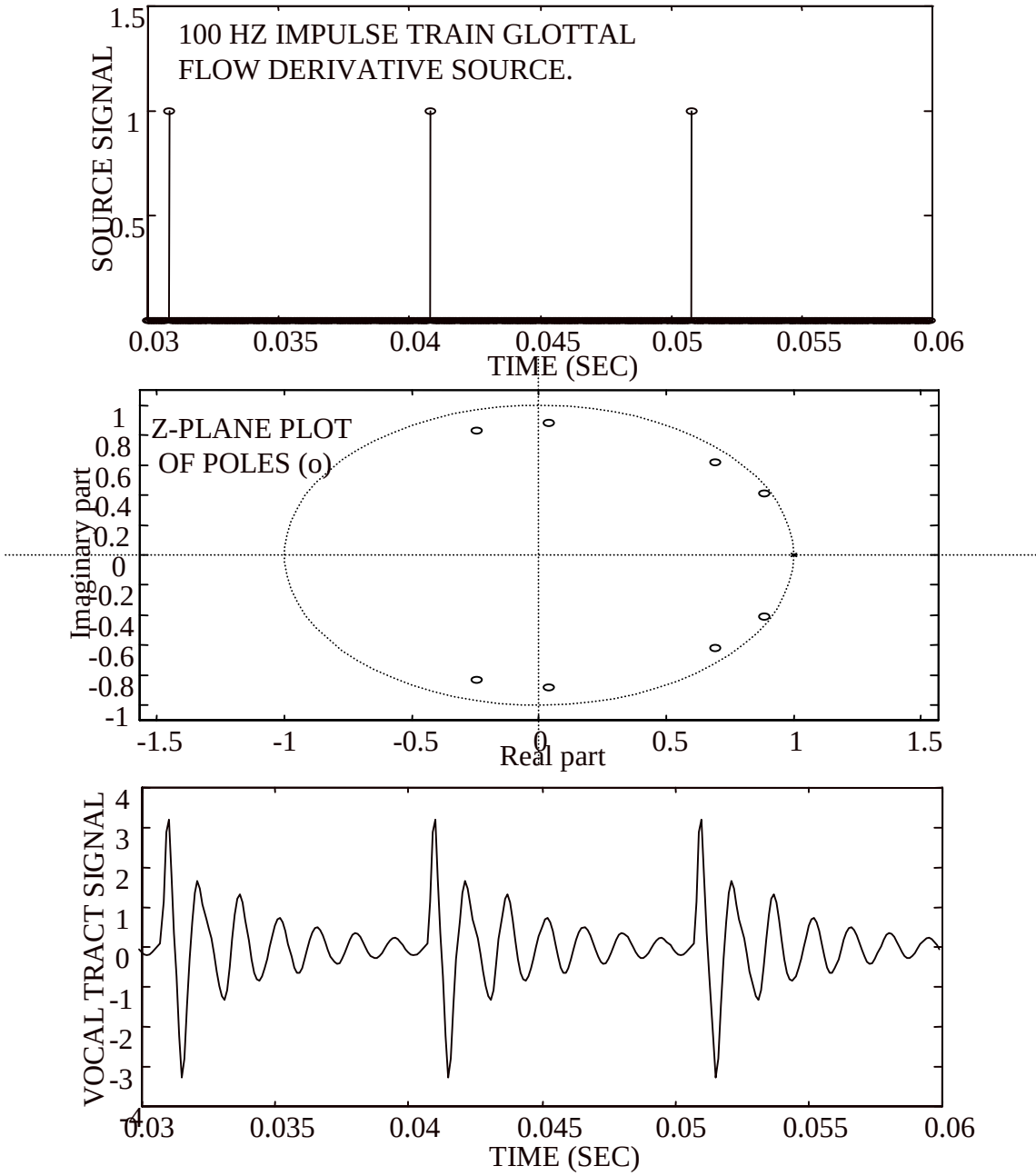


Figure 2.3. Synthetic impulse train response of an 8-pole vocal tract model for /a/. The top plot displays the input impulse train, the middle plot shows the pole locations for the all-pole resonator vocal tract model, and the bottom plot shows the resulting voice time series.

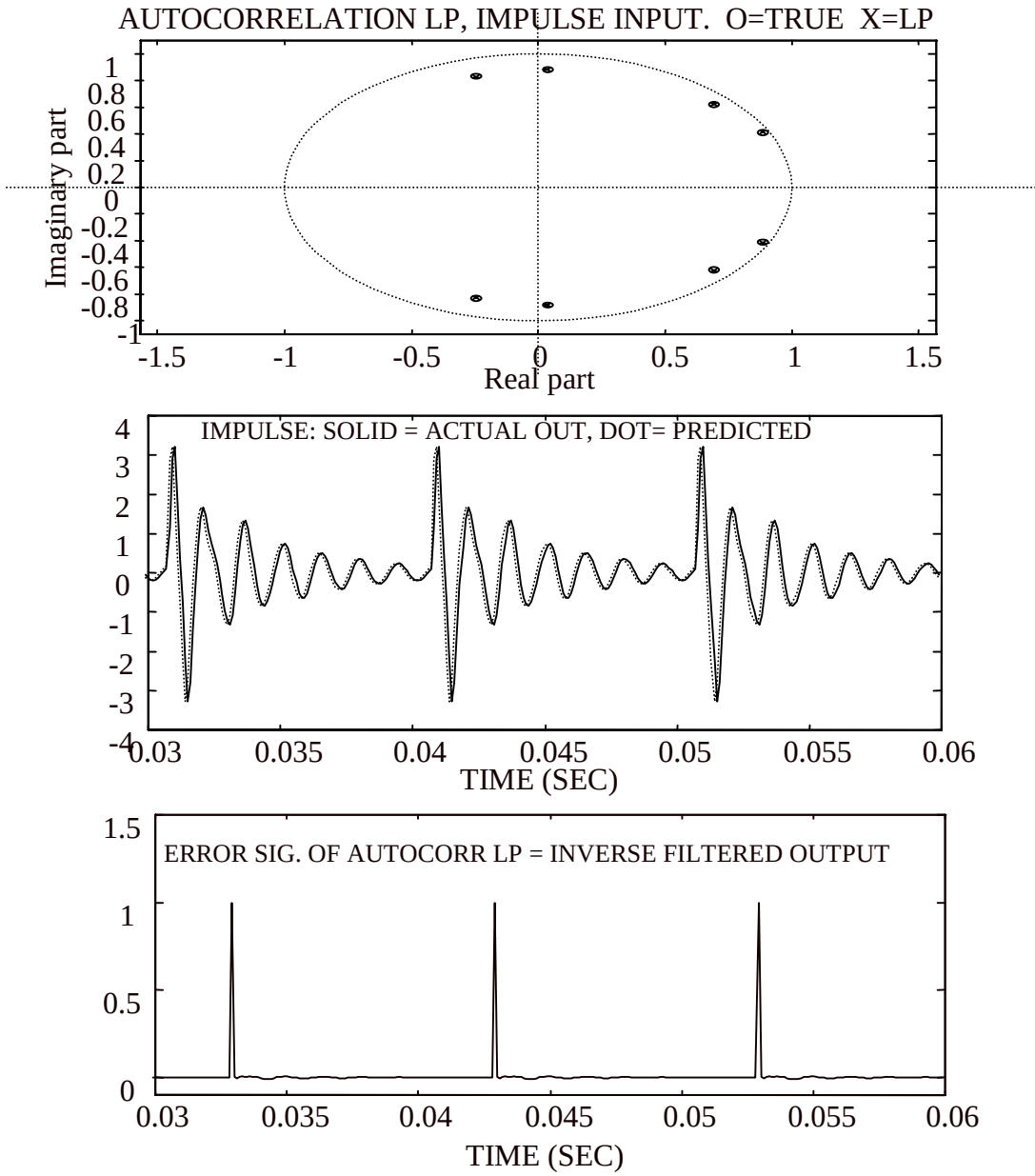


Figure 2.4. Results of autocorrelation LP on the system of Fig 2.3. The top plot repeats the actual pole locations (o) and shows the poles (x) estimated via autocorrelation LP on 10 cycles of the output voice windowed with a hamming window; the estimated poles are within 1% of the actual poles. The middle plot shows the actual voice output (solid) and the predictor output; again, there is close agreement. The bottom plot displays the error signal, which closely agrees with the actual impulse train driving function.

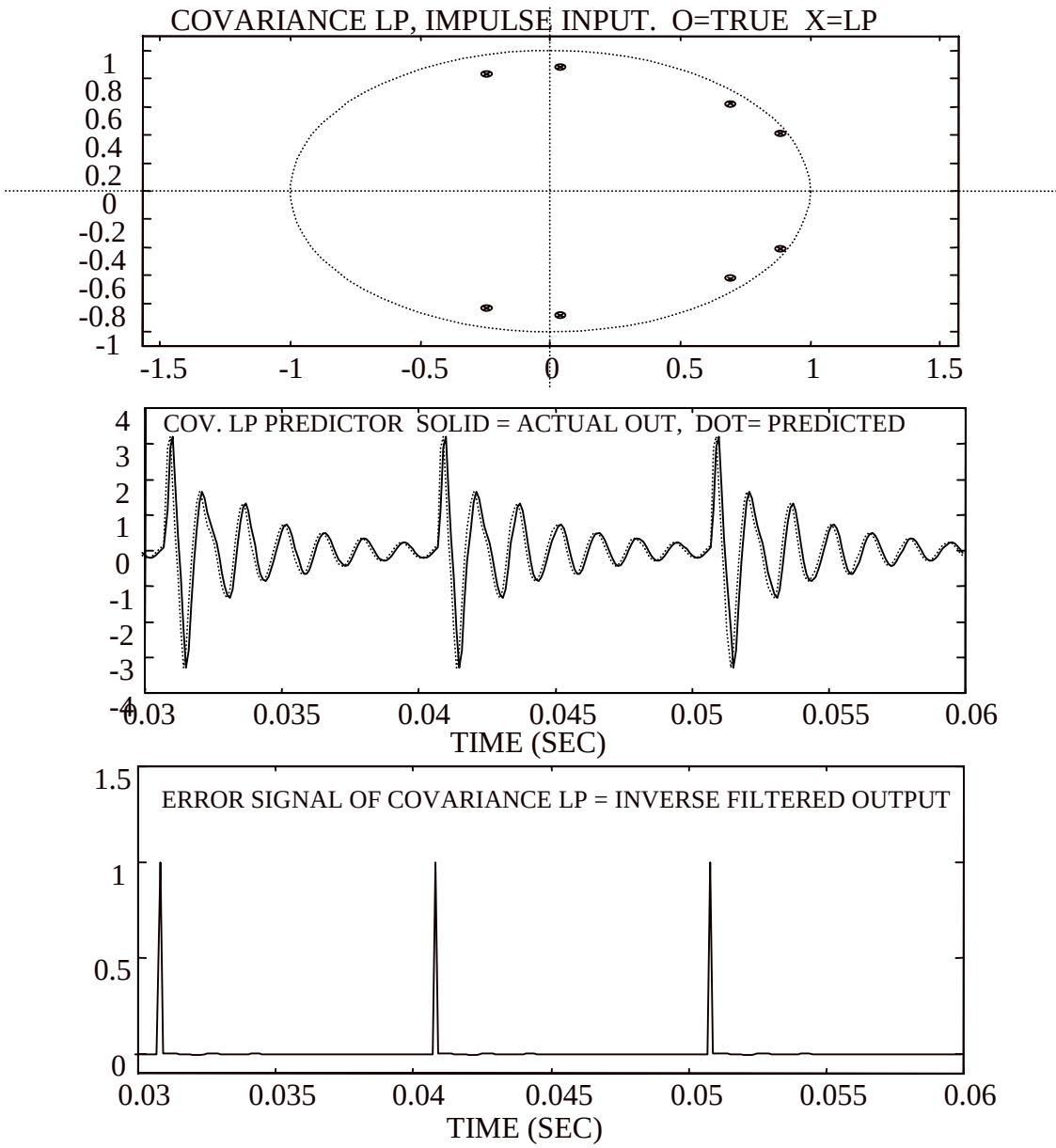


Figure 2.5. Results of covariance LP on the system of Fig 2.3. The top plot repeats the actual pole locations (o) and shows the poles (x) estimated via covariance LP on a 70 sample window of the output voice, which includes an impulse and its response. The estimated poles are within 1% of the actual poles. The middle plot shows the actual voice output (solid) and the predictor output; again, there is close agreement. The bottom plot displays the error signal, which closely agrees with the actual impulse train driving function.

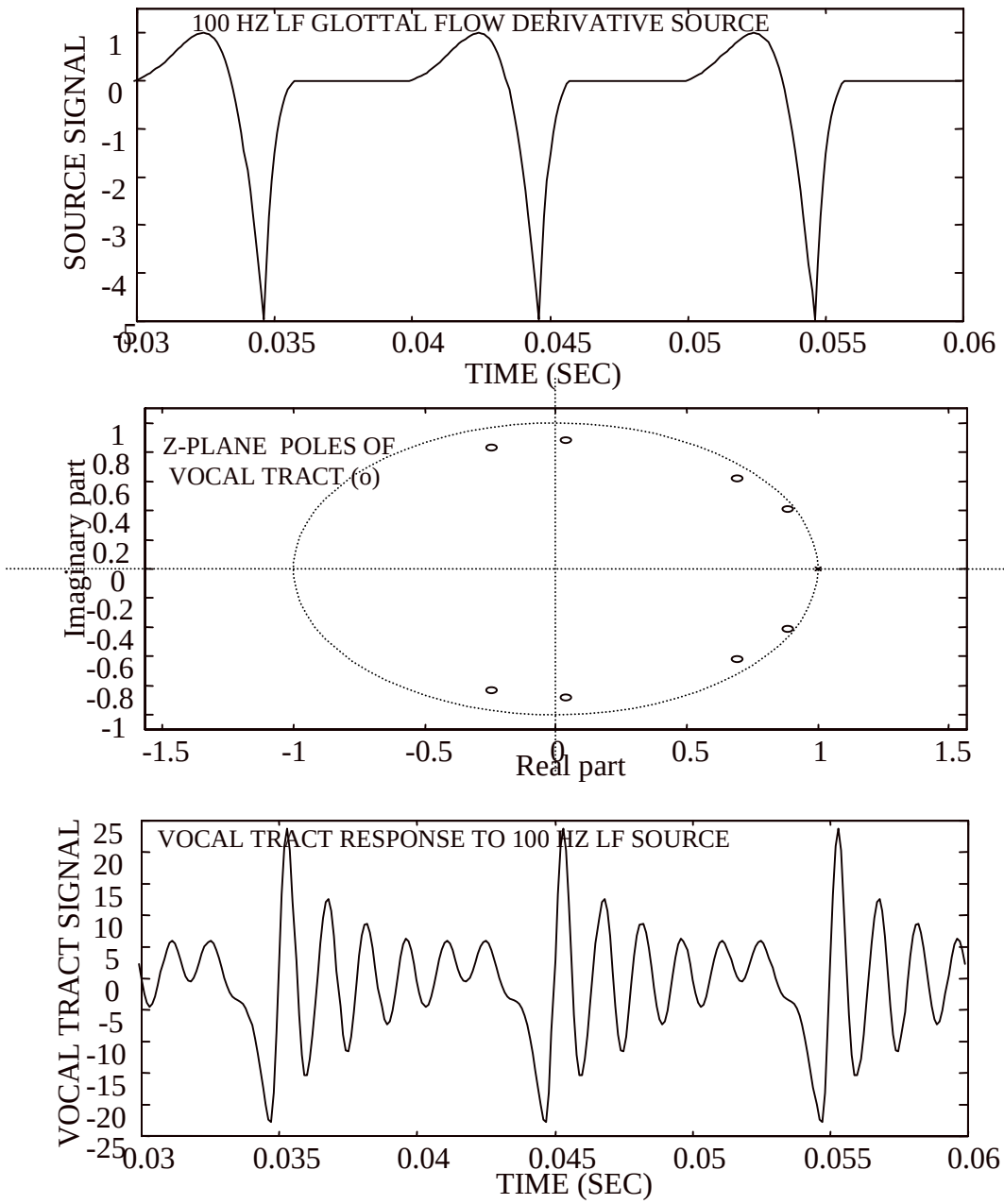


Figure 2.6. Synthetic LF source response of the same system of Fig. 2.3. The top plot displays the input LF source time series, the middle plot shows the pole locations, and the bottom plot shows the resulting voice time series.

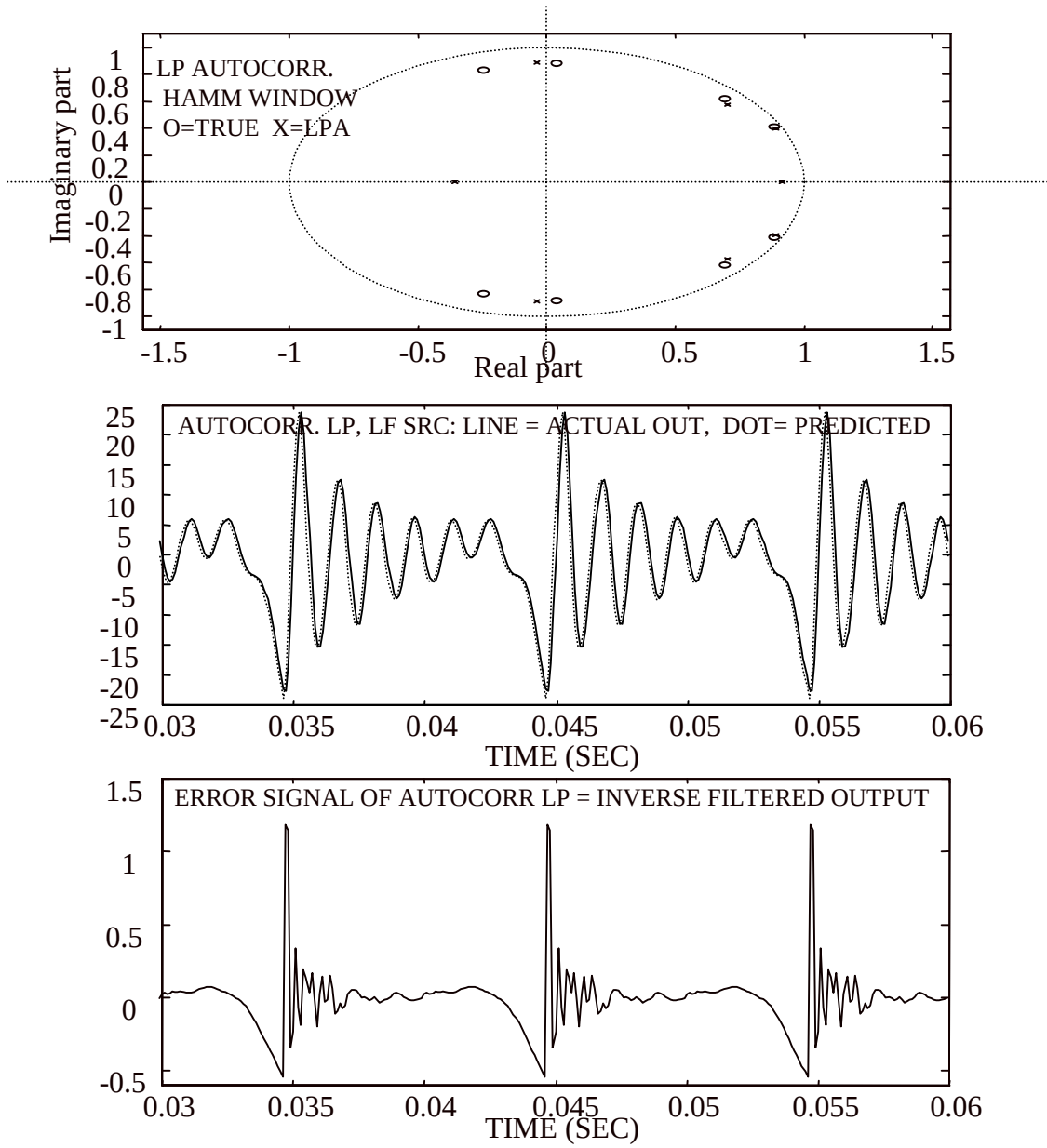


Figure 2.7. Result of autocorrelation LP on the LF source model system of Fig. 2.6; compare directly with Fig 2.4. The top plot of pole locations of estimate (x) and actual (o) reveals considerable error, especially at the higher frequencies; in fact, one of the complex pairs has become 2 real roots. The middle plot, however, reveals LP is still performing its intended primary function of prediction quite well. The bottom plot shows that, because of the error in pole locations, the inverse filter works poorly (compare to top plot of Fig 2.6).

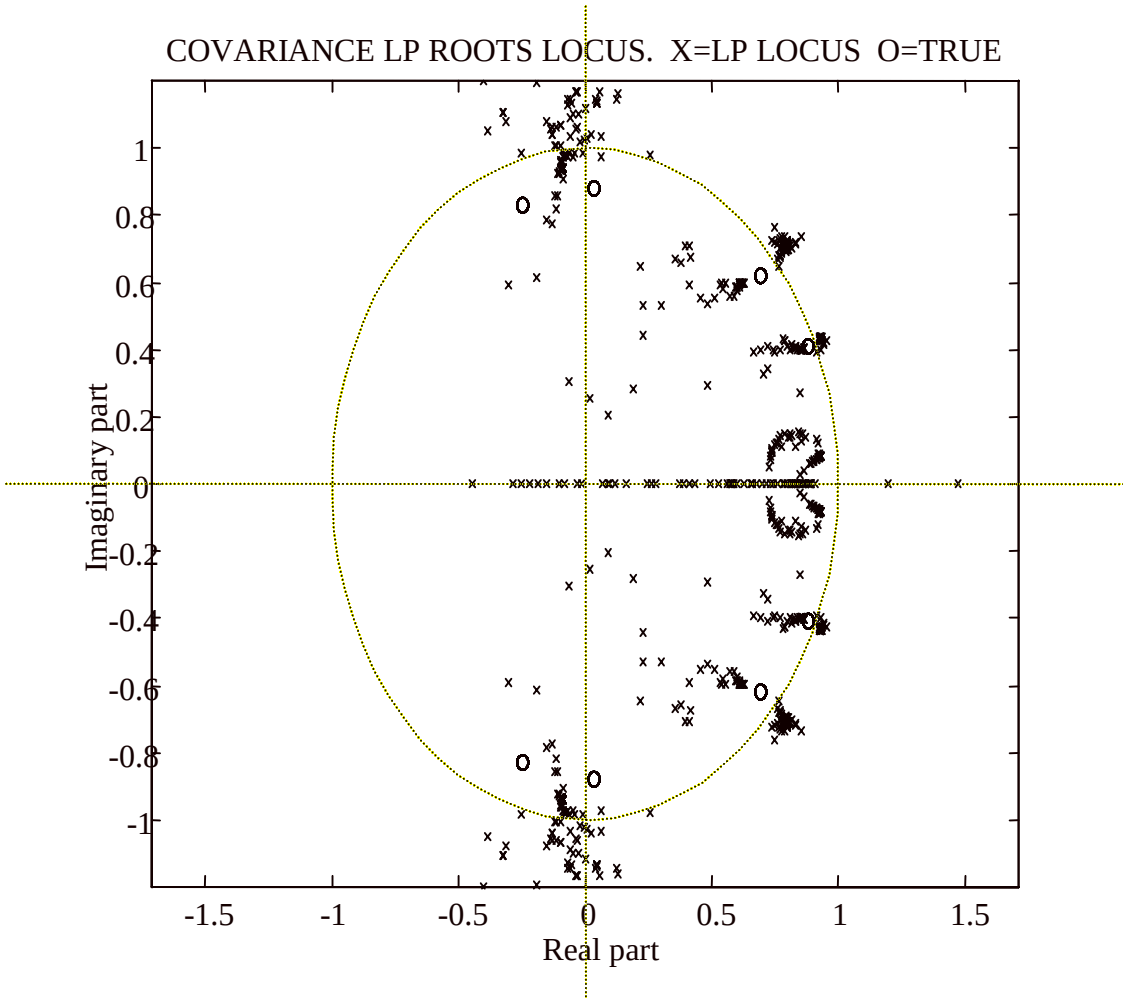


Figure 2.8. Covariance LP poles for a range of analysis window positions. The analysis window length is set to 40 samples, and the position of the window is swept through one complete fundamental period, generating and plotting the resulting LP pole positions for each position. Note that calculated positions (x) never reach true positions (o). The same is true of other selections of window length.

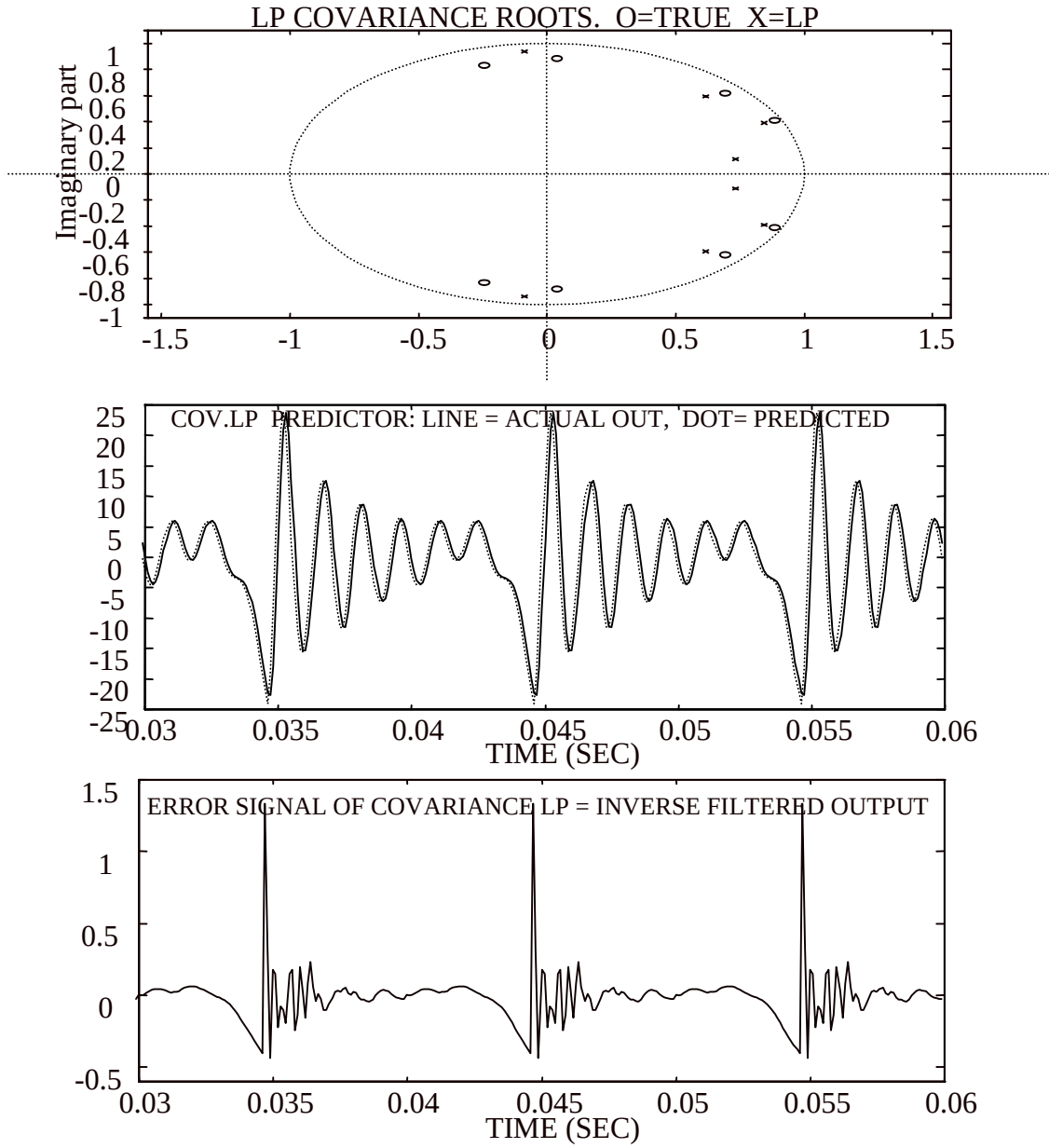


Figure 2.9. Result of covariance LP on non-impulsive system of Fig. 2.6 for an optimal analysis window position; compare with Fig 2.7. The top plot of pole locations of estimate (x) and actual (o) still reveals considerable error. The middle plot again reveals LP is still performing its intended primary function of prediction. The bottom plot again shows that, because of the error in pole locations, the inverse filter works poorly (should be like the top plot of Fig 2.6).

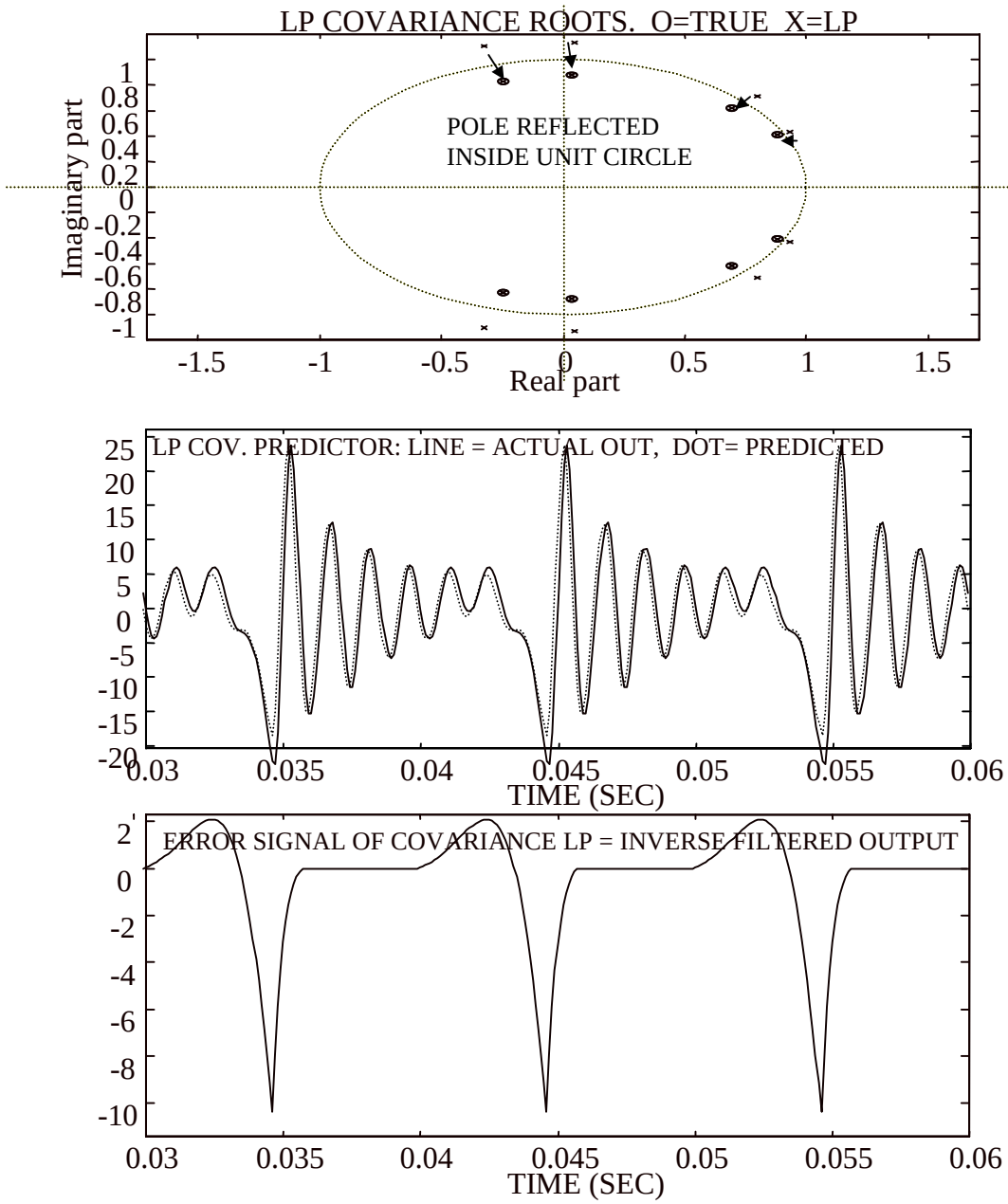


Figure 2.10. Reflection of covariance LP poles inside unit circle. For the same analysis of Figs 2.8, the top plot shows the poles outside the unit circle are inverted to reflect them inside; the resulting new positions coincide almost exactly with the true positions. Prediction (middle plot) is somewhat less accurate, but good inverse filtering is achieved (bottom plot.)

$$U'(n) = \begin{cases} (1/OT) \cdot (2 \cdot n - 3 \cdot n^2 / OT), & n \le OT \\ 0, & n \ge OT + 1 \end{cases}$$

$$U(n) = n^2 - n^3 / OT$$

Here:

n = sample number (discrete time)

U'(n) = derivative of flow

U(n) = flow

OT = "open time" = INT{OPENQ * FSR / FSRC}

where

OPENQ = "open quotient" = pulse width/source period

FSR = sample rate

FSRC = fundamental frequency

Figure 2.11. Equations describing the KGLOTT88 glottal pulse model. [22]. The glottal flow derivative pulse is modeled using an inverted parabola with a linear term added to tilt and lower the end of the pulse

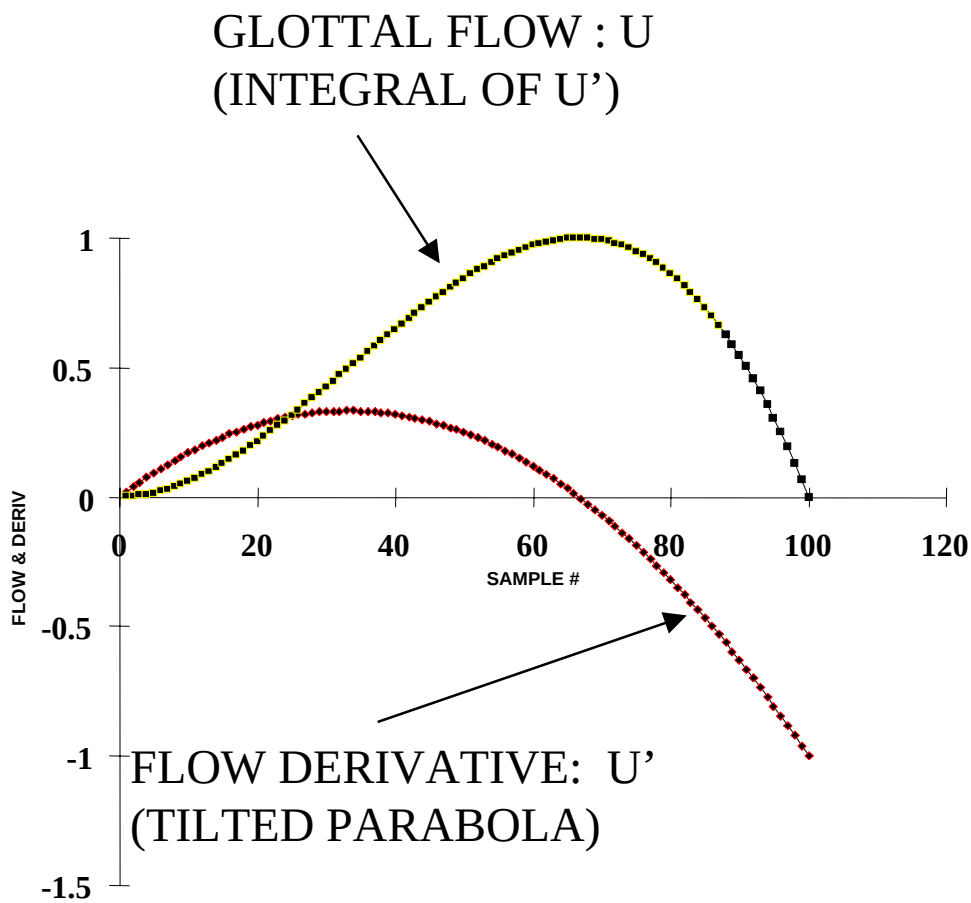
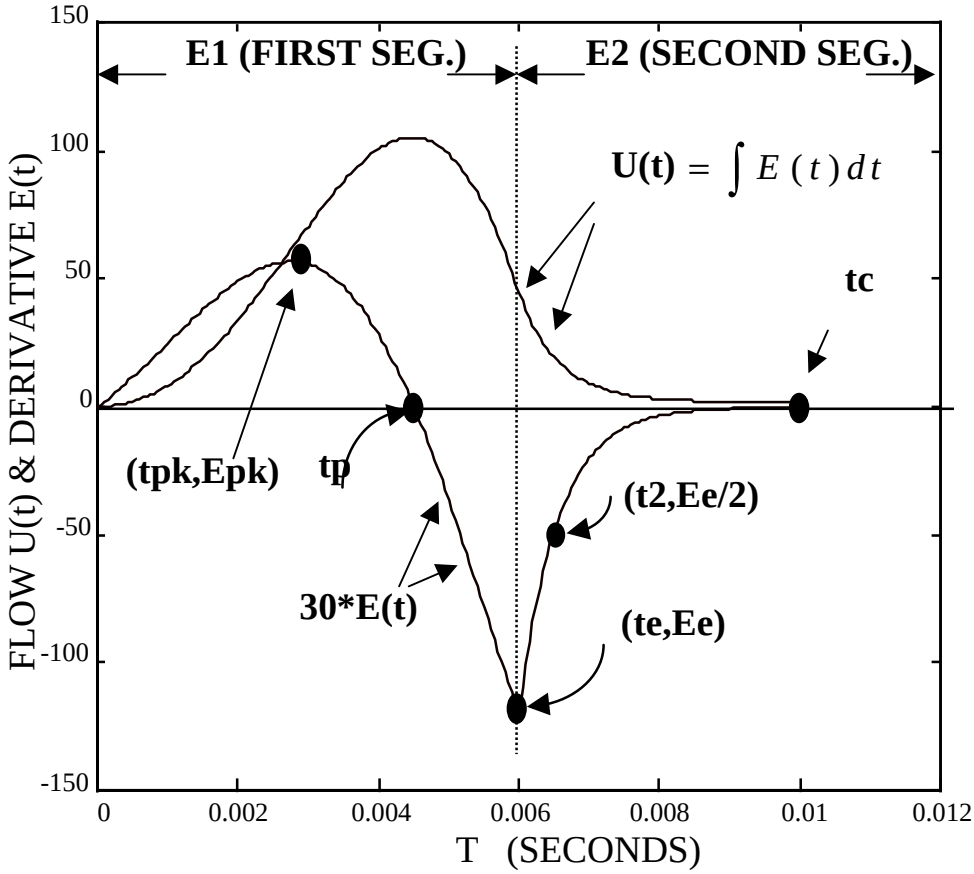


Figure 2.12. Plot of an example KGLOTT88 glottal flow derivative pulse U' and its integral U.



$$E(t)=\begin{cases} E_1(t)=E_0 e^{\alpha t}\sin\omega_g t, & t\le t_e \\ E_2(t)=-E_e e^{-\varepsilon(t-t_e)}, & t_e < t \le t_c \\ E_{2B}(t)=-E_e e^{-\varepsilon(t-t_e)}+m(t-te), & t_e < t \le t_c \end{cases}$$

Figure 2.13. Simplified LF model used to fit the calculated flow derivative [31]. Both glottal flow (U) and its derivative (E) are shown. Four parameters (*tp*, *te*, *Ee*, *t2*) are major features that can define this model. The maxima defines tpk and Epk, which are additional quantities useful in generation of the LF parameters (Figs 2.15 - 2.16). A fifth parameter *m* is optionally added to supply a linear term to the second segment for improved fit, resulting in the second version E_{2B} of the second segment. Parameters ε and *Eo* are found by solving a set of simultaneous nonlinear equations.

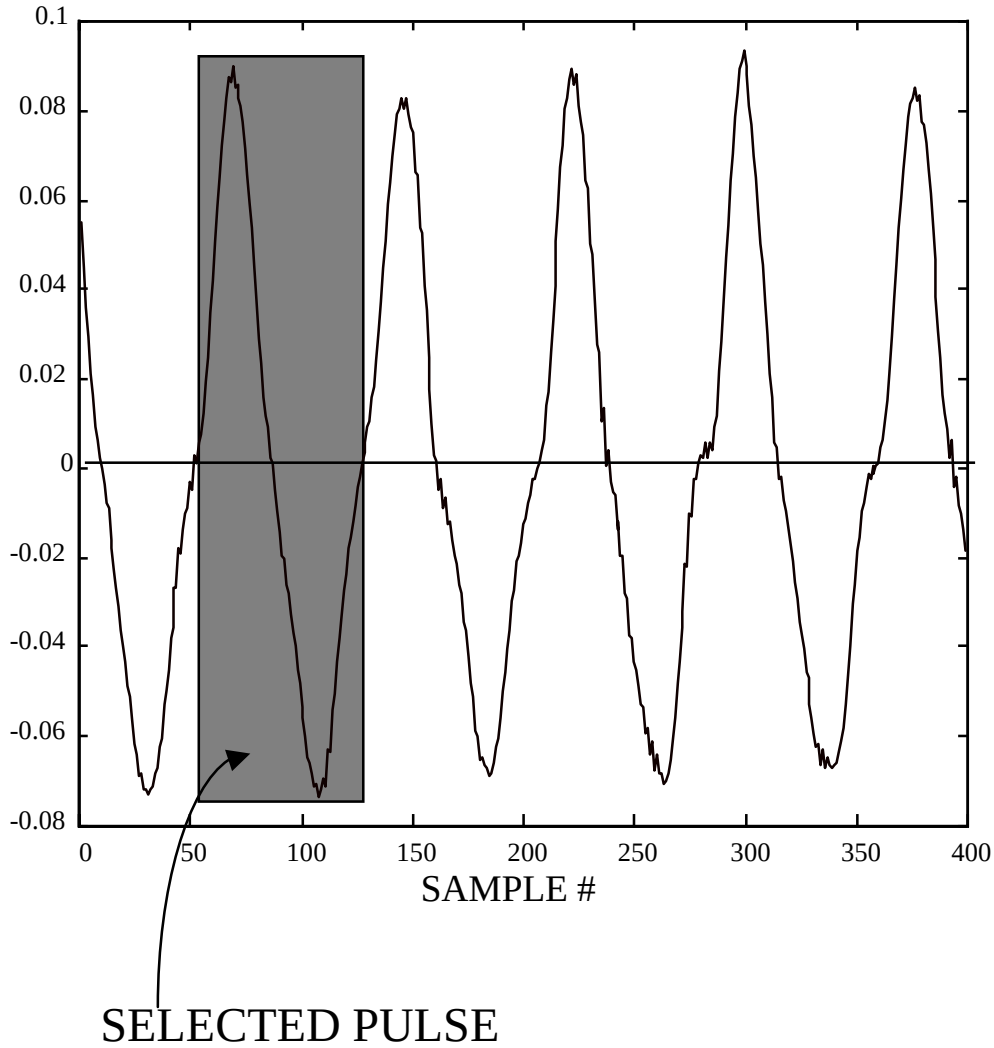


Figure 2.14. Selection of a single pulse from the glottal flow derivative time series. The LF model is least squares fitted to the selected pulse.

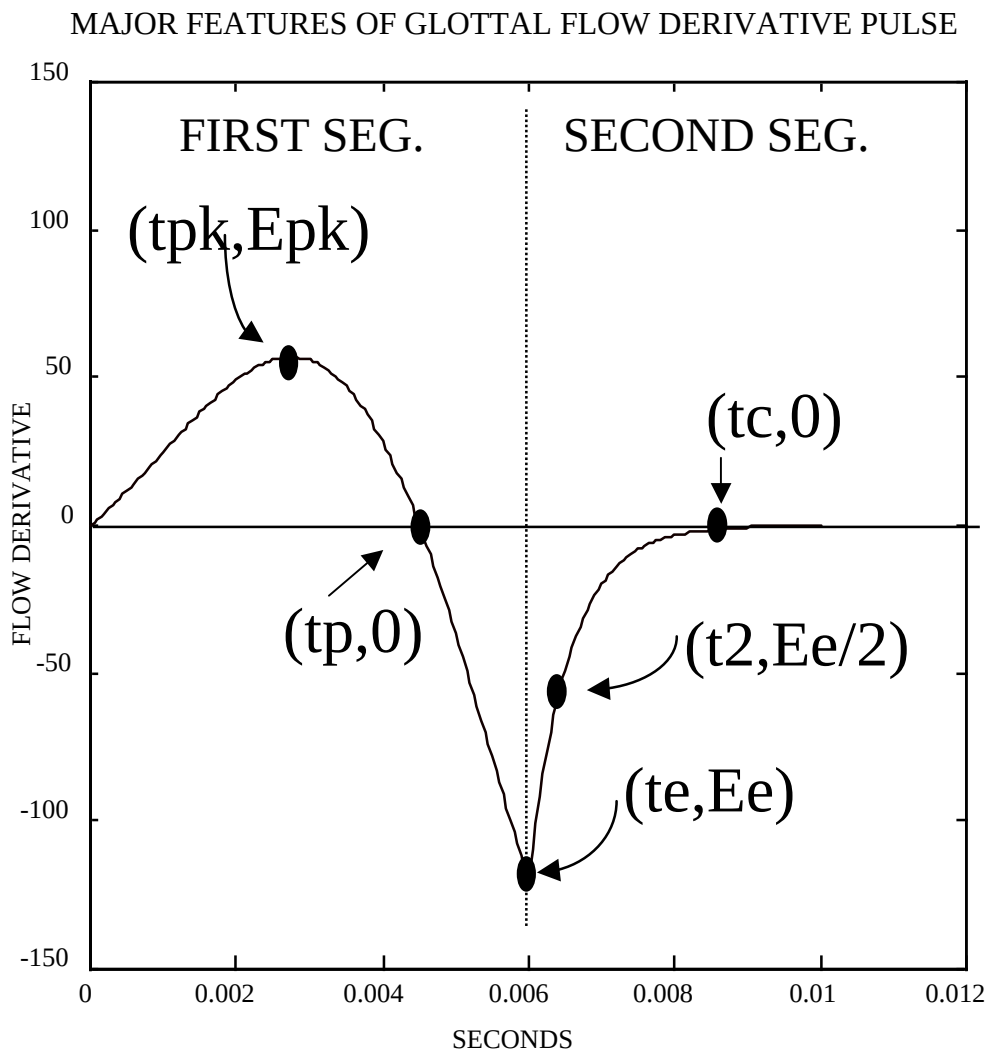


Figure 2.15. Identification of automatically acquired major features of the raw inverse filtered flow derivative. Clearly identifiable points on the pulse provide values for: Epk = positive peak value, tpk =– time of positive peak, Ee = negative peak value, te-= time of negative peak, Ee2 = half value of the second segment (return phase), and t2 = time of Ee2. The starting time is implicitly taken as zero at the start of the pulse data, and tc is set by the end of the pulse data.

1. $$\omega_g = \frac{\pi}{t_p}$$

2. $$\alpha = \frac{1}{(t_e - t_{pk})} \cdot \ln\left(\frac{-E_e}{E_{pk}} \cdot \frac{\sin(\omega_g t_{pk})}{\sin(\omega_g t_e)} \right)$$

3. $$t_{pk} = \frac{1}{\omega_g} \cdot \tan^{-1}\left(\frac{-\omega_g}{\alpha} \right), \qquad \alpha < 0$$

$$t_{pk} = \frac{1}{\omega_g} \cdot \tan^{-1}\left(\pi + \frac{-\omega_g}{\alpha} \right), \qquad \alpha \geq 0$$

4. $$E_0 = \frac{-E_e}{e^{\alpha t_e} \sin(\omega_g t_e)}$$

5. $$\varepsilon = \frac{\ln(2)}{(t_2 - t_e)}$$

Figure 2.16. Equations used to define the first approximation of the LF fit by use of the major features (Fig. 2.15) automatically acquired from the raw inverse filtered glottal flow derivative pulse. The major features give rise to six values: Epk, tpk, Ee, te, tp, and t2, which are then used to determine five approximate values of LF parameter values by evaluating equations 1 – 5 in sequence. Since this system of equations is over determined (i.e., there are more feature parameters than needed), it was found useful to use equation 2, which combines two features as the ratio of Ee/Epk, a physically significant quantity.

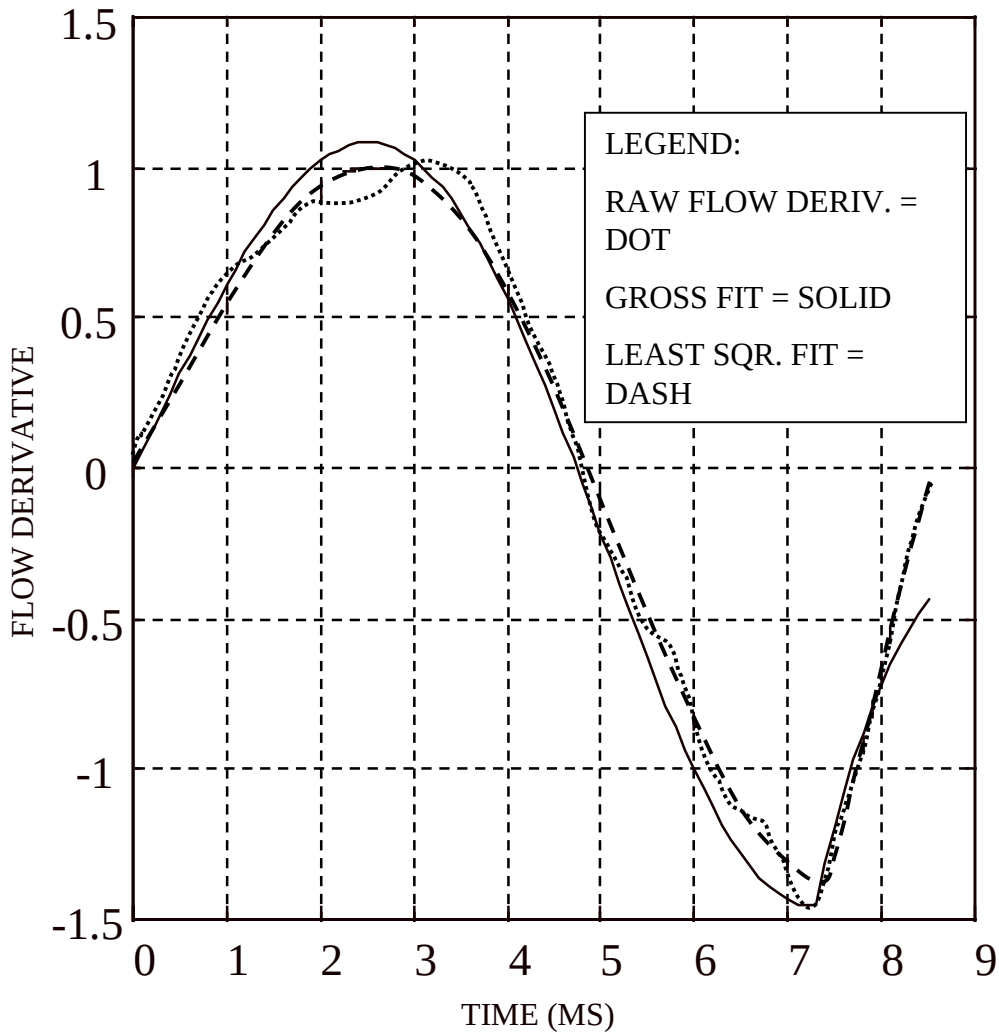


Figure 2.17. Example of fitting the LF model to a selected pulse. The dots represent the raw pulse, the line shows the first approximation used as a starting point for the least squares minimization, and the dash shows the final least squares fit.

1 - 3 solved simultaneously:

1. $$\alpha = \frac{1}{(t_e)} \cdot \ln\left(\frac{-E_e}{E_{pk}} \cdot \sin(\omega_g t_e) \right)$$

2. $$t_{pk} = \frac{1}{\omega_g} \cdot \tan^{-1}\left(\frac{-\omega_g}{\alpha} \right), \qquad \alpha < 0$$

$$t_{pk} = \frac{1}{\omega_g} \cdot \tan^{-1}\left(\pi + \frac{-\omega_g}{\alpha} \right), \qquad \alpha \geq 0$$

3. $$E_0 = \frac{1}{e^{\alpha t_{pk}} \sin(\omega_g t_{pk})}$$

4. $$\varepsilon = \frac{\ln(2)}{(t_2 - t_e)}$$

5. $$m = \frac{-E_e}{t_c} \cdot e^{-\varepsilon t_c}$$ (optional parameter to force final value to zero)

Figure 2.18. Equations used to fit the raw pulse to obtain an optimized set of LF parameters in the least squares sense. The least squares fit is carried out using major feature parameters tp, Ee, te, and t2 as four independent degrees of freedom. Equations 1 - 3 form a simultaneous nonlinear set that are then solved to obtain the LF parameters using the simple sequential solution algorithm. Equation 5 is a parameter used in the optimization with the optional constraint to fix the final value of the pulse to zero.

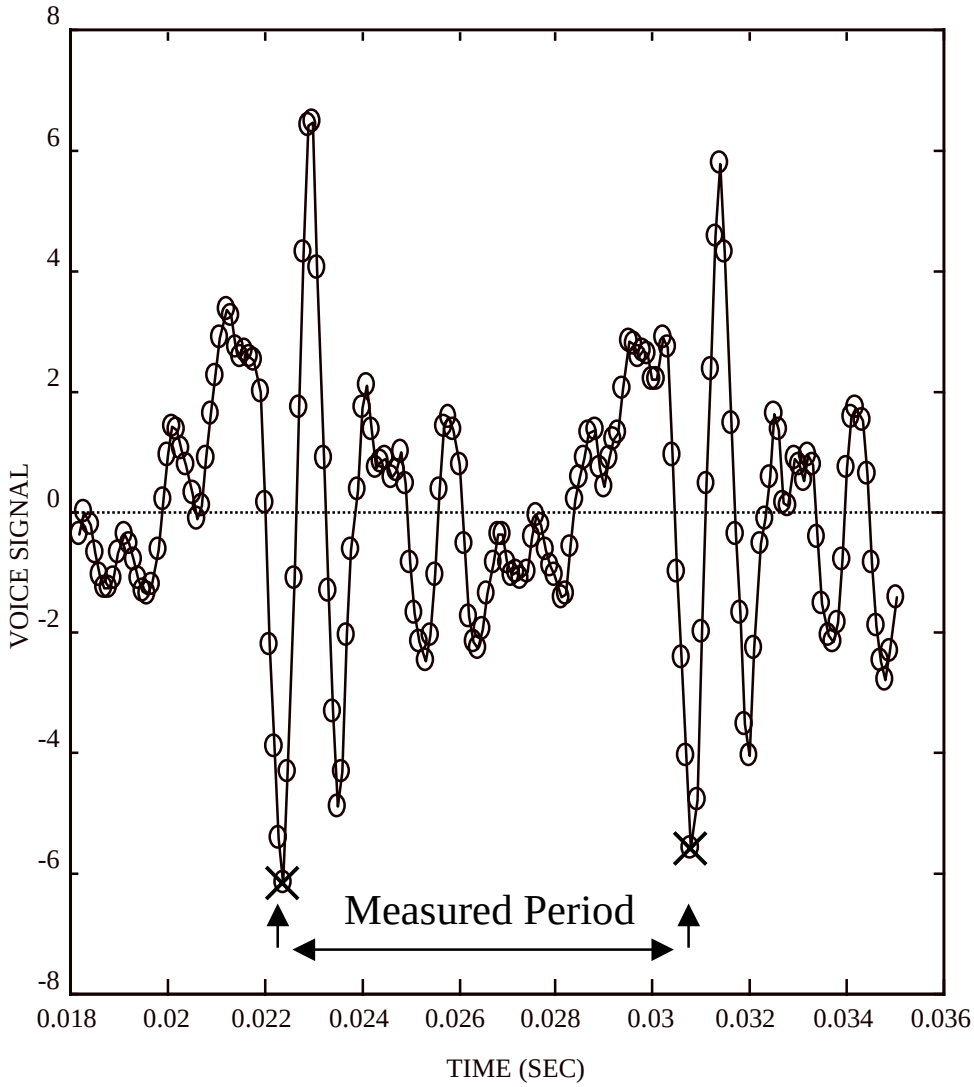


Figure 2.19. Identification of negative peaks for fundamental frequency tracking on a typical original voice signal. The tracking algorithm automatically selects the minima indicated by X and performs a parabolic interpolation to determine the time of minima to a precision of less than one sample. Circles indicate sample points.

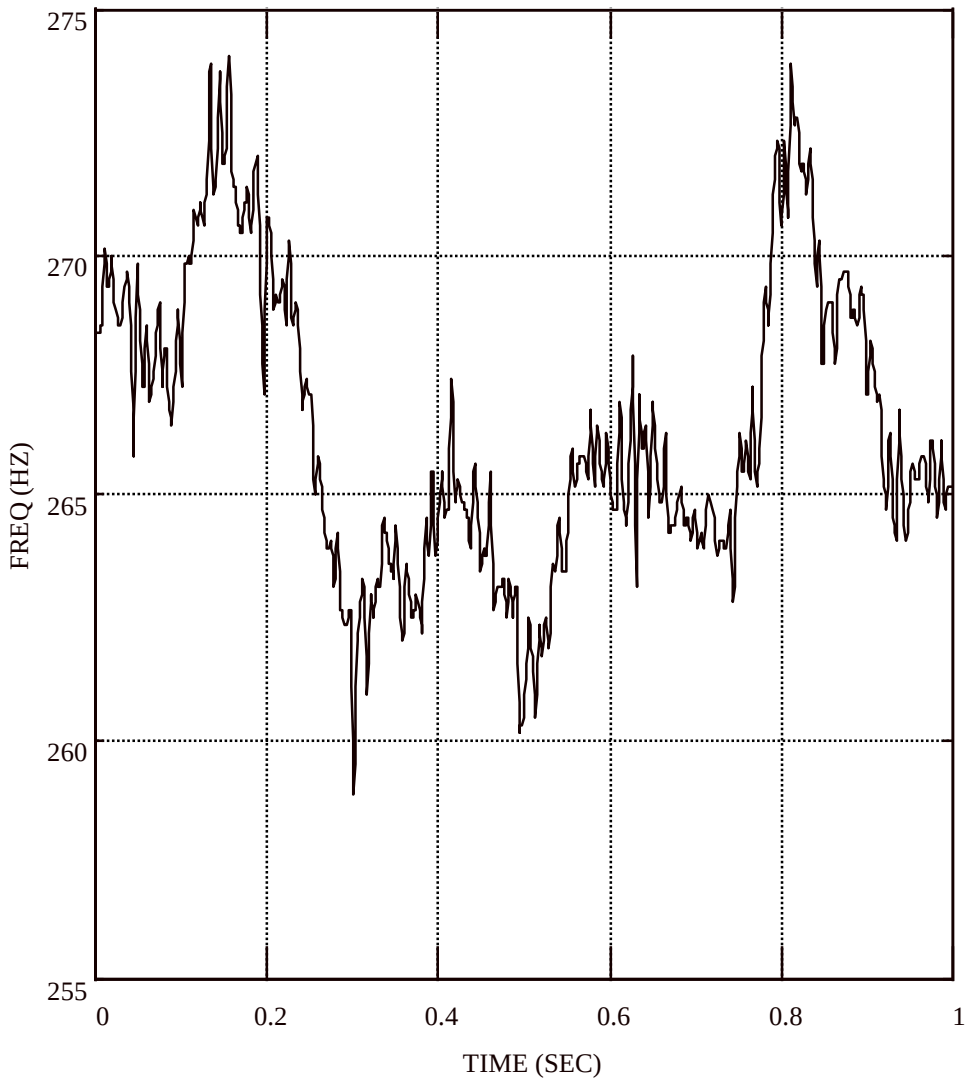


Figure 2.20. A high resolution 1 sec. fundamental frequency track resulting from subsample interpolation. Note lack of fundamental frequency quantization (flat spots) that would occur without interpolation.

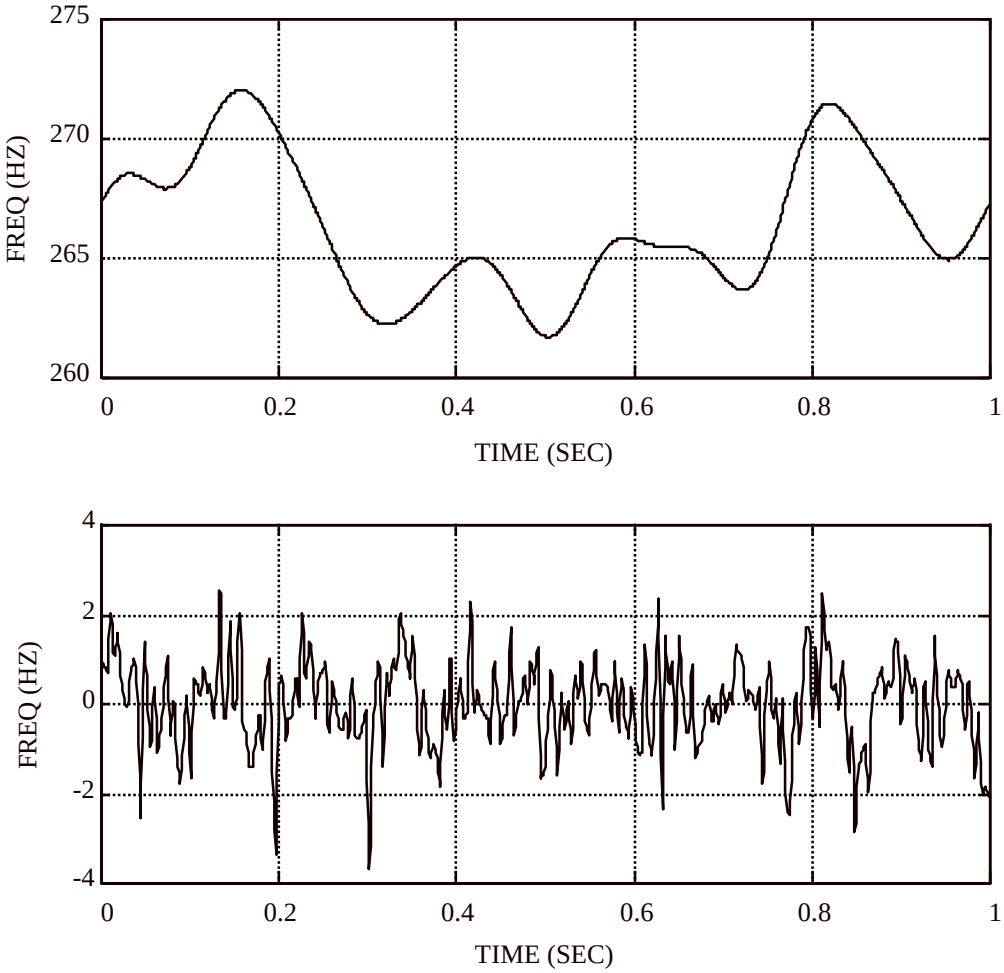


Figure 2.21. The fundamental frequency track of Fig 2.20 is low pass filtered (top part A curve) and high pass filtered (bottom part B curve) yielding the tremor and HFPV time series respectively. A cutoff frequency of 10 Hz is selected.

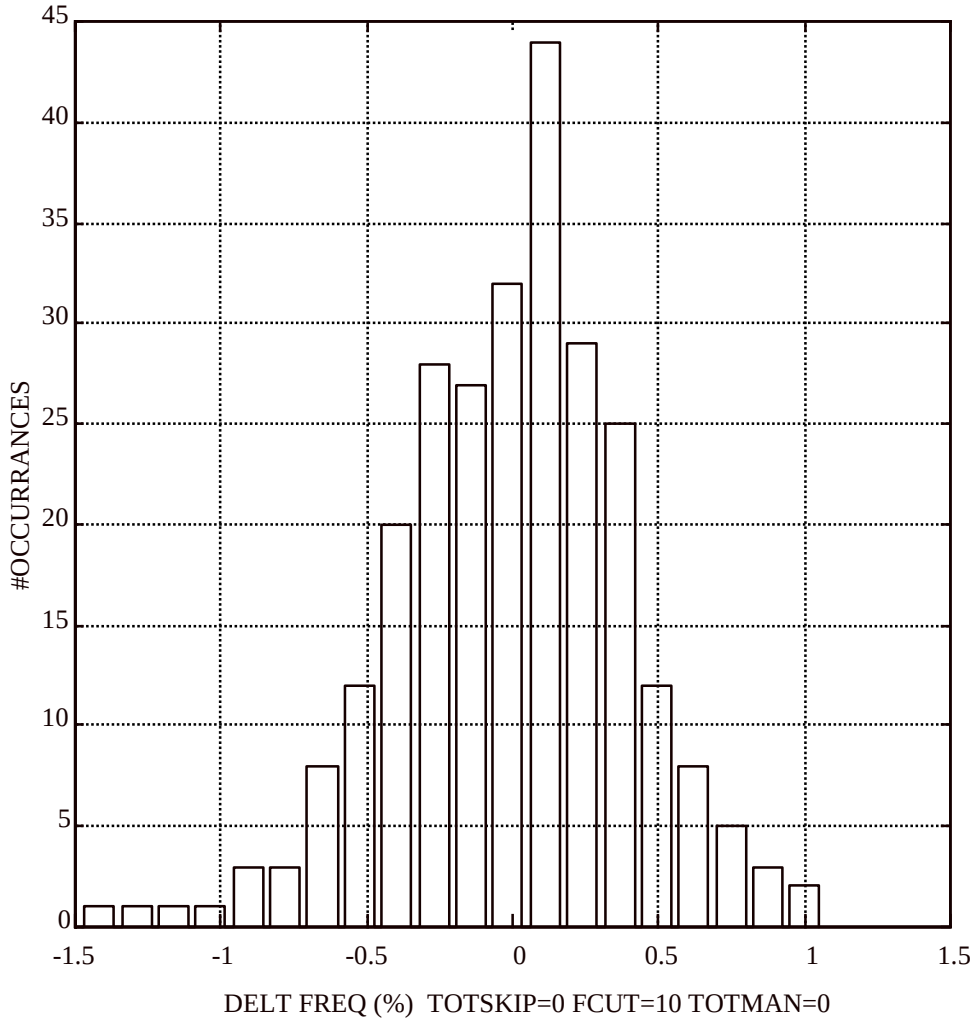


Figure 2.22. Histogram of frequency deviations of the high pass filtered fundamental frequency time series of Fig 2.20. Successful fundamental frequency tracking yields a Gaussian form distribution.

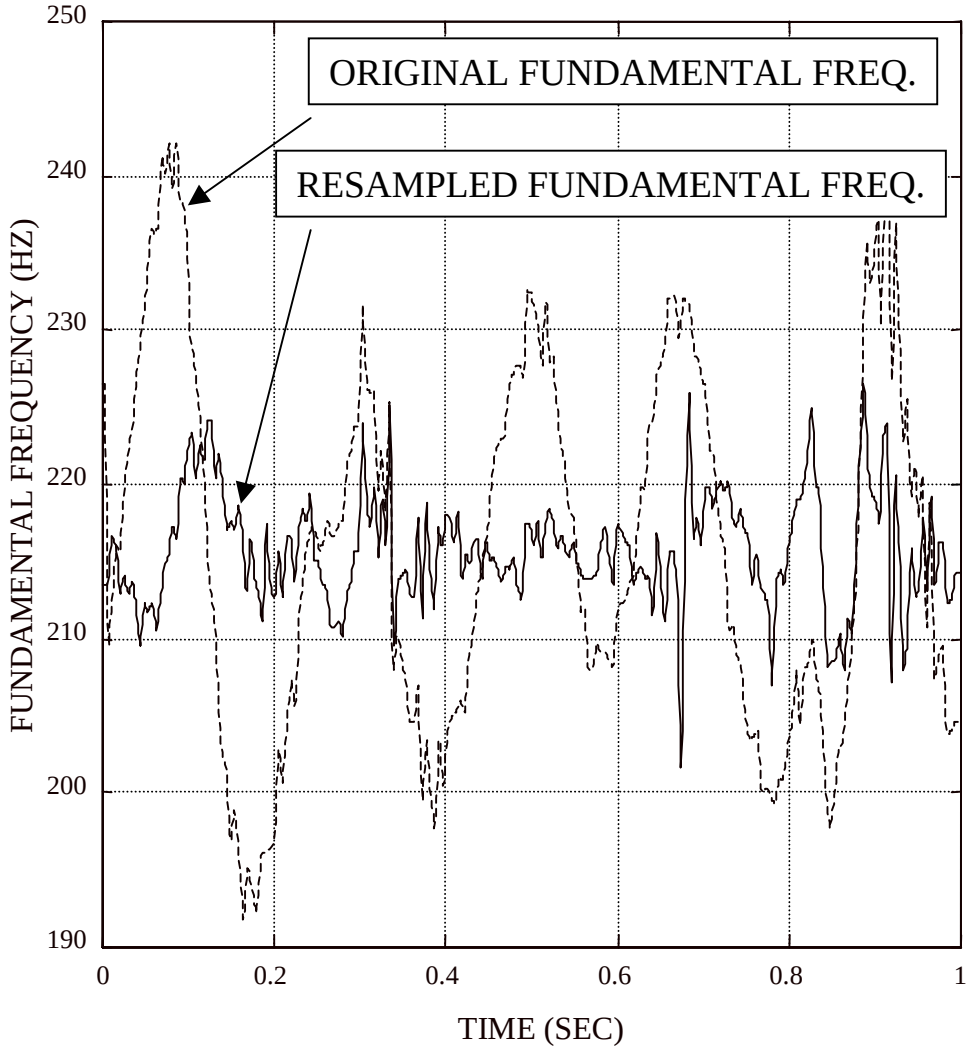


Figure 2.23. Fundamental frequency time series of original voice and voice re-sampled to remove tremor (low frequency variations). Most of the significant fundamental frequency variation is removed

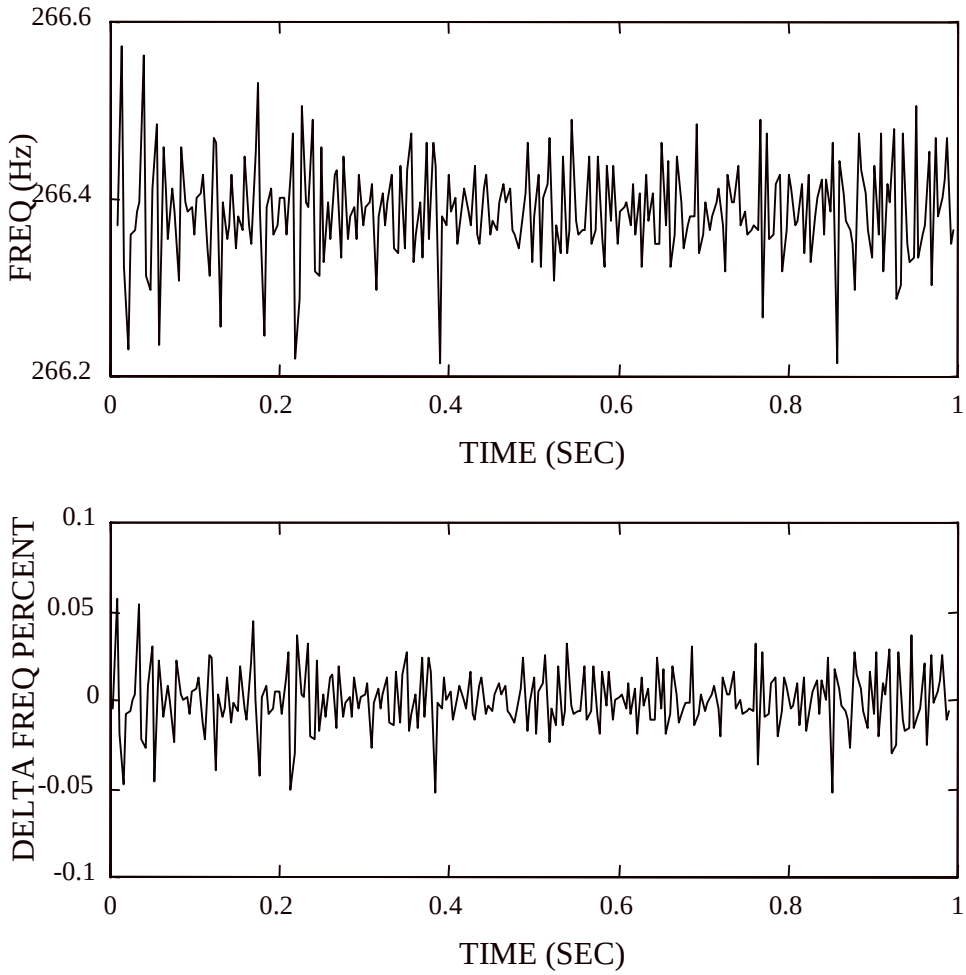


Figure 2.24. Fundamental frequency time series of voice re-sampled to remove all fundamental period variations: both HFPV and tremor. In the upper plot the residual variation after removal is shown in Hz; it is less than 0.4 Hz. In the lower plot the residual is shown in percent.

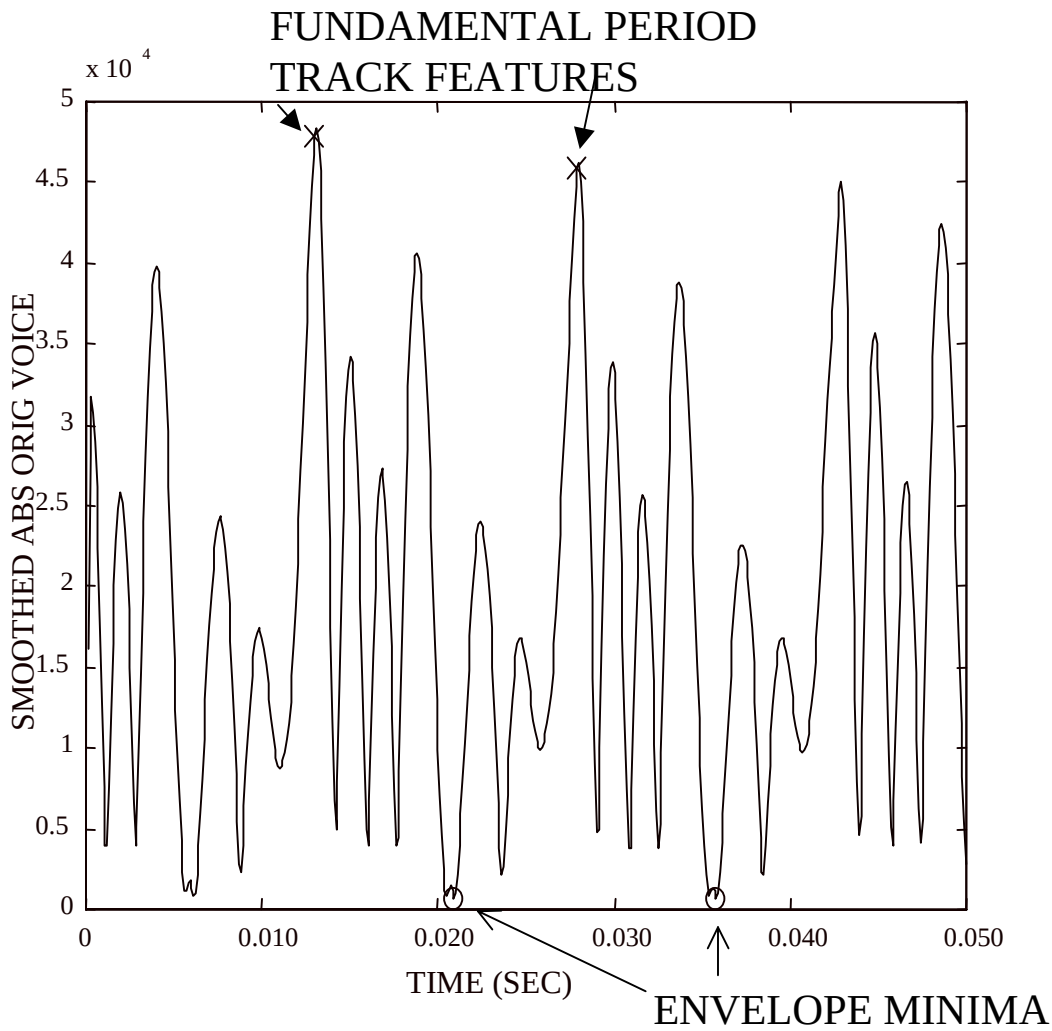


Figure 2.25. Plot of absolute value of original voice signal over 3 periods illustrating fundamental pulse segregation. Fundamental frequency tracking features (X) are used as a starting point for searching for minima (circles), which define the pulse boundaries.

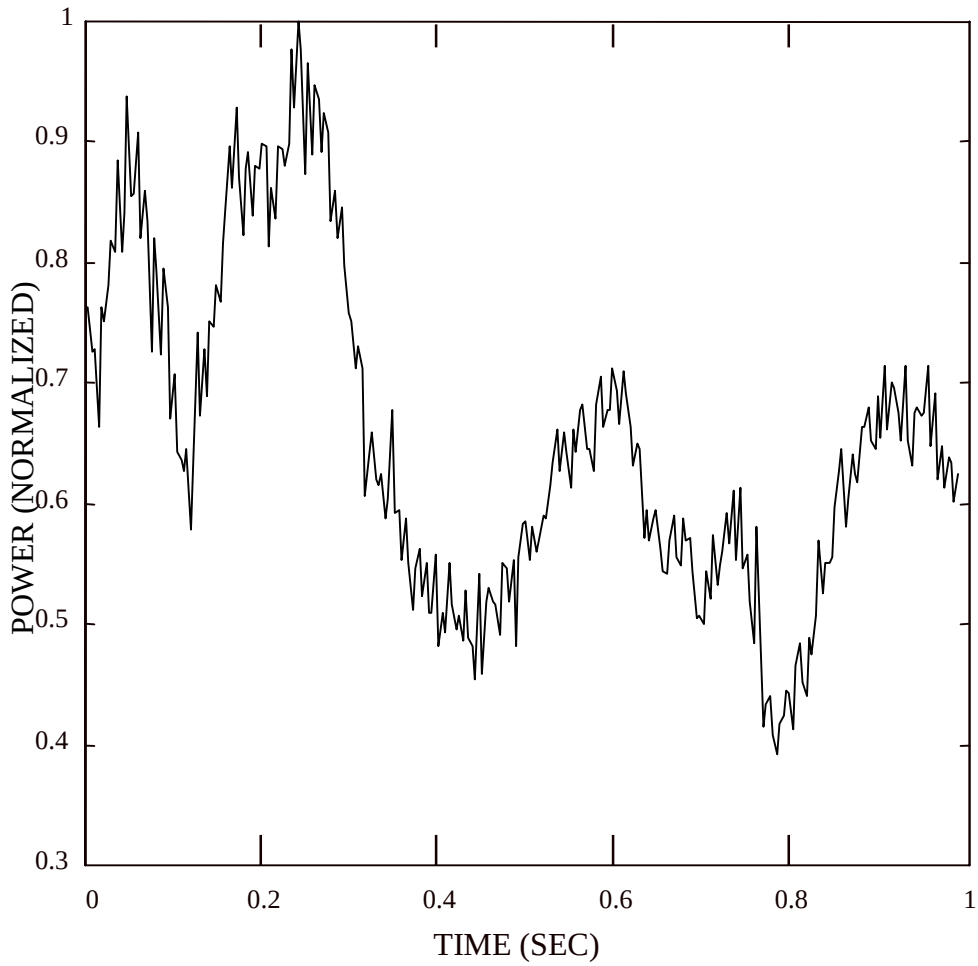


Figure 2.26. Power time history of same signal as Fig 2.25. The results of the pulse identification (Fig 2.19) are used to calculate the energy (sum of samples squared) of each fundamental pulse.

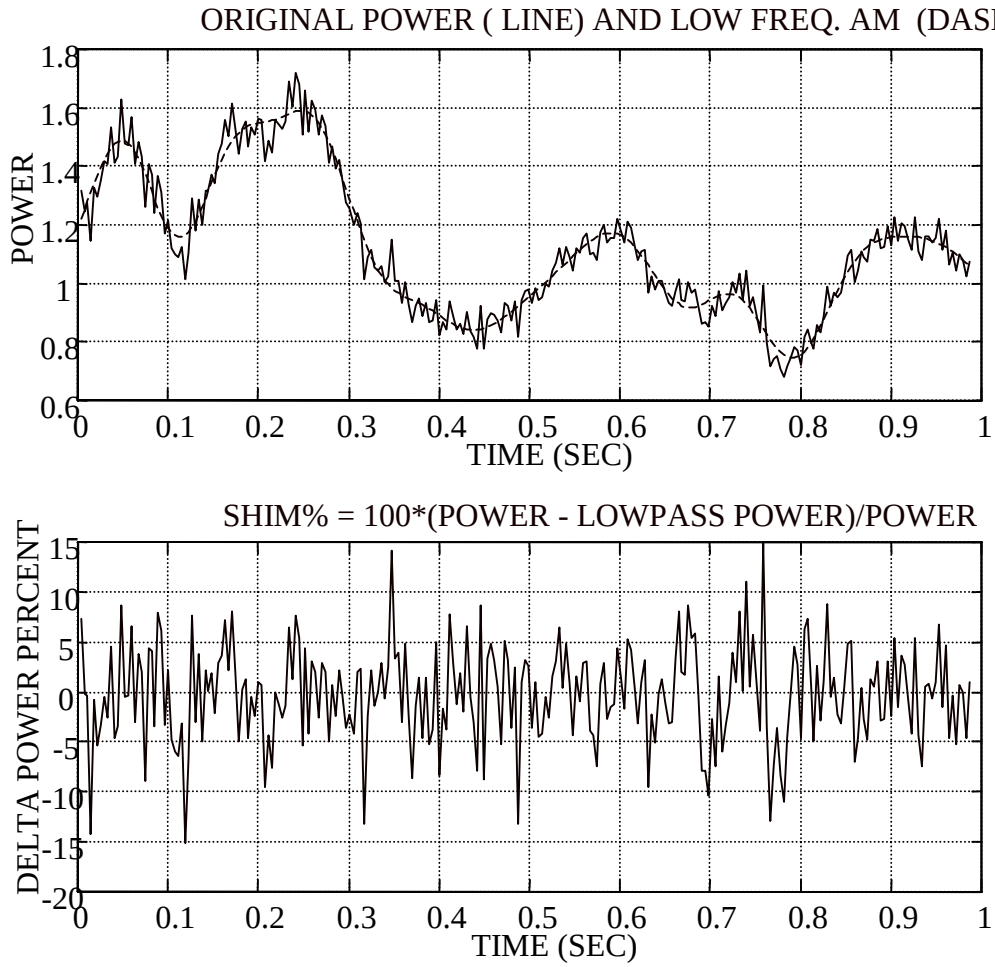


Figure 2.27. Power time history resolved into low frequency volume (top) and high frequency shimmer (bottom) components. The same signal is used as in previous figures.

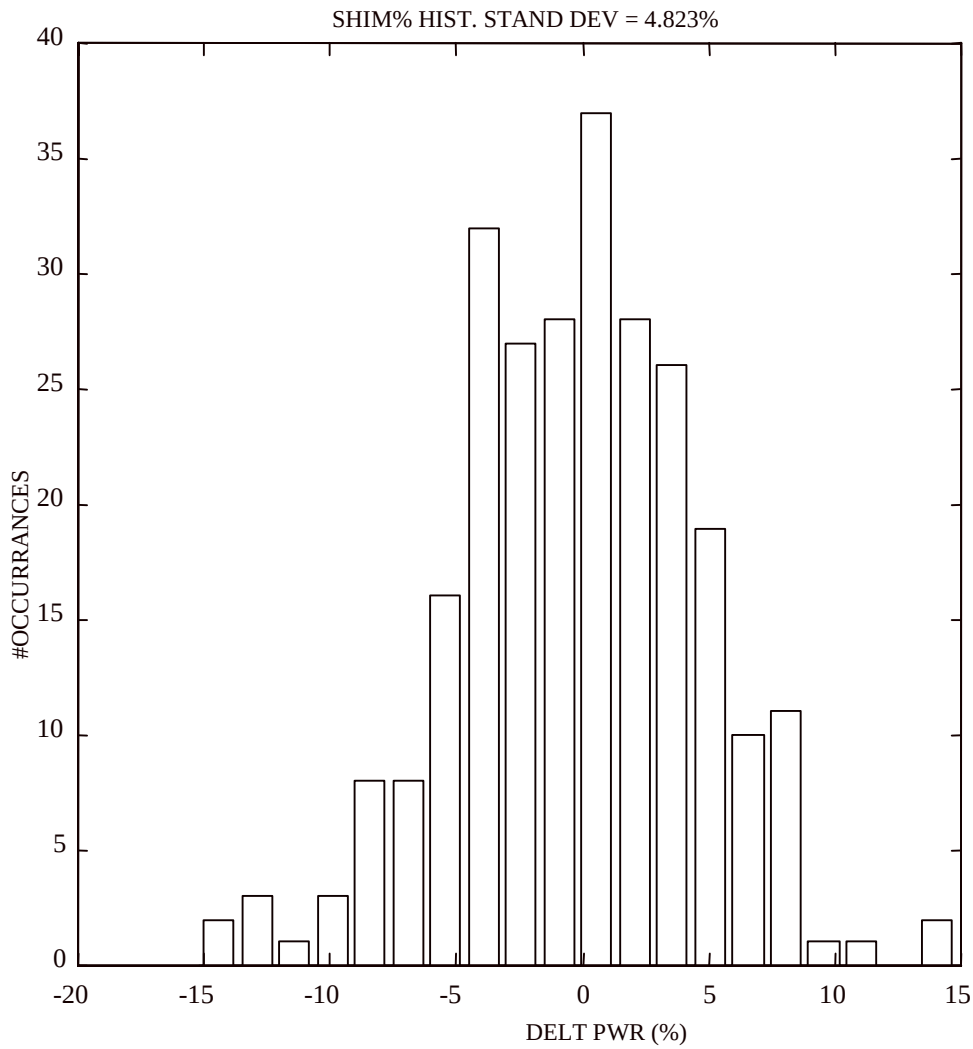


Figure 2.28. Histogram of shimmer values displays Gaussian form.

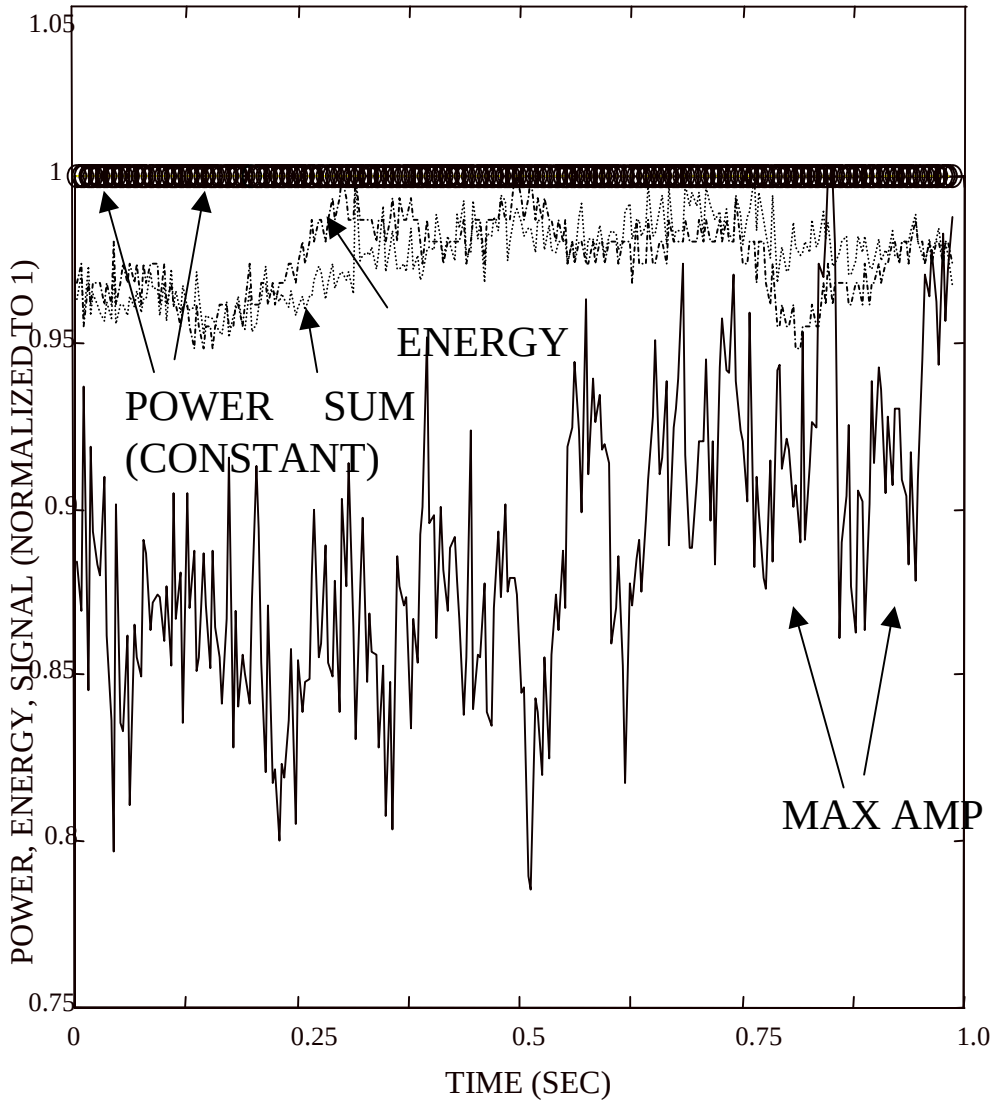


Figure 2.29. Power measures in the AM demodulated voice. The constant value of the power level time series verifies successful processing. Other measures of signal strength such as envelope amplitude show residual variations.

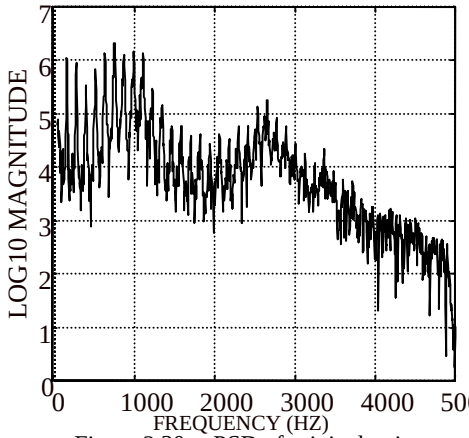


Figure 2.30a. PSD of original voice.

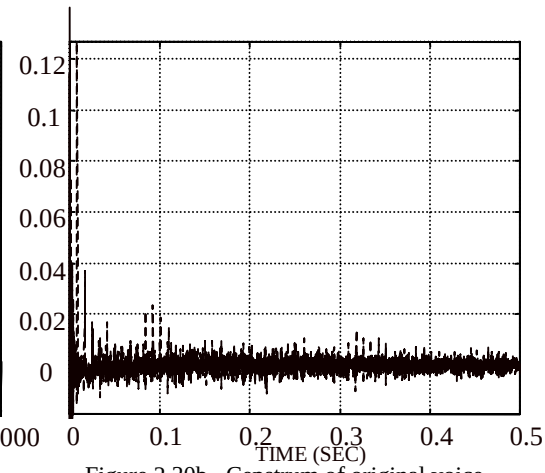


Figure 2.30b. Cepstrum of original voice.

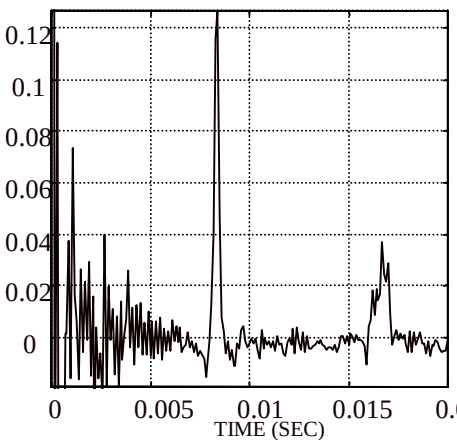


Figure 2.30c. Cepstrum (expanded scale).

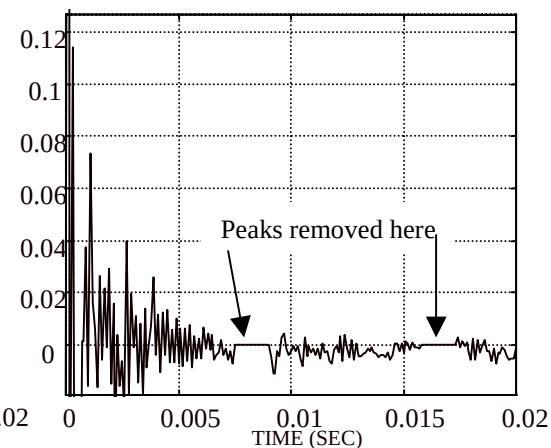


Figure 2.30d. Comb-liftered cepstrum of 2.30c.

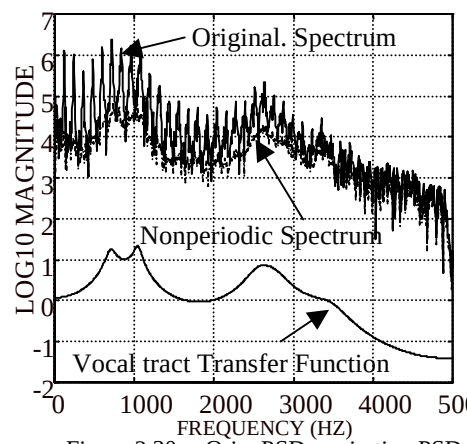


Figure 2.30e. Orig. PSD, aspiration PSD, and vocal tract.

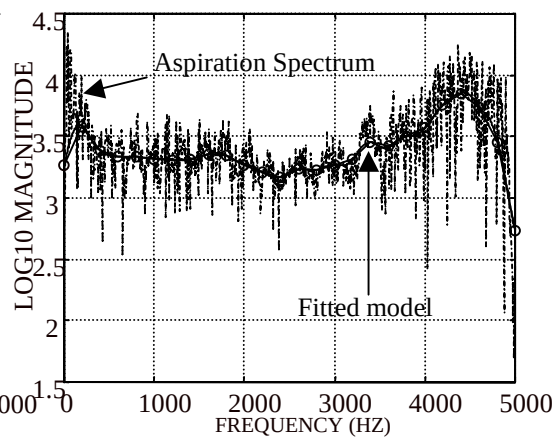


Figure 2.30f. Source aspiration PSD with vocal tract removed and fitted to 25-point piecewise-linear model.

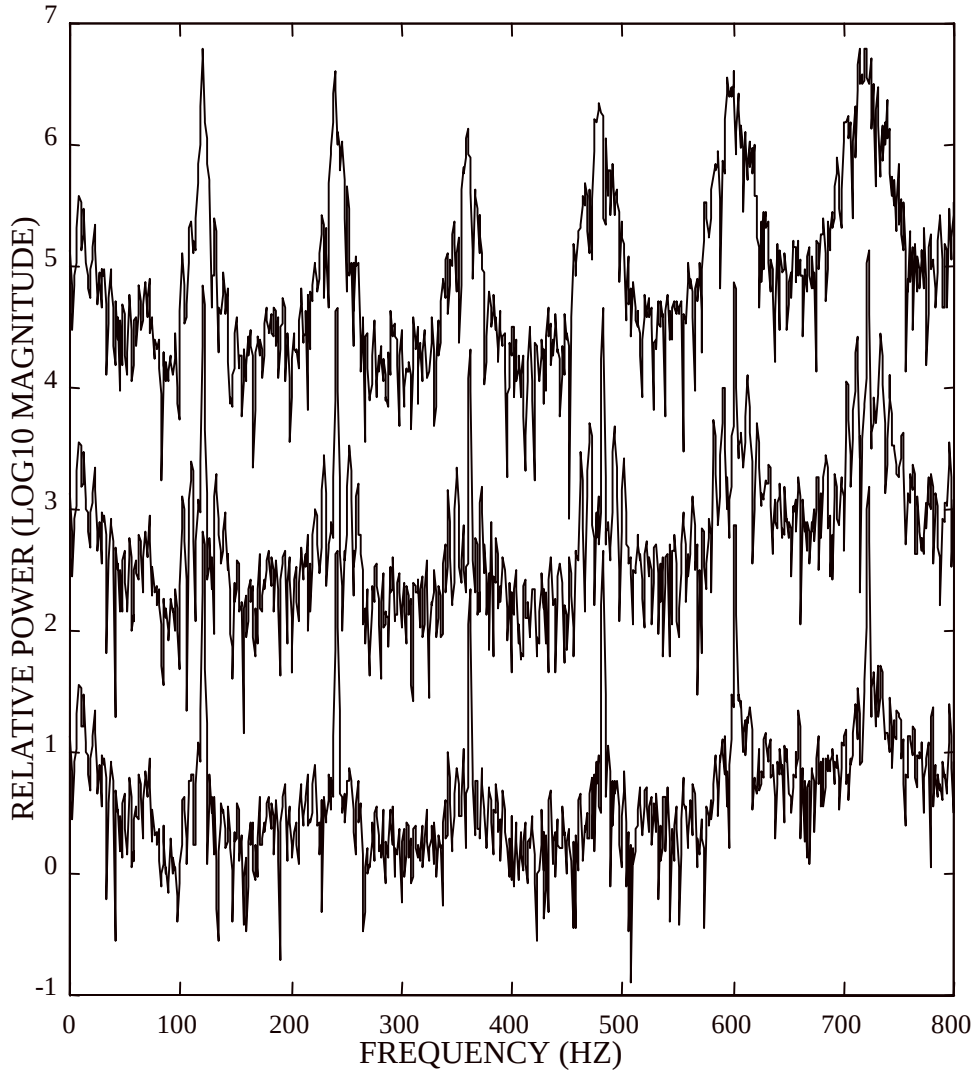


Figure 2.31. Power spectra of original voice (top), voice with FM tremor removed (middle), and voice with all FM removed (bottom).

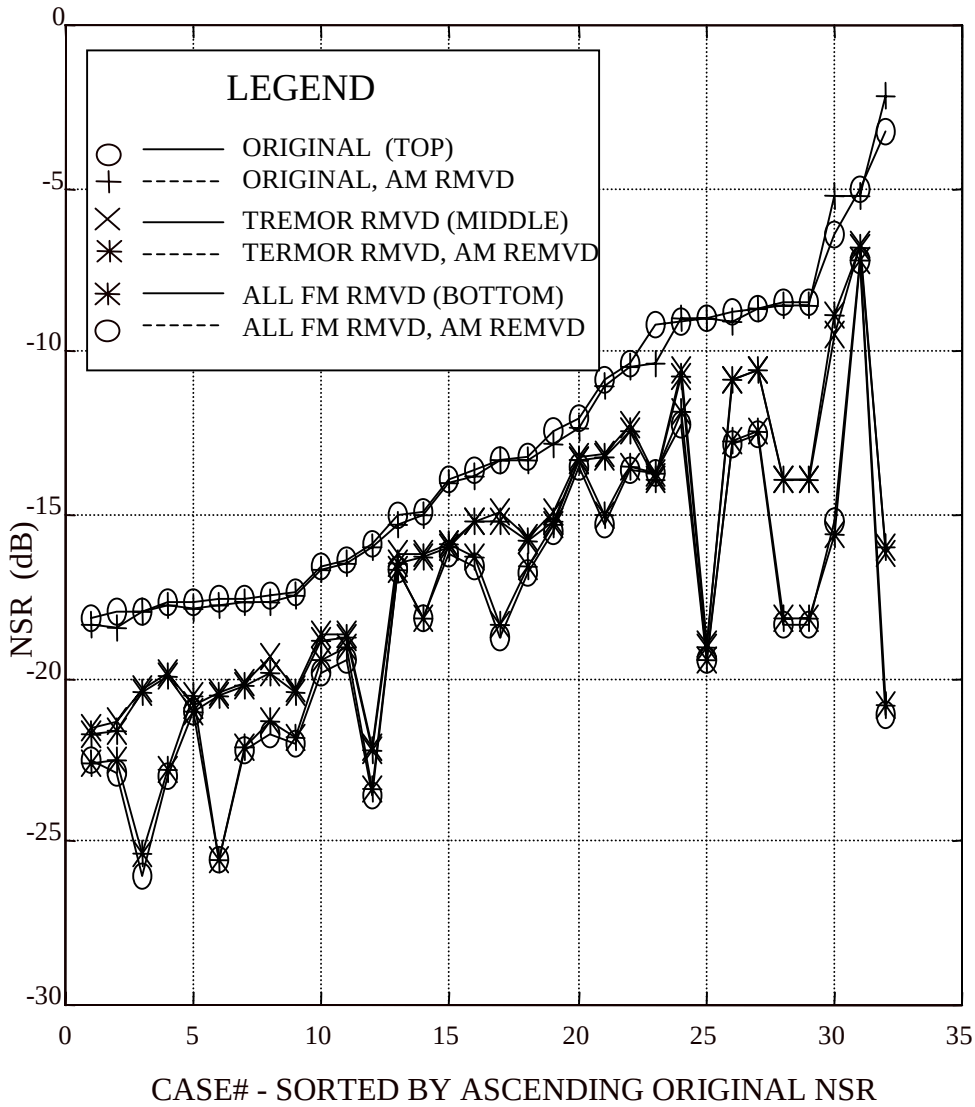


Figure 2.32. Cepstral NSR measurements for six combinations of AM and FM demodulation. The horizontal axis is case number when cases are sorted by original voice NSR. Curves fall into 3 groups: original voice and original voice AM demod. (top pair), tremor demod. and tremor plus AM demod. (middle pair), and all FM demod. and all FM demod. plus AM demod. (bottom pair). AM modulation appears to have a minimal effect.

Chapter 3

External Source Identification of the Vocal Tract

The technique proposed in Section 2.1 to identify the vocal tract and source characteristic relies on a combination of LP analysis, a set of empirical rules, and apriori knowledge of expected parameters for the vocal tract and source waveforms for normal voices. In the case of normal voices with the expected vocal tract formants and glottal source waveforms approximating the LF shape, the rules and apriori assumptions apply and the current technique works fairly well. However, in the case of pathological voices, difficulties have been experienced. In many cases, the automatic LP analysis and inverse filtering algorithms produce parameters which, upon re-synthesis, may result in inaccurate vowel quality. Even efforts to manually tweak these formant frequencies and bandwidths in the synthesizer to match the original voice vowel quality sometimes prove very difficult for expert lab workers.

This may not be surprising when the nature of the source/vocal tract model is considered. The source and vocal tract are represented as two blocks in the usual system diagram of Fig. 1.2. The glottal source is thought of as a time series and the vocal tract as a filter in the frequency domain. The source time series is filtered by the vocal tract, yielding the final voice. Mathematically, however, the source and vocal tract are both simply two indistinguishable signals or systems that are convolved in time. Thus, a perturbation in either the glottal source or the vocal tract could be compensated for by a perturbation of the opposite effect in the other and still produce the same resulting voice waveform. Thus, more than one seemingly “reasonable” source waveform could be combined with different vocal tracts to produce the same result. The same ambiguity problem manifests itself in the inverse (system identification) problem as difficulty achieving proper formant values. Fig.3.1 illustrates a specific example. The convolution of source and vocal tract to yield a voice is shown for three cases.; results from actual computation are shown. In the first case a normal voice /a/ is shown; here the typical LF source [12] is convolved with a typical vocal tract impulse response to yield a typical /a/. In the second case, the source has greatly reduced high frequency content, typical of a “breathy” /a/, but the vocal tract is normal; this gives rise to a vowel low in high frequency content. Even though the normal vocal tract formants are present, they are absent from the voice spectrum, as shown in Fig. 3.1, second row. In the third case, a normal LF [12] source is convolved with a vocal tract with almost no discernable formants. The second and third cases, however, have exactly the same resulting voice. The process of LP formant analysis and inverse filtering attempts to reconstruct the

source time series and vocal tract transfer function from the resulting voice (output) time series alone; in this case it is clearly impossible. Although this example is an exaggerated case, it nevertheless demonstrates the problem of source-vocal tract ambiguity.

One possible solution to the ambiguity is to apply an external stimulus of KNOWN form to the vocal tract and analyze the resulting response. The vocal tract could then, in theory, be identified unambiguously. Once the vocal tract has been properly identified, the voice can be inverse filtered to yield the actual glottal source waveform, which may or may not be the shape expected. In fact, external source identification of the vocal tract has already been applied in various approaches and levels of sophistication for about the last 60 years. Please refer to Section 1.1 for a summary of earlier works.

Given the success of previous investigators, the possibility of applying external source identification to aid in vocal tract identification for pathological voices seemed promising. A demonstration of external source excitation incorporating features from some of the previous efforts was successfully implemented and tested here to illustrate its application to vocal tract identification, which could then aid unambiguous source time series identification for the case of pathological voices.

3.1 Overview of External Stimulation

A simple experimental setup for demonstrating proof of principle of external source stimulation of the vocal tract is shown in Fig.3.2. The system consists of two IBM

PC compatible computers, stimulus amplifier, response amplifier, attenuater, acoustic ducting, microphone, and mounting jig. In operation, the following occurs:

1. External source signals (sinusoidal sweeps, pulses, etc) are computed and output by a D/A converter in PC1.

2. Stimulus signals are amplified and coupled to the output transducer.

3. Stimulus sound is ducted via acoustic conduit to the subject's vocal tract.

4. The resulting vocal tract response, with or without natural vocalization, is collected by a microphone.

5. Both the stimulus waveform (electrical) and the vocal tract response are sampled by A/D converters in PC2 and stored in memory.

This system provides for flexible artificial stimulation and recording of vocal tract responses in addition to the usual simple recording of pathological voice sounds on a single channel with a microphone.

A system model of the experimental setup, which proved useful, is illustrated in Fig. 3.3. In this model, the vocal tract is assumed to be a series element with the stimulus; a more detailed model might include parallel signal paths (representative of reflections of the acoustic waves). When the stimulus is present, S0 is closed. As shown, the stimulus time series is sampled by A/D channel one and passed through the dynamics of the amplifier, transducer, and environmental acoustics, which are modeled in H1(s) and H2(s). The sound stimulus may then stimulate or not stimulate the vocal tract depending

on whether or not the subject's mouth is open; this is modeled by S1 open or closed. Whether the subject is vocalizing or not is modeled by S2 closed or open. The glottal source signal is summed to the input of the vocal tract transfer function. The resulting acoustic signal is collected by a microphone and passed to A/D channel 2 via signal conditioning; these dynamics are represented by H3(s).

Using this setup, many experiments in vocal tract identification are possible. One approach to vocal tract identification proceeds as follows:

1. Time series of stimulus and response is acquired with S0 = CLOSED, S1 = CLOSED, S2=OPEN (stimulus on, mouth shut, glottis open).

2. Time series of stimulus and response is acquired with S0 = CLOSED, S1 = OPEN, S2 = OPEN (stimulus on, mouth open, glottis open). The vocal tract is held in the configuration for the desired vowel, eg. /a/, but the speaker is not vocalizing.

3. Using the stimulus/response of step 1, the transfer function H(s) = R(s)/X(s) mouth shut = H1(s)H2(s)H3(s) can be established.

4. Using the stimulus/response of step 2, the vocal tract model V(s) can then be calculated as V(s) = (R(s)/X(s) mouth open)/H(s). That is, the transfer function of the vocal tract can be recovered by dividing the transfer function of mouth open by the transfer function of mouth shut.

3.2 Validation with a Simple Tube Model

The first step in validation of the experimental setup and validation of the approach is to analyze the formants of a simple quarter wave tube model (shown in Fig.3.2 below the vocal tract depiction), which occur at:

F = C/4L, 3C/4L, 5C/4L ,

…. Where

C = Speed of sound (m/s)

L = Length of tube (m)

For this experiment, two time series were collected:

1. ha(t) is the time series response of the tube model to a 300 – 10,000 Hz chirp sampled at 40kHz.

2. hb(t) is the time series response of the same tube model, physical setup, and chirp, except a rag was solidly crammed into the tube.

Although ha and hb are not time/phase correlated, it is still possible to estimate the magnitude of V(s), ie, formant peaks of the tube.

From the model of Fig. 3.3,

$$|V(j\omega)| = |HA(j\omega)|/|HB(j\omega)|$$

The spectra of ha and hb are combined to form an estimate of the magnitude of V(s); the result is shown in Fig. 3.4. The following may be observed:

1. Formant peaks are prominently revealed, with magnitudes ranging from 60 dB at low frequencies to 10 dB at high frequencies. The peaks occur at frequencies of 660 Hz, 1980 Hz, … expected for an 11 cm tube (using C = 300m/s). Thus, the setup provides sufficient energy coupling to accurately reveal formants.

2. The more prominent contributions of H1, H2, and H3 to HA and HB are visible as the 20 dB wide peak from 0 – 5kHz, the fine 5dB ripples spaced at 94 Hz visible at low frequency (due to resonance inside the acoustic conduit), and 5dB peaks spaced at 300Hz at the higher frequencies. These "background" effects not due to the vocal tract are effectively removed when V is calculated. Thus, the effects of the transducers, conduit, room, etc are effectively minimized by the subtraction of the two spectra, resulting in the relatively smooth response in $V(j\omega)$.

3. When the chirp ends at 10000 hz, V turns to noise, as expected, since there is no stimulus energy available to reveal formants. This is excactly analogous to the failure of a pathological glottal source to reveal vocal tract formants.

4. As frequency increases, the tube's formants peaks shift and change shape, possibly due to the non-ideal realization of the tube.

3.3 Validation with a Vocal Tract

Having verified the setup for a tube model, a normal male (the author) vocal tract was tested next. In this case the open mouth vocalizing /a/ replaces the tube. In order to

achieve the same benefit of background cancellation, the mouth is either open or closed and silent (analogous to the rag stuffed into the tube). A rag was not used in the mouth. In order to compare ES (external source) analysis with the usual LP (linear prediction analysis) and FFT (fast Fourier transform), the sequencing scheme depicted in Fig. 3.8 was used. The external source executed a 0.2s pulse (to allow for possible impulse testing) followed by a 0.3s linear sine chirp from 300 to 4kHz.; the sequence repeats at about 2Hz. Simultaneously, the subject performs the following actions (each for about 1/3 of the test duration of 4 sec):

1. Mouth open vocalizing a normal /a/.

2. Mouth and glottis held in the same position, but not vocalizing.

3. Mouth shut and voice silent.

The subject's state is labeled in Fig.3.8. Both the vocalization and response of the vocal tract to external stimulation is clearly visible. The effects of the formants F2, F3, and F4 are clearly visible in the mouth open chirps in the microphone time series, and are almost absent from the mouth shut chirps (residual peaks may be due to nasal aperture effects or acoustic penetration of the closed lips).

The same spectral analysis and subtraction performed on the tube model is repeated on the vocal tract and displayed in Fig. 3.5. The top curves display the vocal tract and background (mouth open and shut) chirp responses, and the bottom curve illustrates the resulting spectral difference attributed to the vocal tract. Again, excellent signal to noise ratio is observed in the spectrum, with many fine details clearly visible. In

addition to the expected approximate formant frequencies for /a/, additional peaks are observed.

Because the subtraction process does not simultaneously measure mouth open/shut responses, the question arises as to how much movement of the articulators (which determine $V(j\omega)$) occurs between measurements. This problem is addressed in Fig.3.6, which repeats the calculation using two different mouth open chirps. As can be seen, there are slight shifts on the order of 100Hz in F2 and F4, while F1 and F3 remained within 10Hz; other peaks (some of which may not be vocal tract formants) also shifted. In this case, the subject was fairly successful in holding the articulators fixed. In the experiments that follow, this comparison is maintained to estimate time variations. If variations prove excessive, established advanced signal processing techniques that may be applied to the segments of simultaneous voice and ES to identify the vocal tract.

Another question that immediately arises is: How does ES analysis compare with standard FFT and LP (Section 2.1)? To address this question, FFT and LP are applied to the mouth open, voiced, external source quiet segment (0.3s – 0.8s) in Fig.3.8. The result is shown in Fig.3.7. Here the resulting ES calculations of Fig.3.6 are shown in the top curves, and the results of FFT and LP are shown in the bottom two curves. Several observations are possible:

1. All the formant peaks revealed by FFT/LP are present in essentially the same positions as the ES data. In this case, however, there are frequency shifts of greater magnitude than observed in Fig. 3.6; F2 appears about 200 Hz lower in the LP.

2. The ES data contains many substantial peaks above F4 that are almost invisible in the FFT/LP data. These are apparently resonances that are not excited by the glottal source.

Figures 3.9 and 3.11 continue this comparison with similar results. These shifts may be due to an even greater involuntary shift of articulators that may occur from the vocalizing state to the silent state.

In summary, it appears that the ES source data provides very precise vocal tract resonance information. Peak positions are even more sharply defined than in the FFT/LP data. Resonances not revealed with FFT/LP are shown with ES analysis. Problems with articulator drift may be overcome with more advanced testing/processing approaches.

3.4 Improvement for Pathological Source

The next question that naturally arises is how does ES perform on pathological voices? In order to obtain some experience applying ES to a simulated pathological condition, ES analysis was applied to normal male and female voices with the “breathy” condition in which the glottis does not obtain complete closure. Under these conditions higher frequency formants may be less stimulated by the more sinusoidal source waveform of the breathy condition, and FFT/LP (Section 2.1) may not reveal these peaks for determination of synthesizer settings (Chapter 4). This in turn leads to the problems of achieving proper vowel quality in simulation, discussed in the introduction to this chapter.

In testing the breathy condition, another task was added to the subject's set of actions shown in Fig.3.8. The subjects simulated the breathy condition at the start of the test, saying "haaaa" softly, attempting to achieve a sinusoidal driving function with consequent minimal high frequency formant excitation. The subjects transitioned to the normal /a/, effectively pronouncing "haaa-AAAH". This was followed by the usual mouth open, voice silent, mouth shut sequence. The result contains ES quiet segments with a breathy segment (samples 20000-35000) and a normal /a/ (samples 50000-65000), and then the voice silent/ES active segments. All these segments are recorded in one time series, as shown in Fig.3.10 (which is again the author's voice.)

The results of ES breathy testing are shown in Figs. 3.9 (normal) and 3.9b (breathy). In the normal voice of Fig.3.9, there is again fairly good agreement between ES and FFT/LP data. The FFT shows fundamental frequency harmonics of the modal /a/ well into the higher frequencies above 2500Hz; LP reveals clear peaks for F3 and F4, which are within about 100 Hz of the ES peaks. In Fig. 3.9b, the breathy FFT/LP data are shown; in this case harmonic peaks disappear at a much lower frequency. Here the LP peaks are almost absent, and the presence of F3 and F4 are revealed only by aspiration noise peaks in the FFT. The ES data, of course, remains the same, clearly revealing F3 and F4. In this case, ES is far superior to LP.

The same test is repeated with a normal female voice in Figs 3.11 (normal) and 3.11b (breathy). Again, fundamental frequency harmonics disappear at a lower frequency

in the breathy case. In this case, the LP does somewhat better in the breathy case, but the peaks of F3 and F4 are still reduced.

3.5 Summary

A major limitation of formant analysis and inverse filtering for pathological voices is the reliance upon the (possibly spectrally deficient) source to reveal formant information. With use of only the glottal source input to the vocal tract, segregation of the source waveform from the vocal tract using LP and inverse filtering may be ambiguous in the case of pathological voices, resulting in difficulty in achieving good vowel quality in synthesis. Without source energy representation across the entire spectrum of interest (which is supplied by the sharp return phase of a normal voice), resonances in pathological voices may not be detected with LP and other techniques. By externally stimulating the vocal tract with a spectrally rich source (chirp, impulse, white noise, etc), all resonances are clearly detected. Information obtained from ES analysis may then be combined with other approaches to yield more accurate formants and inverse filtered waveforms. The application of ES testing was successfully demonstrated via a progressive series of experiments. Formants of a known physical form (tube) were detected at the expected frequencies. Application of ES testing to the vocal tract revealed high resolution detection of the expected formants. In addition, comparison of ES test results with traditional LP and FFT analysis of the same voices reveals ES testing produces higher resolution, more detail, and additional resonances not detected at all by

LP and FFT. Experiments with simulated pathological (breathy) voices demonstrate a particular case in which ES testing improves formant estimation. Problems of articulator movement during ES testing may be solved by any of several technical approaches. Thus, the value of ES testing for pathological voice analysis is illustrated.

Two limitations or differences from typical FFT/LP vocal tract identification were observed. The typical 12 dB per decade decrease in signal observed in the FFT of a typical voice is not seen. This is expected as the chirp does not fall off in intensity with increasing frequency as the typical glottal source does, and because this ES measurement technique cancels out any variation in source intensity with frequency via spectral subtraction (Section 3.1). In addition, there are extra resonances in the ES data; the source of these peaks has yet to be identified.

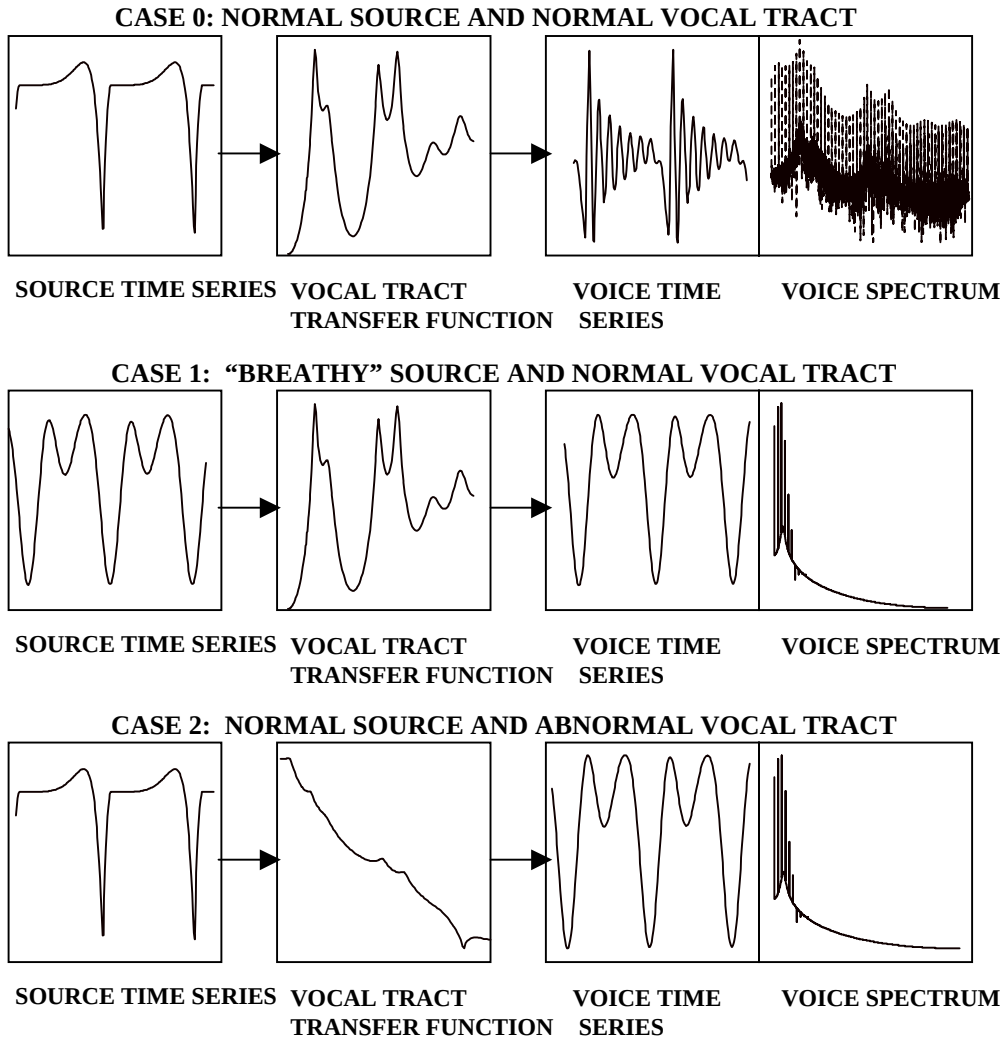


Figure 3.1. Ambiguity between the source and vocal tract models is illustrated with three examples. In case 0, a normal glottal source and vocal tract give rise to a normal voice; the normal LF glottal flow derivative waveform, normal vocal tract log spectrum transfer function magnitude, and normal voice time series and spectrum for /a/ are plotted. Case 1 shows a pathological glottal source missing the normal sharp return; it is convolved with a normal vocal tract model to give rise to a pathological /a/ with sinusoidal appearance and missing high frequency formants. This voice is perceived as "breathy." Case 2 combines the normal glottal source with a vocal tract model with greatly attenuated high frequency formants; the resulting voice time series and spectrum is exactly the same as case 1. Thus, working backwards from the resulting time series of cases 1 and 2 (as is attempted in inverse filtering and formant analysis), it is impossible to arrive at the correct vocal tract and source model without additional assumptions or information.

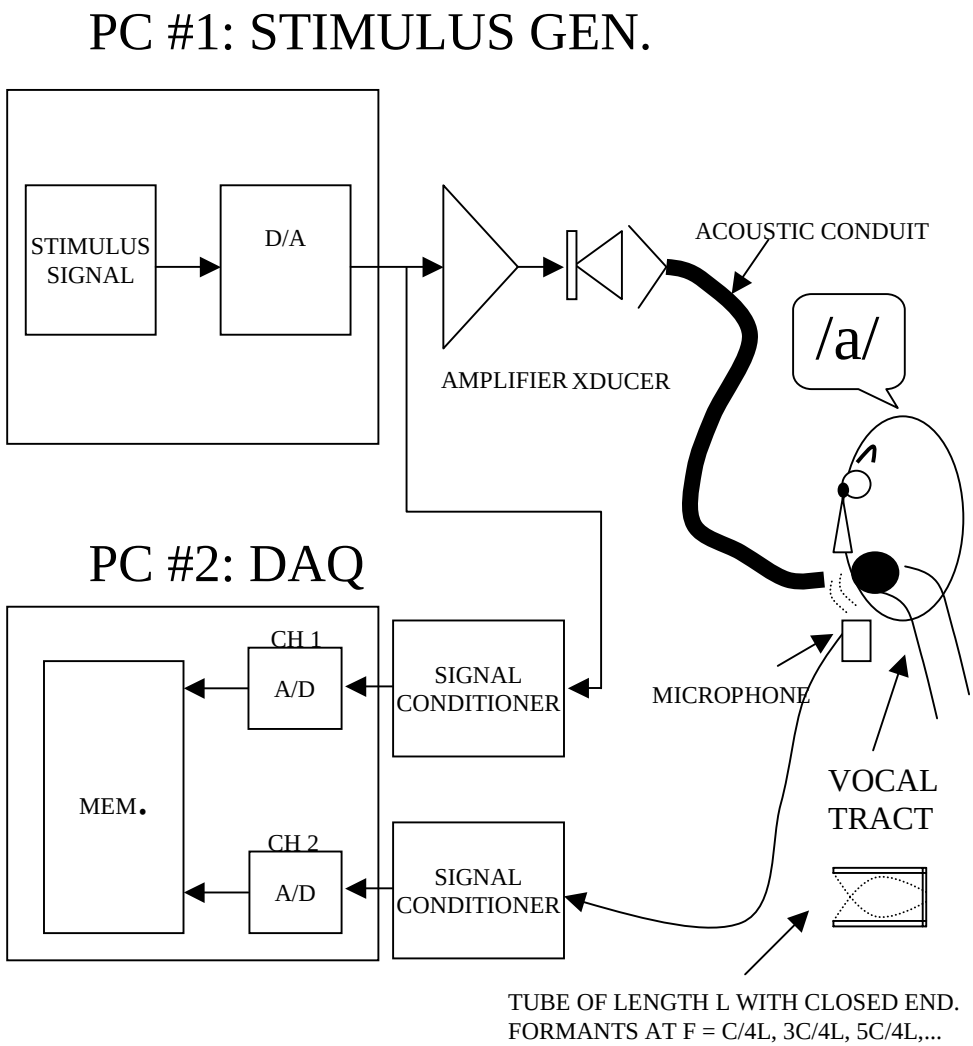


Figure 3.2. System setup for external stimulation of the vocal tract. Two PCs are used: PC #1 generates a stimulus signal to excite the vocal tract. An audio amplifier and transducer transform the stimulus to sound, which is then conducted via an acoustic conduit to either a quarter wave tube model or the subject's vocal tract. PC #2 is used to record the stimulus and response signals. A sample of the stimulus is acquired on channel 1 of an analog to digital converter and is stored to memory. The response is recorded with a small microphone, conditioned and acquired on channel 2 of the analog to digital converter. The tube model is used for proof of principle and system validation.

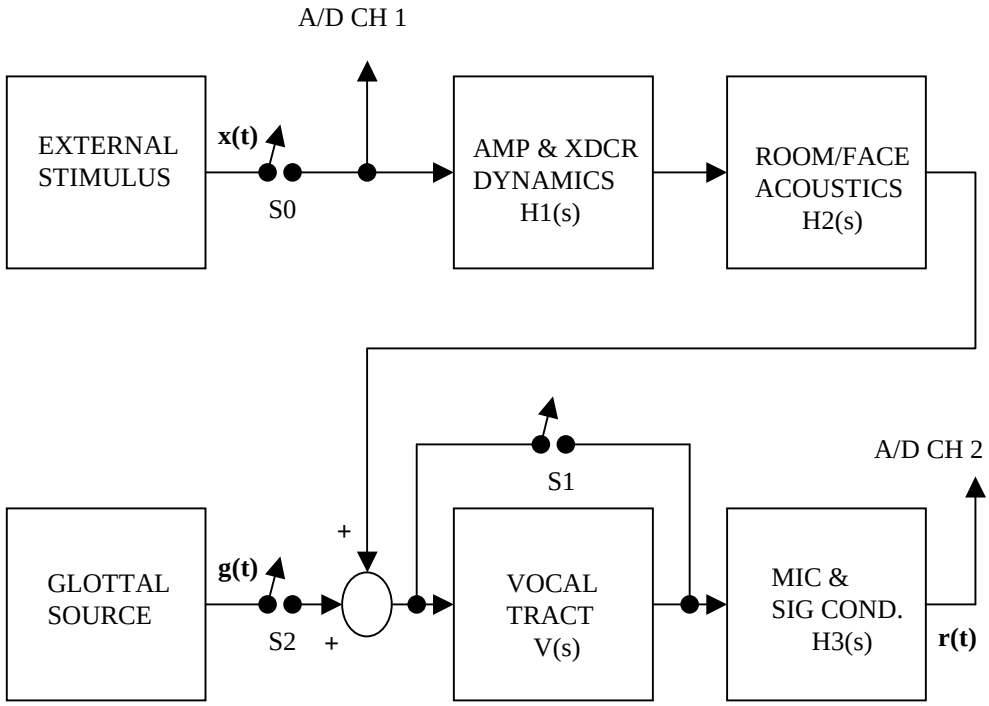


Figure 3.3. Model of external source formant analysis system. In addition to the normal glottal source g(t) and vocal tract V(s) pathway, an external source x(t) is added to stimulate the vocal tract. Stimulus electronics, transducer, and environmental acoustics are modeled in H1 and H2; microphone and signal conditioning are modeled in H3. Switches S0 and S2 represent presence/absence of x(t) and g(t); S1 represents presence/absence of the vocal tract (mouth open or shut).

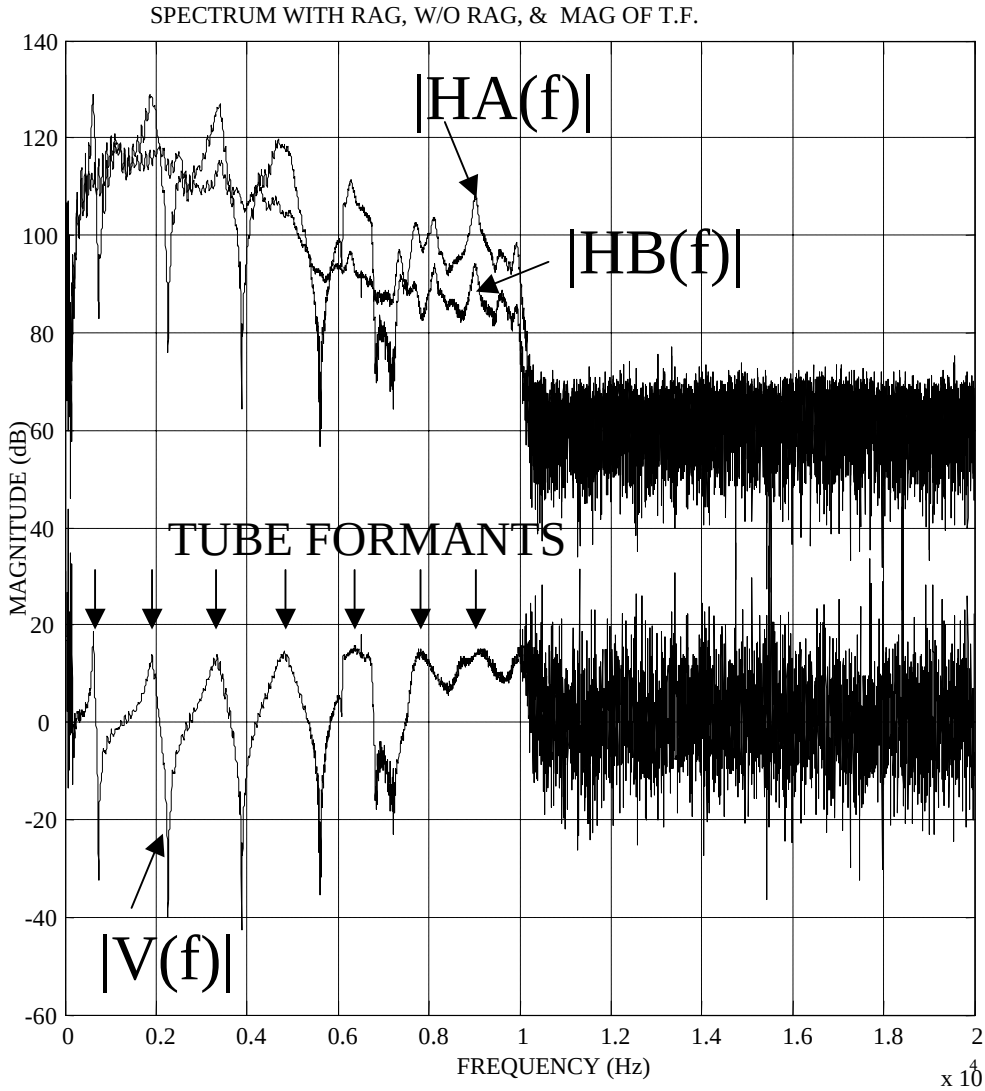


Figure 3.4. Validation of external source testing setup using a simple quarter wave tube closed at one end. HA is the spectrum recorded by the microphone near the open end of the tube when a 0 – 10KHz stimulus sinewave sweep is executed. HB is the "background" spectrum recorded when the tube is stuffed with a rag, effectively eliminating the tube formants. V is HA – HB, revealing the expected quarter wave tube formants. All spectra become noise above 10KHz at the end of the sweep.

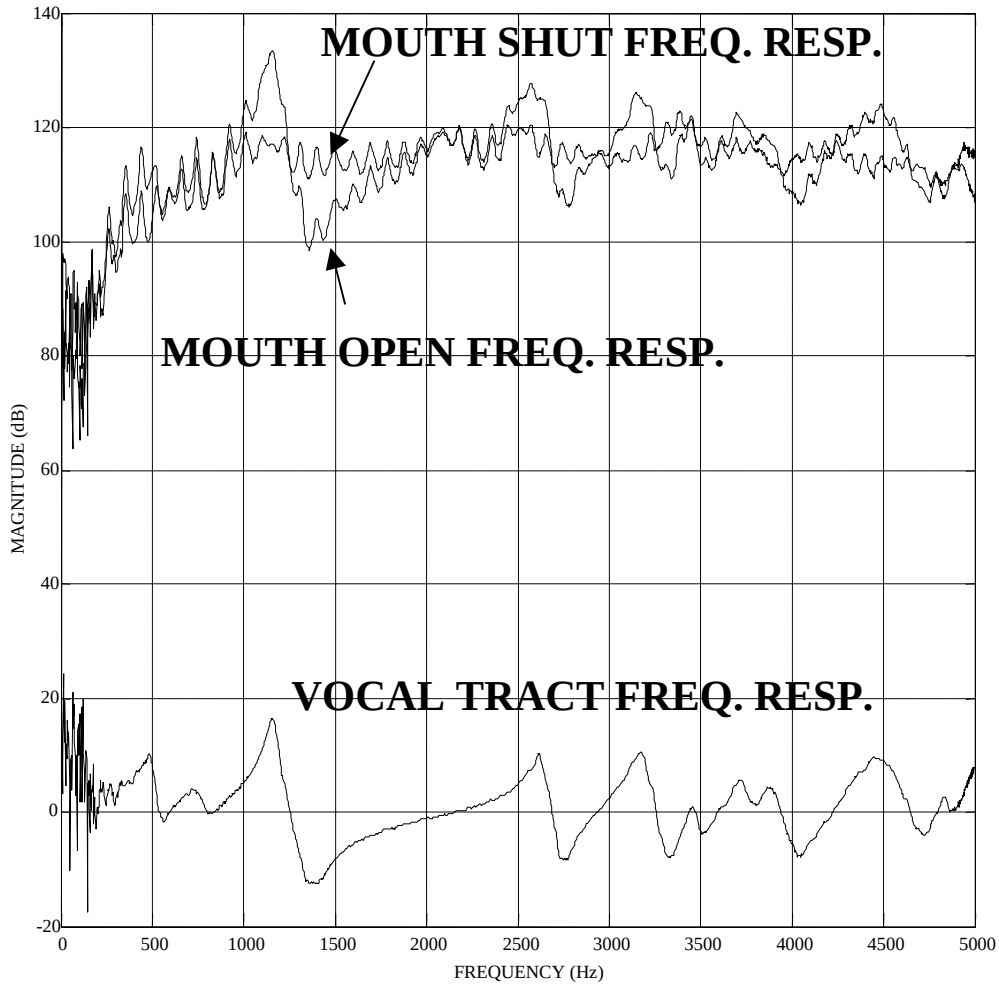


Figure 3.5. Vocal tract transfer function for /a/ with a normal male voice. Analogous to Fig 3.4, the chirp response spectra of mouth open/shut (top curves) are subtracted to yield the vocal tract frequency response (bottom). The expected formants for /a/ are present, plus additional peaks. Ripples at 94 Hz spacing in the top curves due to acoustic conduit resonances are effectively cancelled by the subtraction process.

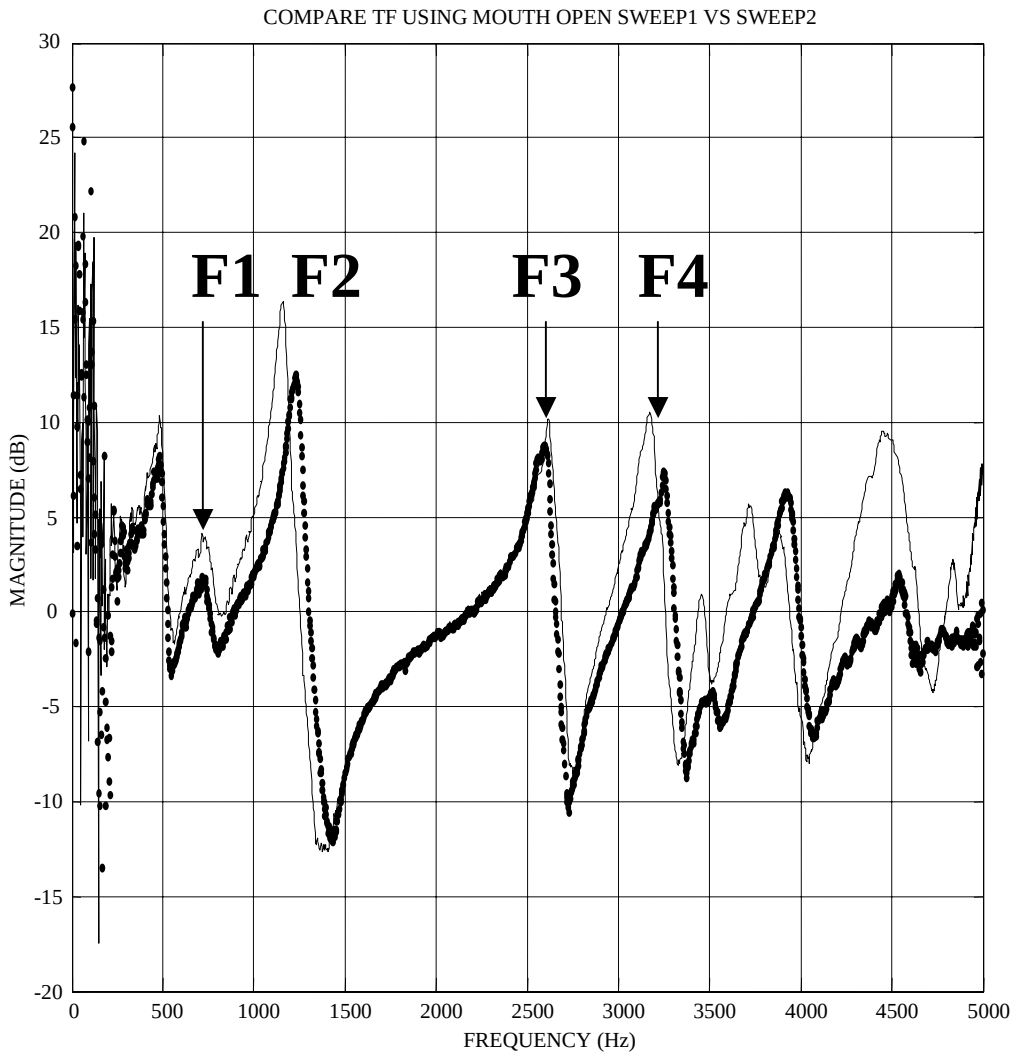


Figure 3.6. Drift in formant peaks during testing. In order to estimate effects of articulator movements, the vocal tract calculation is repeated using different chirps from the same time series. In this case, the peaks remained fairly constant with about 100 Hz change at 3200 Hz (3%).

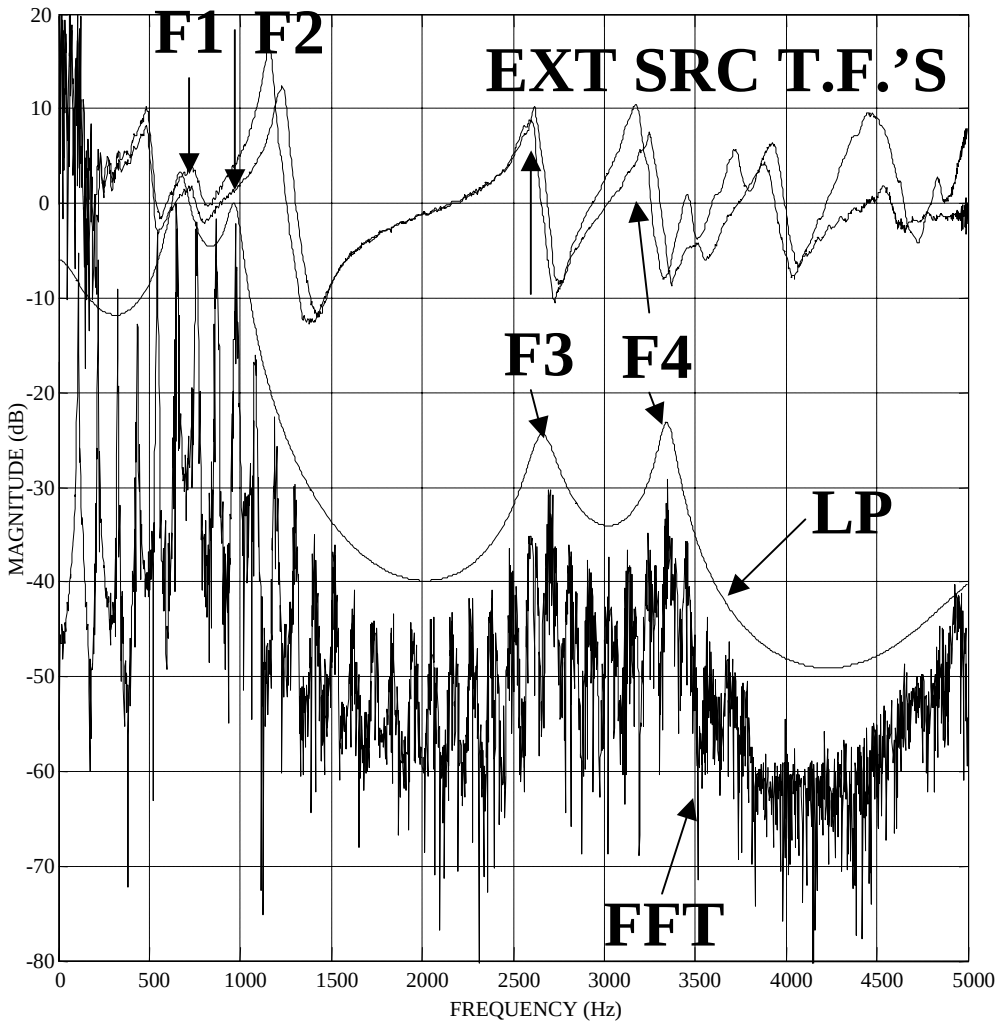


Figure 3.7. Comparison of traditional FFT/LP analysis with the ES analysis of Fig. 3.5. Formant peak shifts between FFT/LP and ES resonances are probably due to involuntary articulator movement from the vocalizing to non-vocalizing transition. Many resonances are shown in the ES data that are not revealed by FFT/LP.

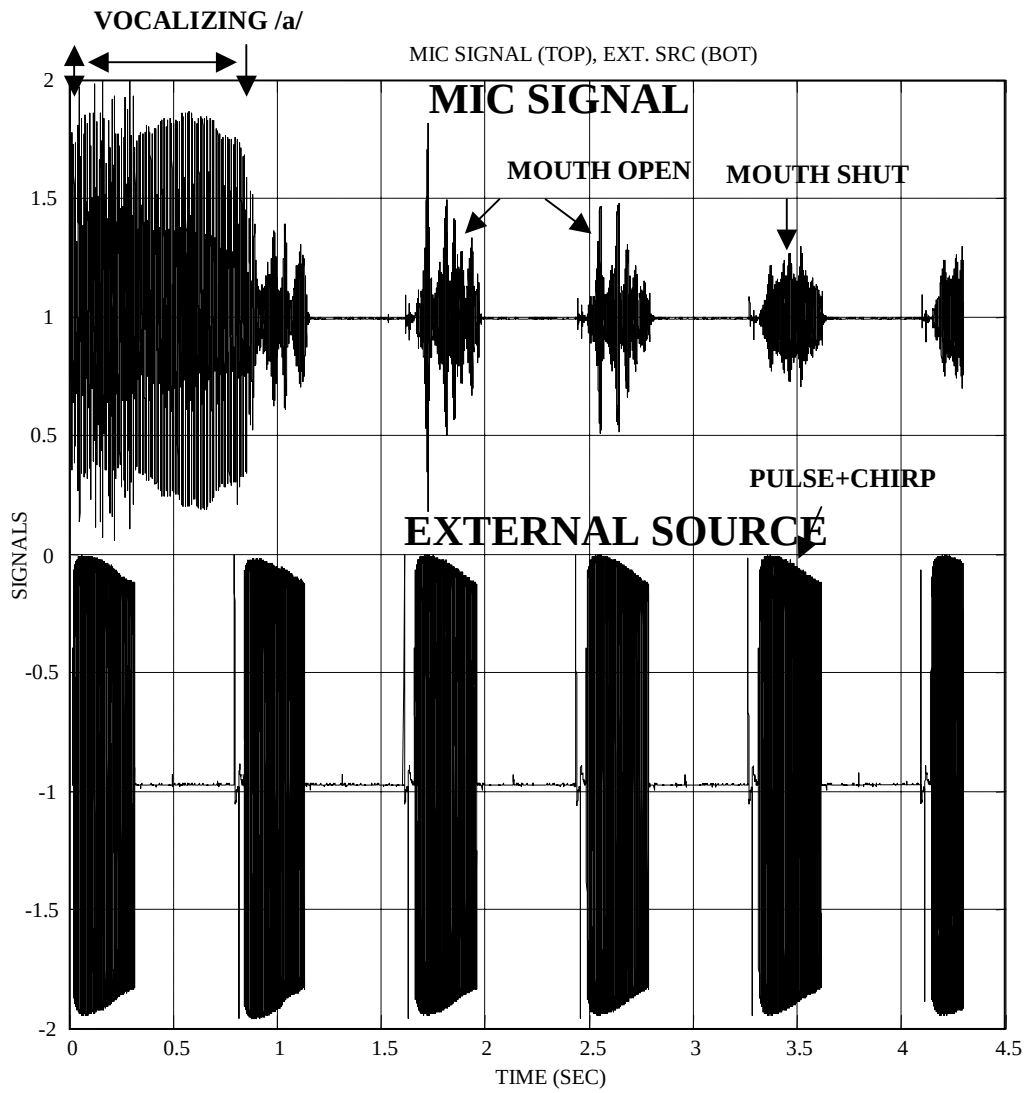


Figure 3.8. Time history of vocal tract audio output (top) and the ES (external source) stimulus (bottom). Periodic pulse + sine chirps are interspersed with periods of silence, during which vocalization (alone) can be recorded (0.3 s to 0.8 s). The resonances of the vocal tract are clearly visible in the mouth open responses at 1.7 s and 2.6 s. Note that the first chirp occurs during vocalizing; this was done to allow collection of data for future analysis using correlation techniques to separate the responses to chirp and glottis.

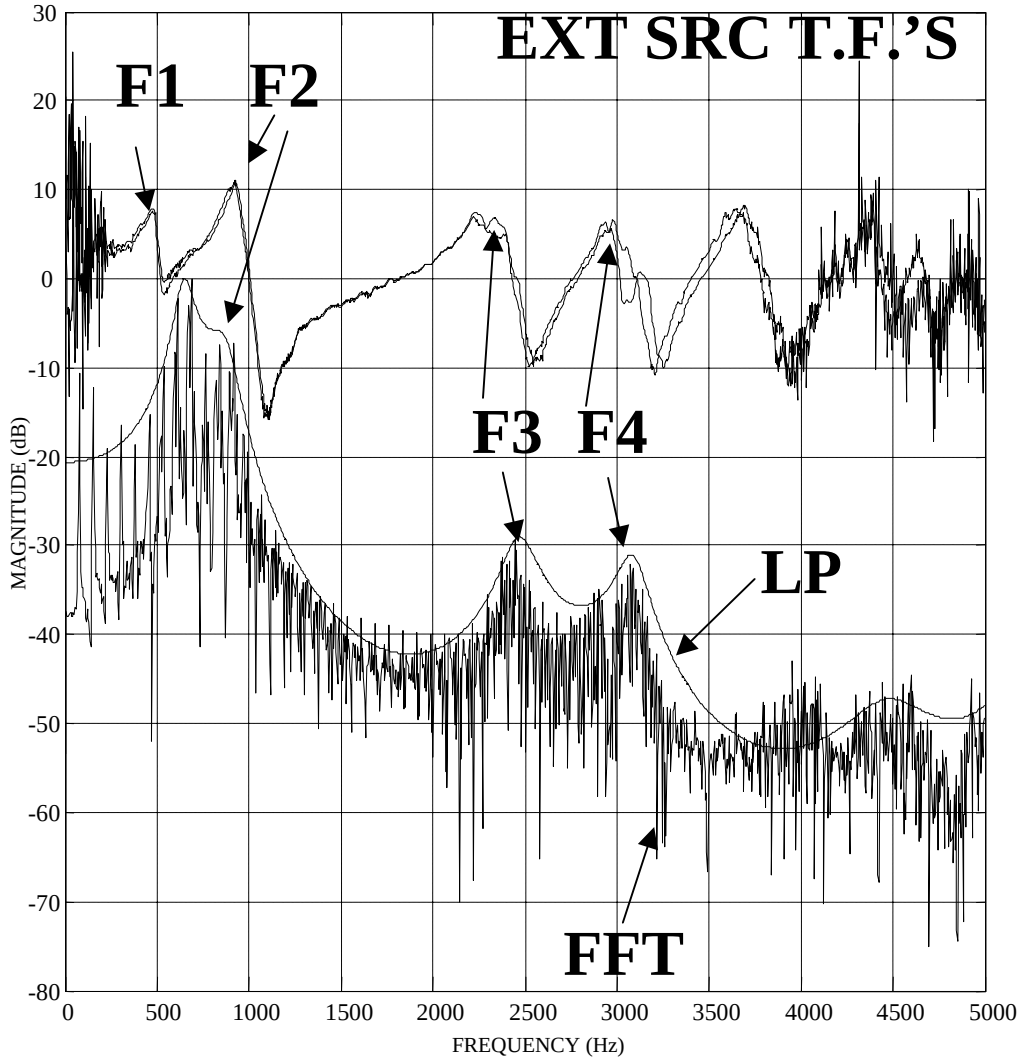


Figure 3.9. FFT/LP/ES analysis of another instance of normal male /a/. Compare with Fig 3.9b which is a breathy /a/ from the same time series.

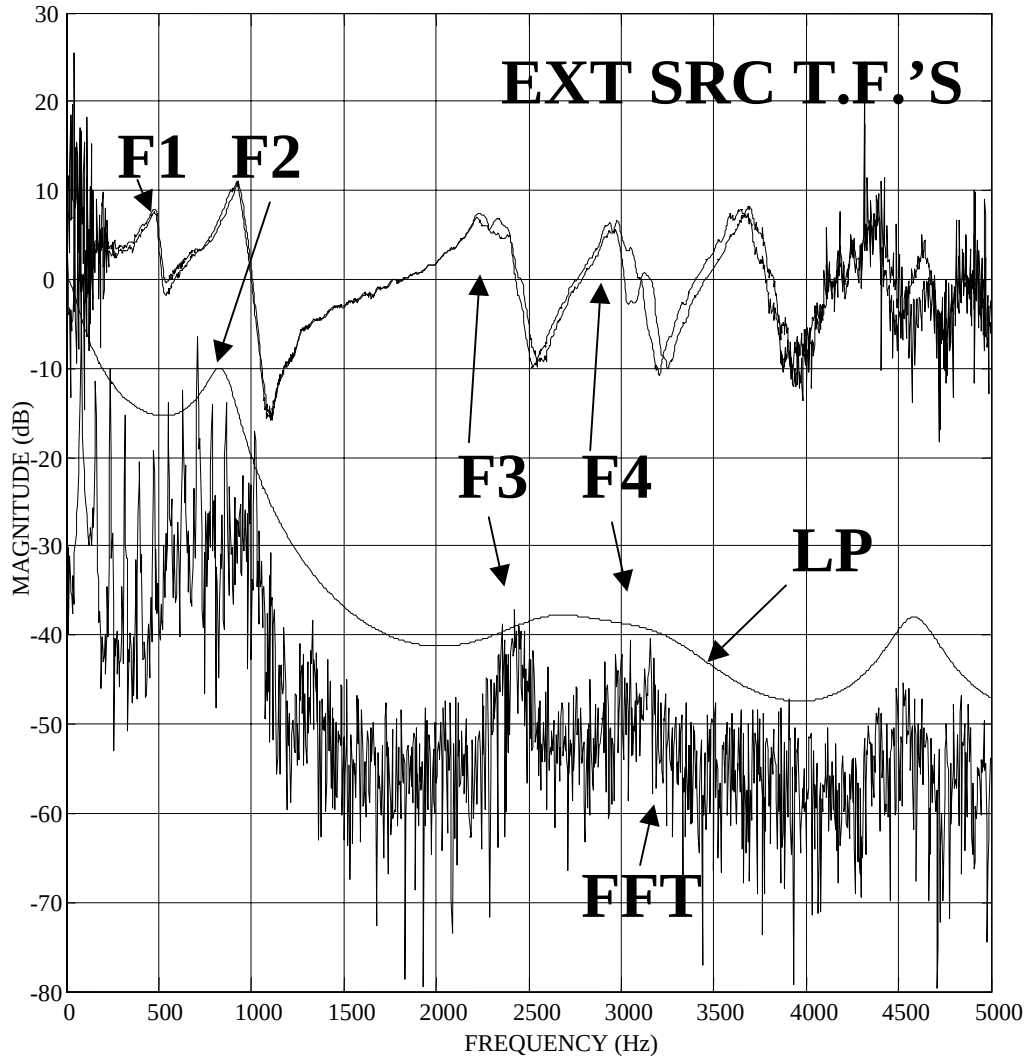


Figure 3.9b. FFT/LP/ES analysis of a breathy /a/ from the same time series as Fig. 3.9.

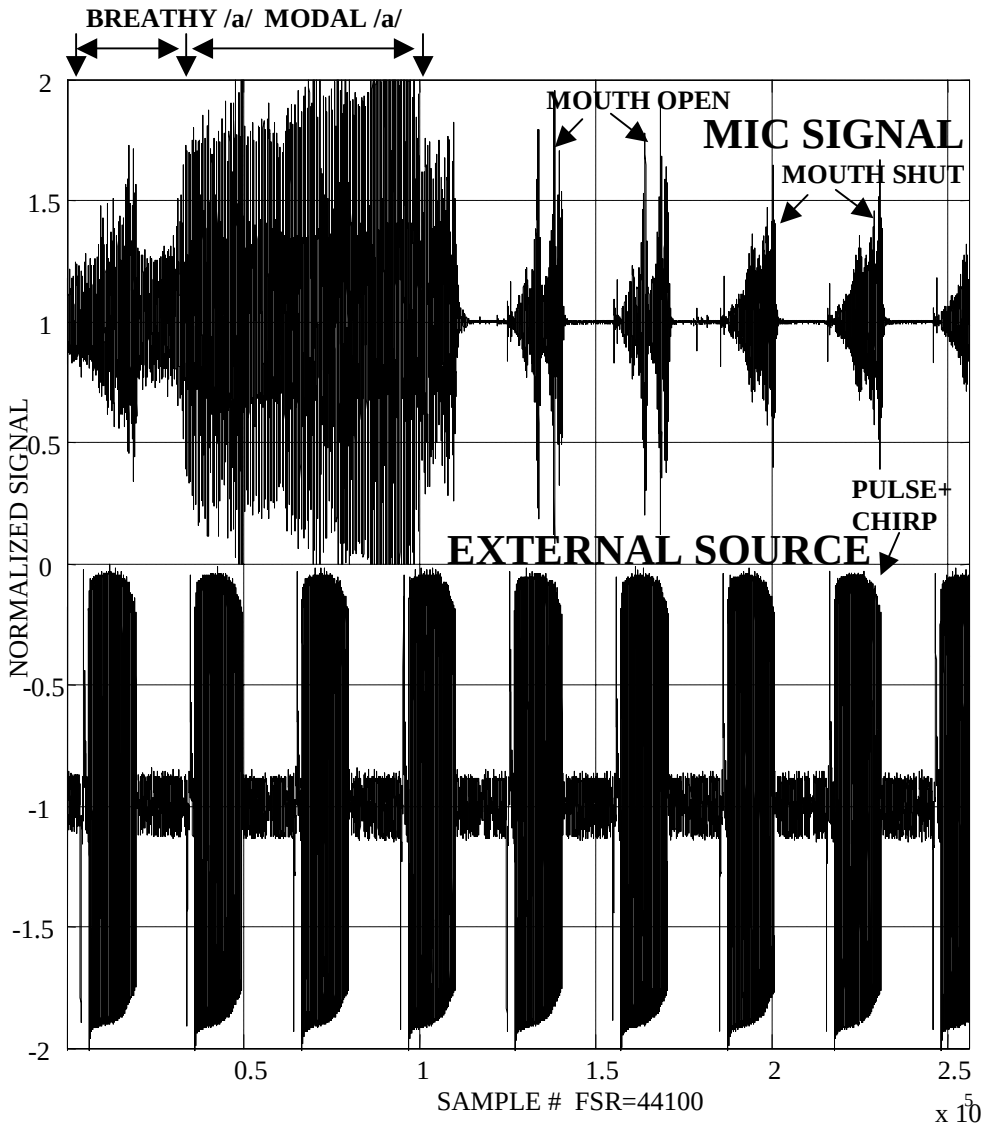


Figure 3.10. Time history of vocal tract output and ES (analogous to Fig. 3.8) with an added breathy segment at the beginning. The voice is the author's.

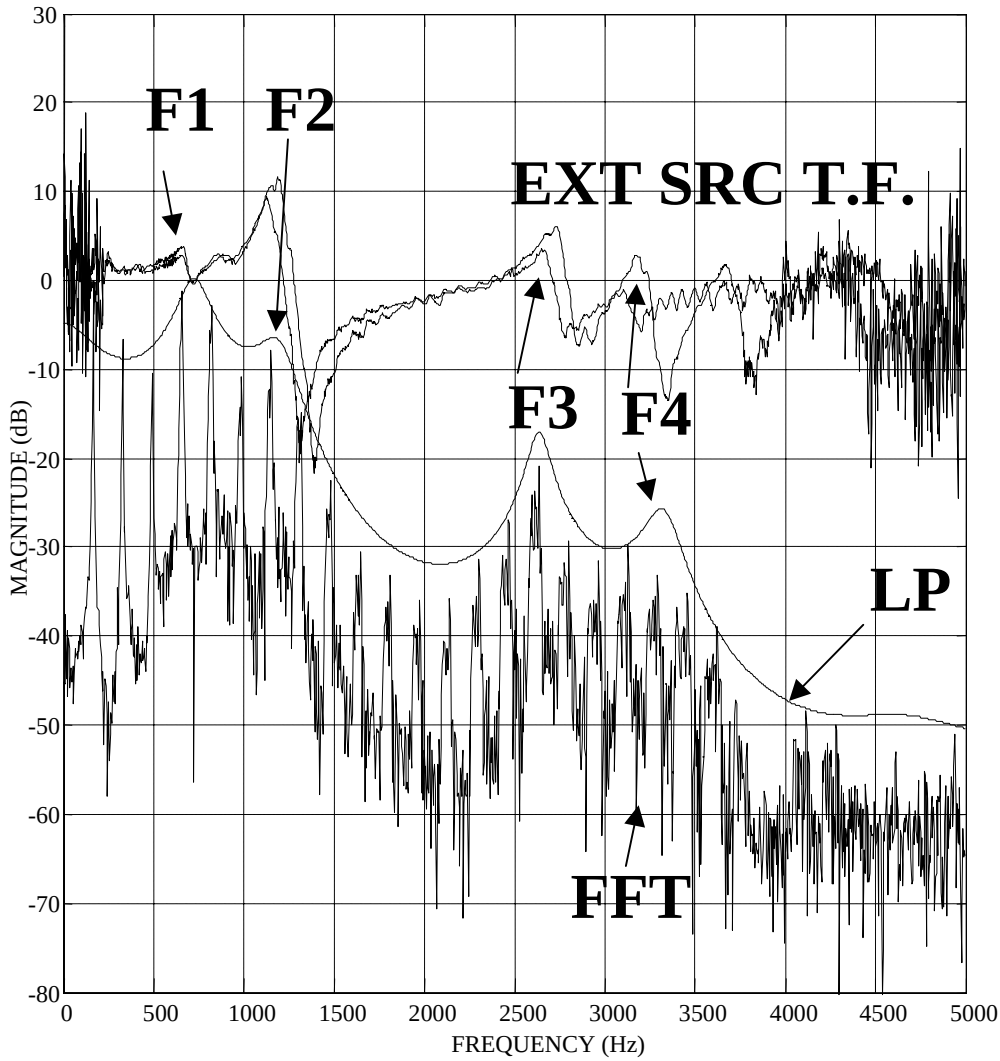


Figure 3.11. FFT/LP/ES analysis of a normal female /a/. Compare to Fig 3.11b which contains a breathy /a/ from the same time series.

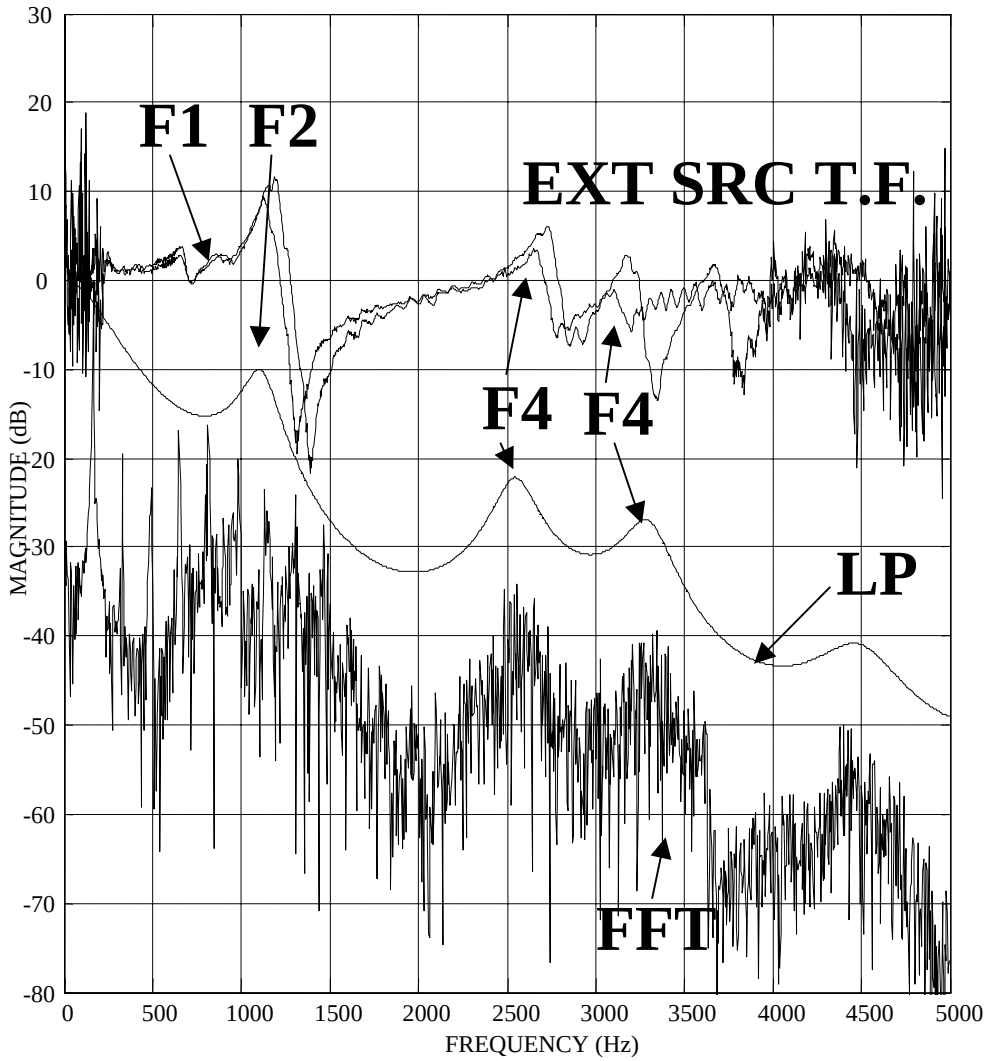


Figure 3.11b. FFT/LP/ES analysis of a female breathy /a/. Compare to Fig 3.11 which contains a normal /a/ from the same time series.

Chapter 4

Speech Synthesizers

An essential step in the study of pathological voices is re-synthesis; clear and immediate evidence of the success and accuracy of modeling efforts is provided by comparing the original and synthetic versions of the pathological voice. The effects of variations of each of the model parameters may be quickly evaluated perceptually by generating synthetic voice samples with an easily controlled synthesizer. Tests may be performed to validate analysis results, and experiments may be performed to determine the effects on the listener of variations and interactions of model parameters. This chapter describes the implementation of two synthesizers used in the study of pathological vowels: a real-time hardware synthesizer, and a software synthesizer implemented in MATLAB [26]. In contrast to most voice synthesizers, the synthesizers implemented in this study were specialized and optimized for the generation of pathological vowels for the purposes outlined in Chapter 1. Ordinary speech synthesis for commercial purposes generates connected speech of (hopefully) reasonable quality; here, the synthesizers were

designed with the model parameters described in vowel analysis (Chapter 2), and they evolved to high levels of fidelity. In some cases, the synthetic and original vowels are indistinguishable.

4.1 Hardware Real-time Synthesizer

The first efforts at synthesis were directed at construction of a real-time vowel synthesizer based on the source-filter (Fig. 1.2) model of voice production. The primary motivation was to provide a system capable of immediate feedback when changes were made to model parameters, such as formant frequencies and bandwidths, fundamental frequency, and noise levels. In order to quickly implement the system, readily available components were employed; a PC (personal computer) platform with an X86 (Intel 80X86 processors) CPU was selected. Hardware subsystems were implemented on a generic ISA (Industry Standard Architecture) bus [33] prototyping adapter PCB (printed circuit board) using commonly available IC's (integrated circuits). Software for real-time performance was written in high-speed assembly language and was designed to disable the MSDOS (Microsoft Disk Operating System). In this section, the design philosophy of the real-time synthesizer is presented, along with an overview of its hardware and software implementation.

4.1.1 Real-time and Control Concepts

In order to provide immediate and accurate synthesis with real-time response, two requirements must be met:

1. The results (synthesized sound) must be produced quickly. The results of all calculations must be performed within a time budget set by the sample period selected (usually 0.1 ms.). Updated output waveform sample points are computed based on time varying model parameters such as fundamental frequency, noise levels, etc., and the new values must be determined within the time available (one sample period of 0.1ms or one "frame time" of about 20 ms). This speed requirement was achieved with X86 processors running assembly language code.

2. The results must be supplied deterministically. That is, successive executions of the calculations must occur on a repeatable schedule, and there must be no variation in the regularity ("jitter") in the production of sample outputs to the output device (speaker or headset). Precise periodic output was achieved by keeping the computation periods within the cycle or frame budget and using crystal-controlled hardware clocking to latch outputs from the D/A (digital to analog converter). Non-deterministic behavior due to operating system activity was achieved by deactivating MSDOS, which launches interfering tasks.

Another design principle incorporated into the real-time synthesizer was the capability to perform all computations within a single sample period; the CPU is phase-locked to the sample period. Most voice synthesizers using PCs either do not produce real time output at all (e.g., Klatt [32]) or allow the CPU to run unsynchronized, filling buffers

for later clocked output. Such asynchronous computation is acceptable only if synthesizer output is to be generated, as resulting latency delays will be below perception limits in well-designed systems. However, if real-time control with dynamic inputs and feedbacks are to be incorporated, the CPU must process all feedback signals and new sensor inputs and generate updated control outputs within a single sample period, as illustrated in Fig.4.1 and Fig.4.2. Buffering and variable delays would render control algorithms unpredictable. By keeping all computations within a single sample period, the system is rendered expansible to a variety of possible closed loop controls applications. For example, slidepot controllers could be used as input devices to control a bank of formant frequencies and bandwidths, allowing very rapid evaluation of vowel quality variations.

Single cycle computation capability was maintained throughout the development of the real-time synthesizer, despite the ever-increasing computational demands of the addition of new features to the synthesizer. The evolution of the X86 processor line from 286 to 386 to 486 to Pentium kept throughput up to the increasing computational tasks, as shown in Fig. 4.3.

4.1.2 Functional Overview

An overview of the functions of the alpha implementation of the hardware synthesizer is displayed in Fig. 4.4. This version was a simple proof of concept test bed, using commercially available adapter cards and external prototyping electronics rather than the custom wirewrap adapter card of the final version. The alpha implemented up to

five formant resonators with a simple impulsive source and generated acceptable vowel sounds. An 80286 platform was used as the processor. Timing was provided by a Metrabyte CTM05 adapter card using an AM9513 counter-timer to provide a 10 kHz clock output for ISR (interrupt service routine) interrupts to the processor to initiate each control cycle and to strobe the calculated output into the D/A (digital to analog converter) data latch. A reconstruction filter (low pass filter to remove “steps” from the D/A output), audio amplifier, and loudspeaker completed the issue. The impulsive source and resonators were implemented in X86 assembly language, using 16 bit fixed point arithmetic for all digital signal processing. The 16 bit fixed point arithmetic resonator calculations had to be painstakingly scaled to prevent integer over/underflows.

Many upgrades and refinements resulted in the real-time synthesizer in its current form, as shown in Fig. 4.5. The impulsive source was generalized to include the KGLOT88 [22] and the LF [12,31] models for the glottal source flow derivative waveform. These resulted in an immediate improvement in the realism of the synthetic vowels. In addition, provision was made for inclusion of an arbitrary wave for the source, allowing offline generation of any desired waveform. Aspiration noise simulation was added to the model, again improving realism. Numerator zeros were added to the vocal tract model to allow simulation of nasal effects. An AGC (automatic gain control) function was implemented to compensate changes in output signal magnitude due to parameter changes, thus preventing overflow of the D/A but retaining full use of its input range. Provisions were added to allow time-scheduled variations in control parameters. A

flexible scheme for saving and recalling control parameters and signal time histories was implemented. Provisions were made to generate nonperiodic variations in frequency (jitter) and amplitude (shimmer), and to generate diplophonia (alternate cycle variations in period and amplitude). All hardware functions were consolidated on an ISA prototyping adapter card implemented with wirewrap. The synthesizer was initialized under MSDOS 6.2 on a Pentium 120 MHz or higher platform. In the following sections, the hardware and software implementations of the real-time synthesizer are briefly described.

4.1.3 Hardware Implementation

In order to achieve true real-time performance, it was necessary to provide hardware extensions to the PC platform. Efforts to use components native to the motherboard alone proved unsuccessful. The resulting hardware extensions resulted in a final issue of 17 integrated circuits on a wirewrap PCB adapter card plugging into the PC system bus. Interface of the synthesizer hardware to the PC is via the 16-bit ISA bus, as shown in Fig.4.6. The hardware components of the real-time synthesizer consists of the following subsystems:

1. An independent clock generator. In order to create precisely timed output samples from the D/A, it is necessary to supply an accurate timing signal to latch each new output from the synthesizer calculation. Software timing schemes are unreliable and are subject to jitter (variations in cycle to cycle period lengths). An easily usable source of timing

signals for this purpose was not available on the PC motherboard. Instead, an independent programmable, crystal-controlled generator was used. The generator is initialized with the divider ratios to produce the desired clock frequency (usually 10kHz).

2. D/A converter. This device converts binary 16-bit integer output from the synthesizer into an analog voltage to drive the reconstruction filter. A 2R ladder type converter was selected for its easily used high speed parallel interface to the data bus.

3. Timing and control. This subsystem provides the "glue logic" and control signals for synchronizing the CPU and with the various hardware subsystems; it was implemented in SSI (small-scale integration) TTL (transistor-transistor logic) and PAL (programmable array logic) devices.

Briefly, in operation the hardware subsystems perform as follows:

1. The clock generator is programmed with the proper control words and divider ratios to supply the basic synthesizer sample rate (usually 10kHz). This rate then determines the critical cycle time budget for one cycle control (usually 0.1ms).

2. The PC's operating system is then disabled and the PC's interrupt controller is reprogrammed to look at interrupts from the synthesizer.

3. The timing and control circuitry is then enabled via the control port to begin sending the clock pulses as interrupt signals to the CPU.

4. As each output sample is generated by the real-time software, it is written into the D/A latch when it becomes available (which must occur prior to the end of each control cycle).

5. The next clock pulse simultaneously initiates a new control cycle and enables the latched D/A value as an analog output.

6. The hardware continues to generate timed analog outputs until the system is stopped.

4.1.4 Software Implementation

The real-time synthesizer software performs generation of output time series simulating the voice signal. It is responsible for initializing the system, performing user interface, calculating each new voice sample output within the cycle budget (usually 0.1ms), and detecting error conditions such as cycle overrun (the software takes too long to calculate the next sample). A brief overview of the software is displayed in Fig. 4.7.

The core of the program is the ISR (interrupt service routine), which responds to each external clock pulse to begin each new control cycle (right side of Fig. 4.7). Each cycle performs all necessary calculations to generate the next output sample for the synthetic voice; these tasks include:

1. Update the current time counter.

2. Generation of current values of any time varying parameters, such as variable formant frequencies or user supplied schedules.

3. Generation of the current value of the source signal in the source-filter model, according to the current time and the interpolated value on the type of waveform being generated and any random noise component

4. Updating the 2nd order digital resonators that simulate the vocal tract.

5. Generation of the next voice signal output value.

6. Performance of the AGC (automatic gain control) algorithm.

7. Writing the final scaled output signal into the D/A converter input latch.

8. Detection of control cycle overrun. The user is warned when the CPU takes longer to calculate the next output sample than there is time available in the control cycle budget.

As discussed in Section 4.1.3, the ISR is invoked by a hardware interrupt to the CPU from the external real-time hardware clock, which is free from any jitter (period variation). The ISR must complete all necessary computations before the next interrupt. All ISR code was written, therefore, to attain maximum speed. Assembly language was used to directly control the X86 processor hardware to avoid inefficiencies imposed by a compiler. Despite often-heard comments to the contrary, C language proved to impose a fairly consistent speed penalty of about a factor of 10. In addition, high-speed coding techniques were adopted wherever possible. These included use of table lookups (rather than computations), replacement of loops with simple repetition of code blocks, use of pointers to minimize data movements, and the use of the numeric coprocessor for all arithmetic. By use of these techniques and the advances in throughput of the X86 processor line, it was possible to achieve a flexible vowel synthesizer with instant response to user inputs.

The remainder of the real-time synthesizer software, illustrated in the left side of Fig. 4.7, consists of initialization and background tasks. The initialization tasks prepare the PC for real-time operation and set up control parameters for the synthesizer. Prior to operation, it is necessary to disable the extant PC operating system (MSDOS) and

interrupt control scheme, which is inherently useless for hard real-time applications due to its intrinsic non-deterministic (inconsistent) behavior. The operating system's interrupts are disabled and replaced with the hardware interrupt from the crystal-controlled clock by re-initializing the 8259 interrupt controller (or its ASIC functional equivalent in later PC's) and other operations. Similarly, conditions for normal operation are restored at the termination of the synthesizer execution, thus eliminating the need to reboot the PC.

In addition to initialization, the background software handles the GUI (graphical user interface) to process keyboard commands entered during operation of the synthesizer. Upon completion of the ISR, control returns to the background task, which performs lower priority functions, including:

1. Detection of keyboard inputs.
2. Writing new display values to the CRT.
3. Loading input data.
4. Writing output data.

The background software was written in C, to take advantage of its ease of use and more powerful programming features, since speed is a secondary requirement here.

4.2 Software Synthesizer

Following successful implementation of the real-time hardware synthesizer, effects concentrated on refinements of the synthesis algorithms with the goal of achievement of ever-higher levels of fidelity and match to the original voice (see Fig. 6.1). Concentrated efforts were applied to the source waveform and the nonperiodic features of the voice, including FM, AM, and aspiration noise. In order to more quickly evaluate algorithms for synthesis and to allow for the implementation of complicated procedures at the sacrifice of real-time response, a purely software version of the synthesizer was implemented in the MATLAB [26] interpreter. MATLAB permits matrix objects to be created and manipulated with high-level functions designed for controls and signal processing applications, as well as supplying easy to use graphical output. It has become an industry standard for signal processing analysis.

In the following section, an overview of the software synthesizer is presented: the basic design philosophy and computational approach are described. Algorithms used to implement synthesis of the innovative pathological voice model parameters described in analysis in Chapter 2 are then described.

4.2.1 Functional Overview

In contrast to the real-time synthesizer, the software synthesizer operates offline in what used to be called “batch” mode. Each time synthesis is invoked, all control parameters are loaded and then computation is initiated and runs asynchronously to completion (generation of the complete output time series). The resulting output time

series is stored in memory; the user then initiates audio output to the loudspeakers via the PC's sound card adapter (which has MATLAB drivers). Offline generation permits non-causal (out of time sequence) signal processing techniques, which eases the process of synthesis by permitting various segments of the output to be generated in computationally convenient stages rather than in the strict output sample time sequence. For example, an AGC (automatic gain control) is not necessary, because the whole output time series can be scanned for its maximum and scaled accordingly before it is output to the D/A. Also, different components of the result such as a complete one second periodic LF source time series and a complete spectrally-shaped aspiration noise time series can be completely calculated in different modules and later recombined in the appropriate ratio depending on the desired NSR (noise to signal ratio).

An overview summarizing the major features of the software synthesizer is displayed in Fig. 4.8. After invocation, the synthesizer inputs a set of user-specified control parameters defining the first synthetic vowel to generate; initializations are performed, and the GUI (graphical user interface) is started. The first computation of the output time series then performed; as shown, the major portion of this effort is the generation of the source time series, which may incorporate nonperiodic features such as high and low frequency AM and FM and aspiration noise. Computation of the output time series may take from 1 second to minutes depending on the features enabled. In its normal operation state, the synthesizer waits for user keystrokes to invoke various tasks, such as modifying a model parameter, generating audio output of the synthetic vowel, or

loading/storing parameters or time series to files. The three main GUI's of the synthesizer are shown in Fig. 4.9a-c. The primary screen (Fig. 4.9a) allows the user to access sliders and buttons to control I/O, invoke playbacks, and vary model parameters controlling fundamental frequency, power, and the nonperiodic features. Two sub-screens allow interactive modification of LF parameters (Fig. 4.9b) and formants (Fig.4.9c).

4.3 Summary

The function and implementation of two synthesizers has been explained. A hardware-based real-time synthesizer was constructed with the capability of generating immediate responses to changes in voice model parameters. The real-time synthesizer was designed with an expansible architecture allowing easy addition of new features and model parameters, deterministic real-time performance, and the capability to perform real-time closed loop control by performing all system computations within a single sample period. A software synthesizer based on MATLAB was implemented to provide the capability to quickly code and evaluate complex algorithms.

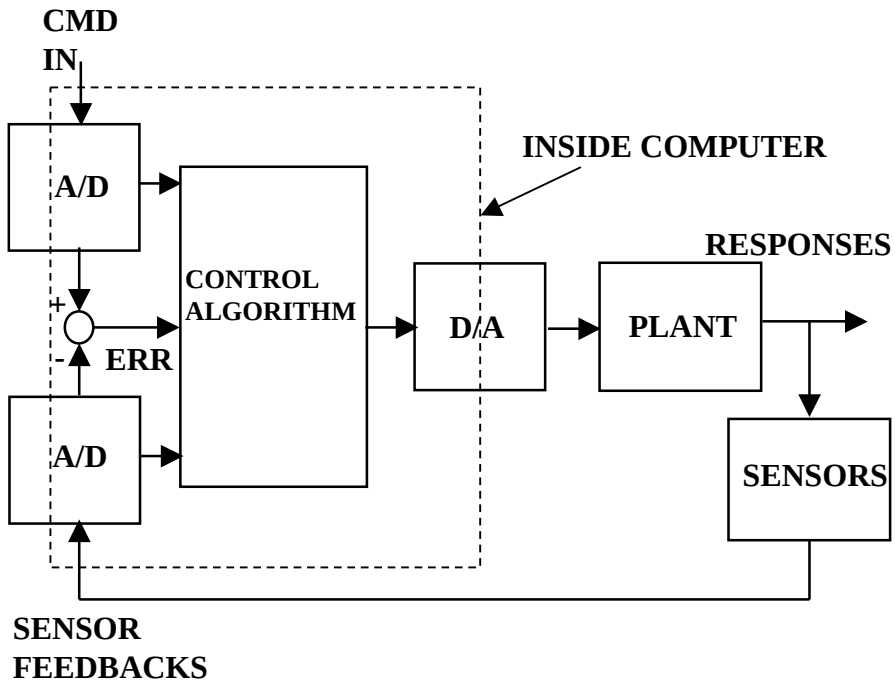


Figure 4.1. Real-time digital control loop. The plant (eg., loudspeaker) is controlled by a control algorithm (eg., vowel synthesizer). User inputs (eg., from a virtual reality glove), sensor feedback (eg., measured noise level), and calculated error signal are input to the control algorithm and must generate updated control outputs within a single sample period

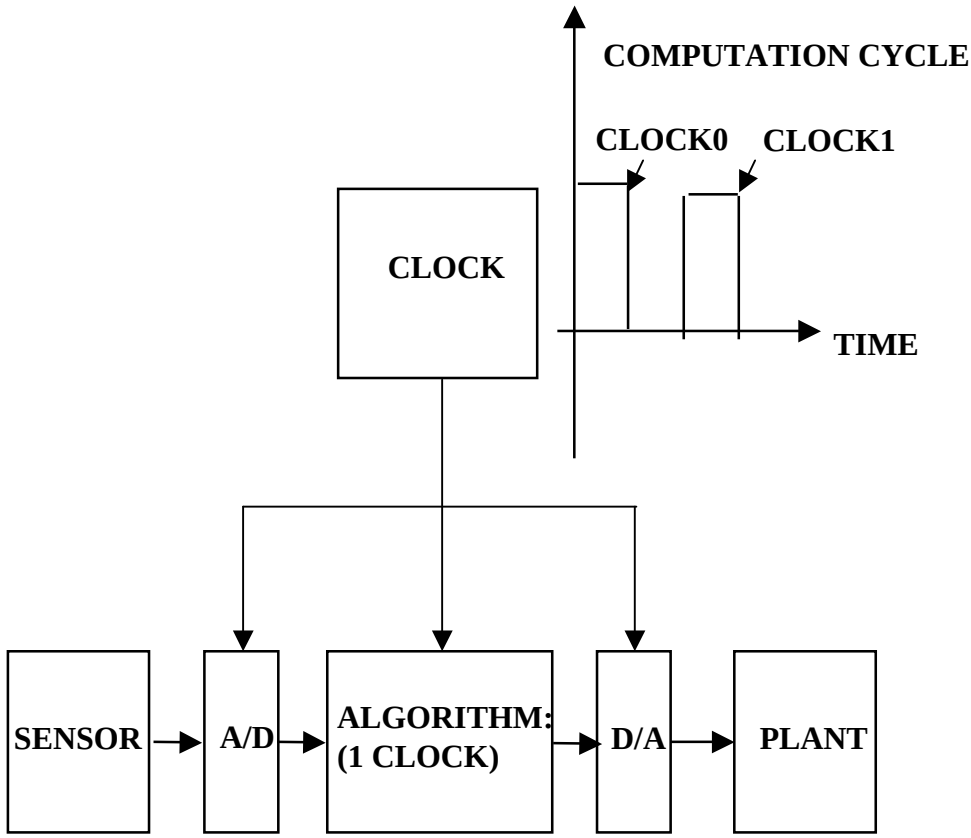


Figure 4.2. Real-time control timing. For real-time digital control, the entire computation cycle must be completed in one sample period. The A/D's provide comand and sensor inputs to the CPU on clock0. The CPU performs control calculations between clock0 and clock1. New D/A values are output to the plant on clock1.

CPU CHIP	CLOCK Mhz	CYCLE/ MULT	uSEC / MULT	uSEC / 10 RESNTR	% 100uS BUDGET
8086	**5.0**	**168**	**33.6**	**1848**	**1848**
80286	**12.5**	**168**	**13.4**	**737**	**737**
80386	**33.0**	**53**	**1.6**	**88**	**88**
80486	**66.0**	**13**	**0.20**	**11**	**11**
80586	**120.0**	**1**	**0.0083**	**0.46**	**0.46**

Figure 4.3. X86 processor performance progression. The numerical processing speed of the PC CPU's increased dramatically over the progression from 8086 to 80586 (Pentium) as both clock speed and processor efficiency improved. Column 2 is CPU clock speed, column 3 is number of clocks to perform a multiplication in the formant resonator calculation process (16 bit fixed point for the 8086 – 80286, and 64 bit floating for the rest). Column 4 is the time in microseconds to perform the multiply. Column 5 is the time to perform all the calculations for ten formant resonators. Column 5 is the percentage of a 10 kHz cycle budget (0.1 ms) consumed by calculations for 10 resonators. It is noteworthy that the transition from 16 bit integer to 64 bit floating actually resulted in *less* CPU time.

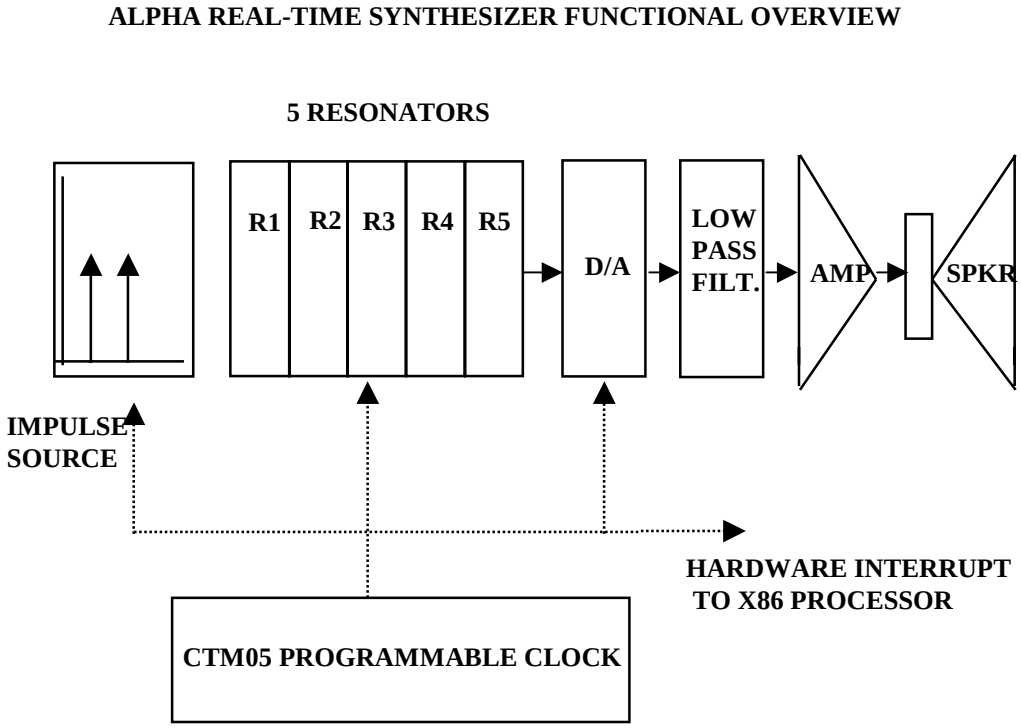


Figure 4.4. Overview of the alpha real-time vowel synthesizer. This first implementation used commercially available adapter cards and prototyping circuitry external to the PC platform. It used a simple impulsive source to excite a bank of 5 second order digital resonators implemented in 16 bit scaled integer arithmetic and programmed in X86 assembly language.

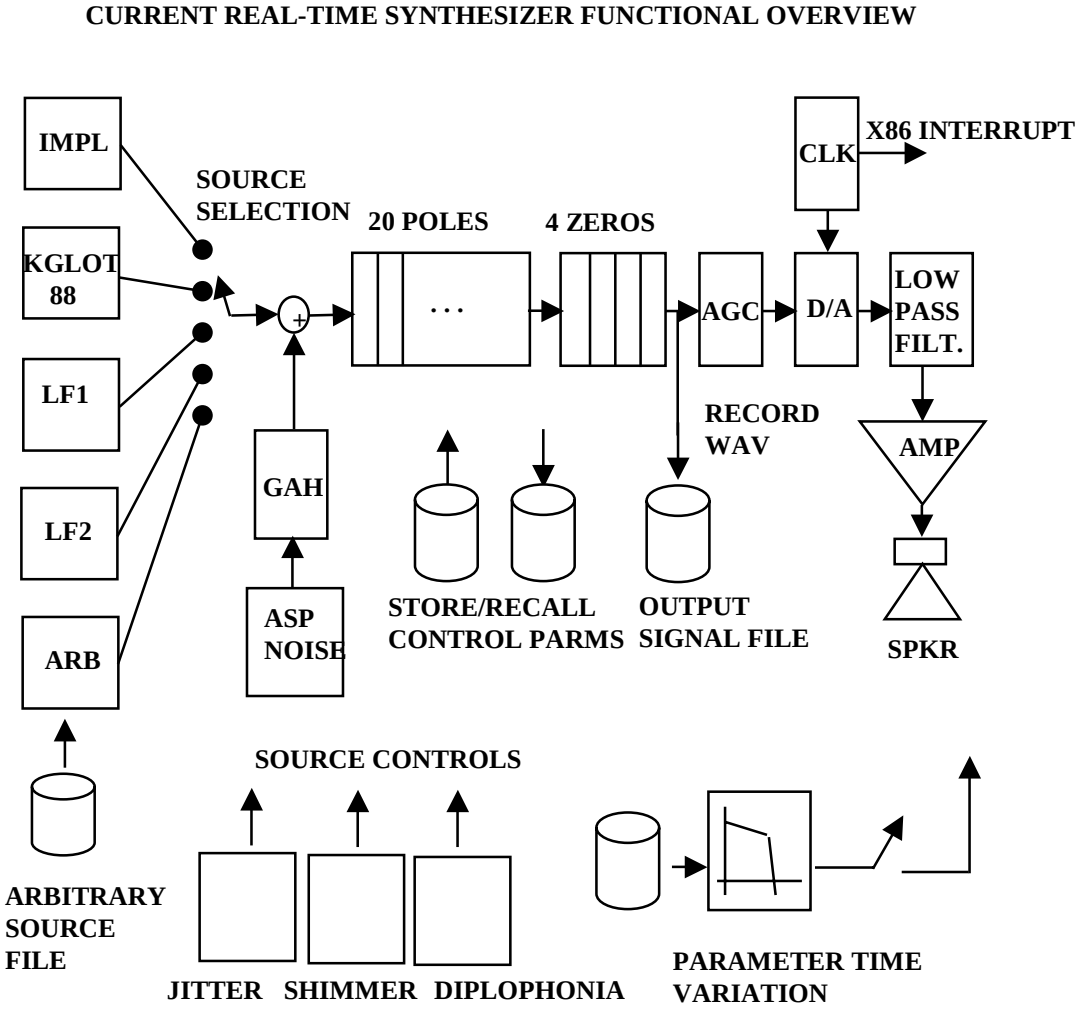


Figure 4.5. Overview of the current real-time synthesizer. Upgrades include flexible source specification (impulse, KGLOTT88, LF, or arbitrary waveform), an aspiration noise source, vocal tract transfer function numerator zeros, an automatic gain control, jitter, shimmer, diplophonia, arbitrary time variation of parameters, and the ability to store and recall time series and parameter sets. All hardware components are grouped on one adapter card.

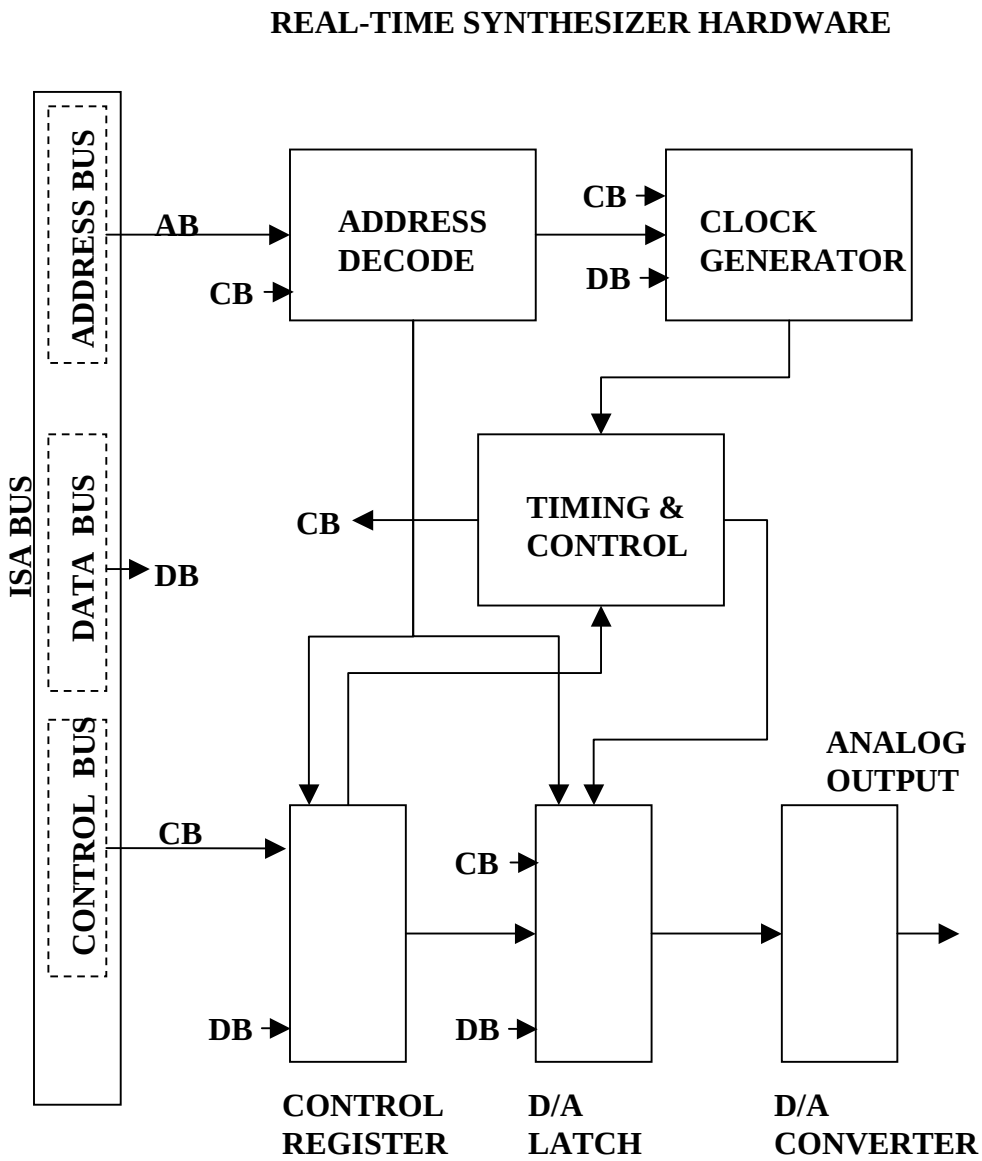


Figure 4.6. Real-time synthesizer hardware. Several hardware features were required to achieve true real-time performance on the PC platform. These include a crystal-controlled clock, a D/A converter, data latches, and timing and control glue logic to interface these components.

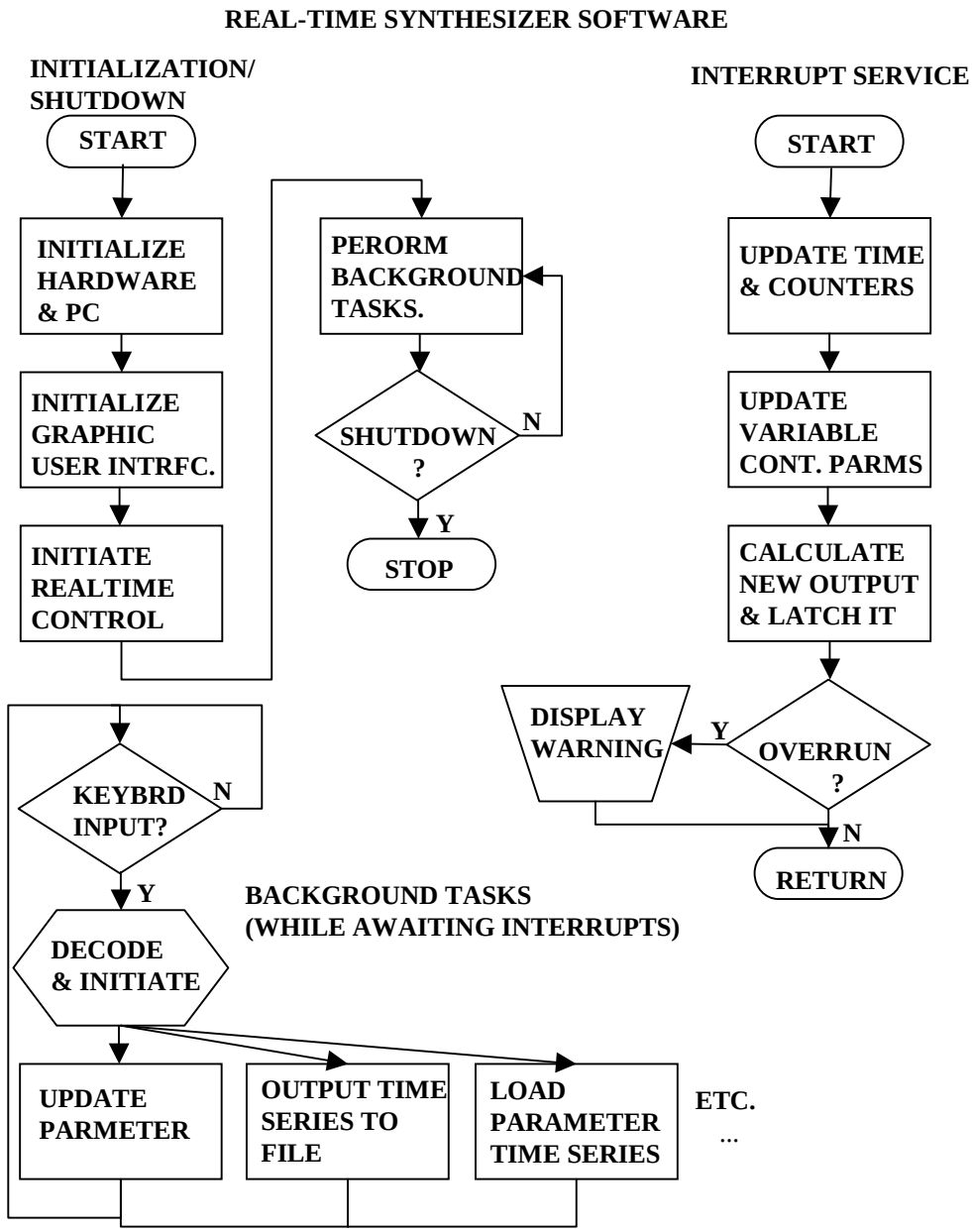


Figure 4.7. Real-time synthesizer software. Programming code may be segregated into two types: foreground (hard real-time tasks), and background (user interface, system management, etc.). The foreground tasks were performed in an ISR (interrupt service routine) and were coded in assembly language for fastest possible execution. The background tasks were coded in C language.

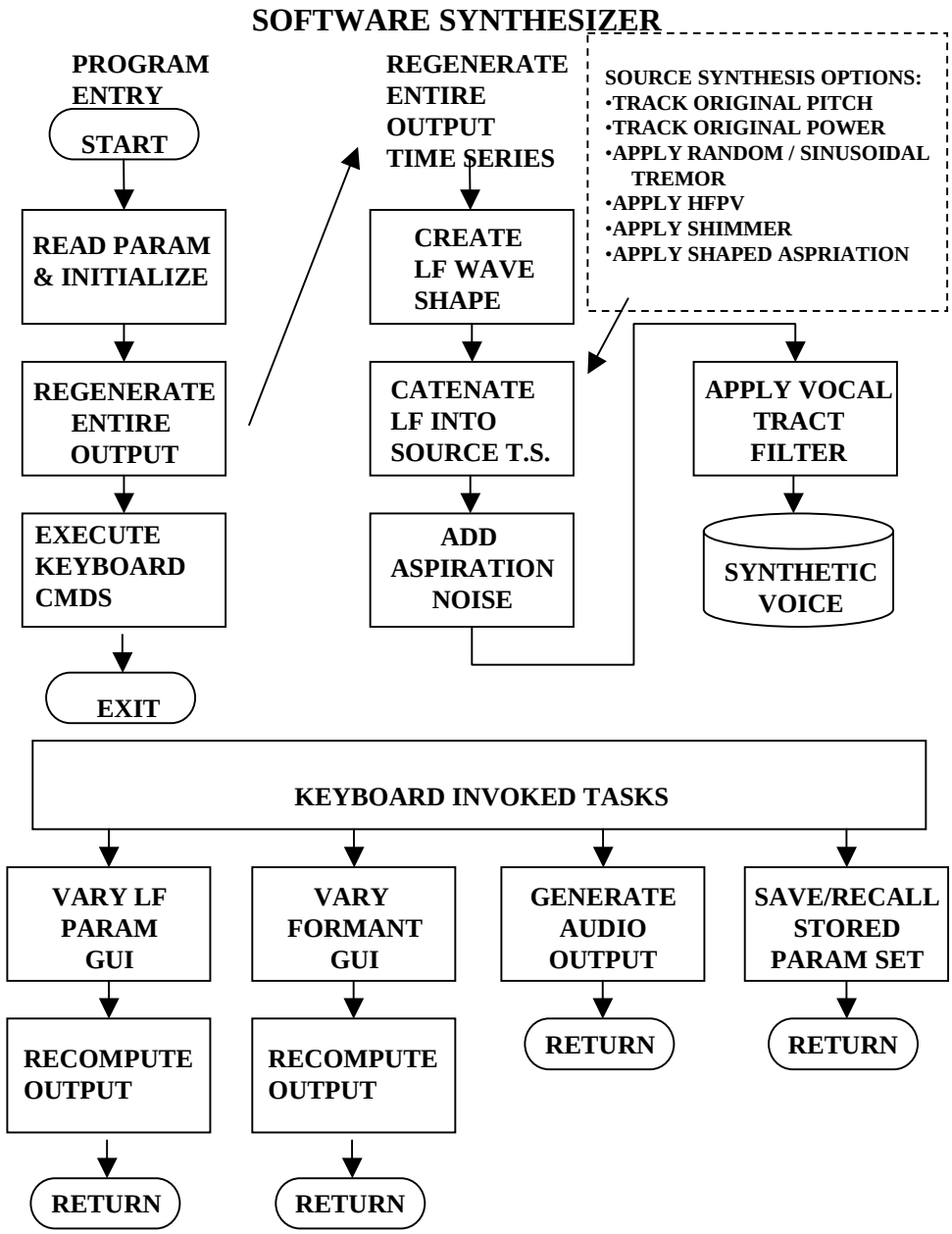


Figure 4.8. Overview of software synthesizer operation. After invocation, the synthesizer loads parameters for the requested (previously analyzed) case and calculates the synthetic version. The program then waits for user input commands to execute functions such as modifying the LF source parameters, modifying formants, or dumping or loading time series to/from disk.

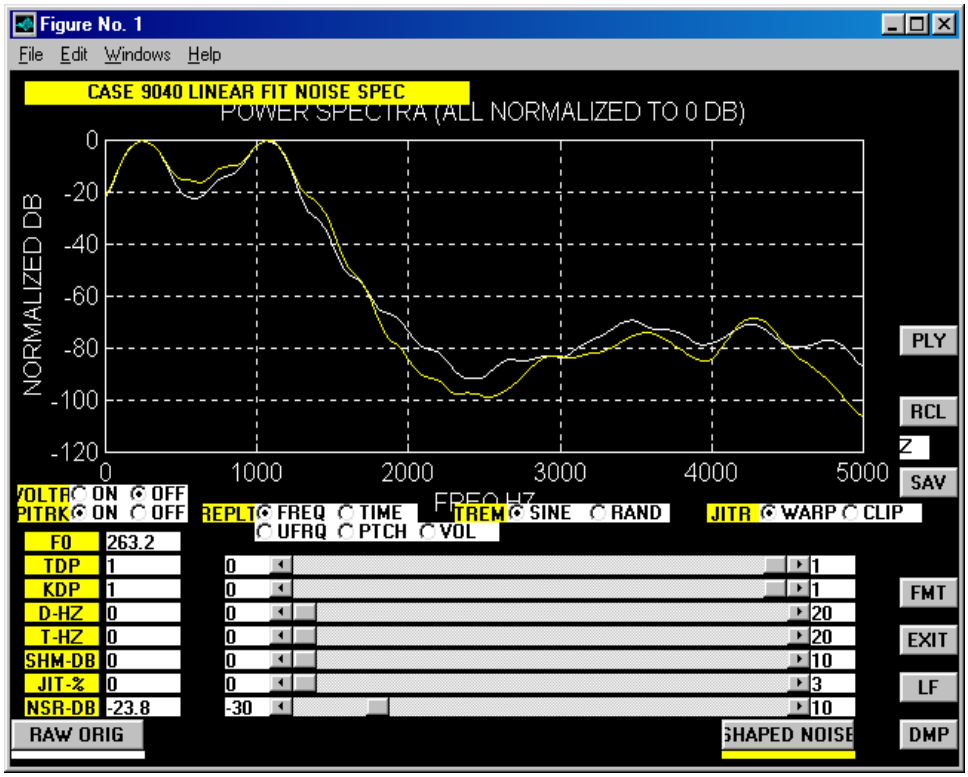


Figure 4.9a. Main GUI (graphical user interface) for the software synthesizer. Provisions are included for user specification via sliders or text of noise levels, HFPV, shimmer, etc. Options are included for tasks such as activating playback of the original or synthetic voice, turning on/off fundamental frequency/volume tracking, plotting spectra or time series, or invoking modification of LF or formants.

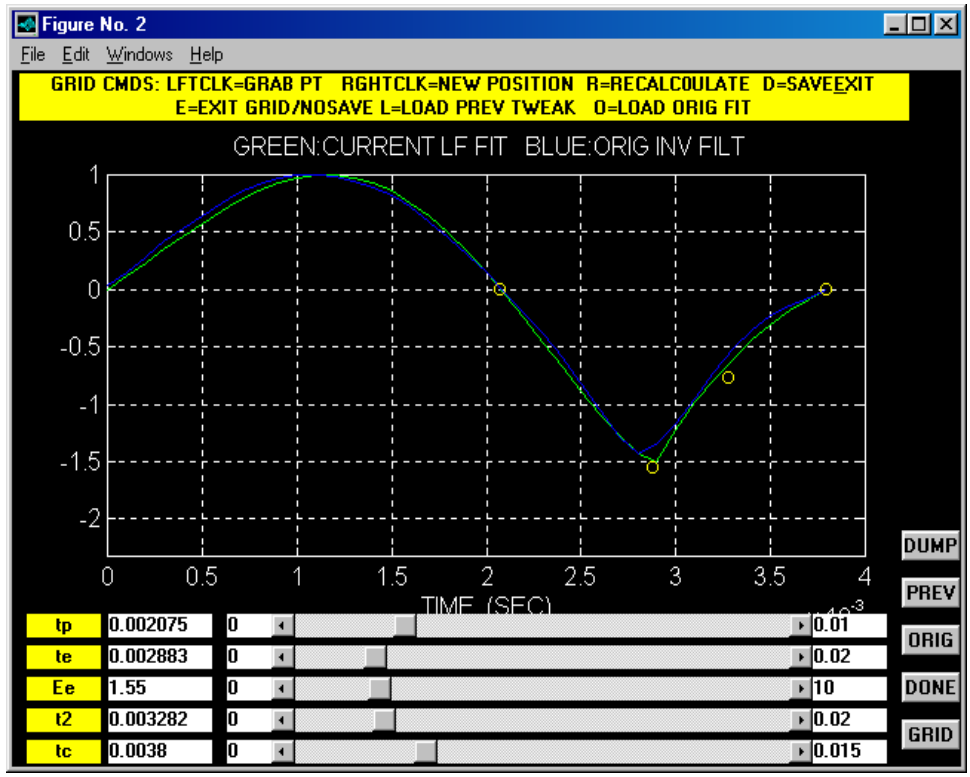


Figure 4.9b. LF modification GUI. This screen allows the user to control the shape of the LF source waveform by varying the LF parameters.

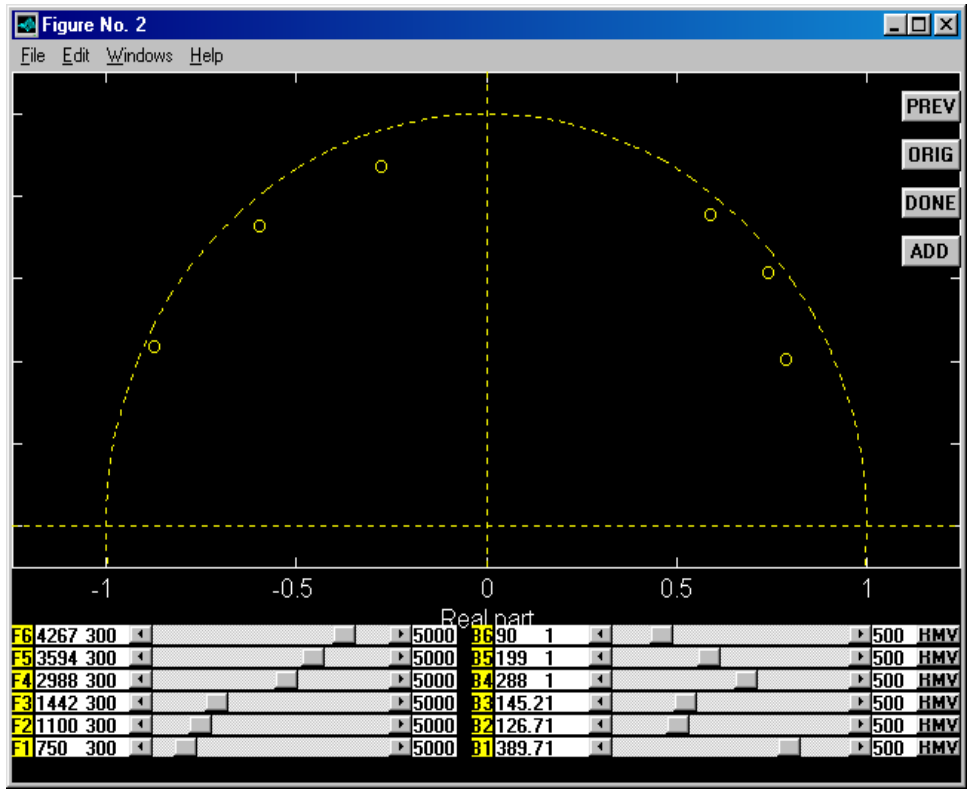


Figure 4.9c. Formant modification GUI. This screen allows the user to move the pole locations specifying the all-pole vocal tract model.

Chapter 5

Synthesis Algorithms and Validation

An essential step in the study of pathological voices is re-synthesis; clear and immediate evidence of the success and accuracy of modeling efforts is provided by comparing the original and synthetic versions of the pathological voice. The effects of variations of each of the model parameters may be quickly evaluated perceptually by generating synthetic voice samples with an easily controlled synthesizer. Tests may be performed to validate analysis results, and experiments may be performed to determine the effects on the listener of variations and interactions of model parameters. In this section, the details of algorithms used to synthesize pathological vowels are described. Experiments confirming the success of synthesis are then explained.

5.1 Synthesis Algorithms

This section describes algorithms used by the synthesizers to regenerate a synthetic version time series of the original pathological vowels. Using the derived analysis model parameters describing the pathological voices (formants, glottal source waveform, aspiration level and spectral shape, tremor, HFPV, and low and high frequency power variation), a synthetic version was calculated for each original pathological voice sample. Most of the steps of the synthesis process have direct analogs in the analysis steps described in Chapter 2. The software synthesizer implements the most current algorithms.

5.1.1 Basic Waveform Generation

The modified LF model [31], with its ease of use and adaptability to a variety of waveforms, is currently chosen as the most useful source waveform model for synthesis of pathological voices. Using the estimated LF parameters as described in Section 2.2.2, a basic waveshape of the glottal flow derivative is calculated (Fig. 2.13 and Fig. 2.15) using a parametric time scale normalized to one pulse period. The amplitude is normalized to unity, and this waveshape is used throughout the simulated voice by concatenation; the LF waveshape is assumed to remain constant in the current implementations of the synthesizers. The effects of fundamental frequency changes due to tremor and HFPV are created by variation in the sample instants chosen for interpolation of the calculated basic LF waveshape, as described in Section 5.1.2.

5.1.2 Source Synthesis – Low Frequency Fundamental Frequency Variation

In order to simulate base (low frequency) variations in fundamental frequency, the source waveshape is effectively stretched or compressed in time such that the period of one fundamental frequency pulse in actual time is exactly the reciprocal of the desired instantaneous frequency. This changes the number of actual time samples interpolated on the LF pulse waveshape. To raise fundamental frequency, fewer samples are selected from the fitted LF pulse; to lower fundamental frequency, additional samples are selected. These interpolation points are chosen equally spaced along the LF waveshape, with their spacing inversely proportional to the desired frequency. The synthesizer provides several options for selection of the base frequency:

1. A constant value, such as the average of the low-pass filtered (tremor) frequency of the original voice (for example, the average value of the top curve in Fig. 2.21).

2. A sinusoidally varying frequency about the mean F0 value. The user selects the frequency of variation, and extent of variation (deviation).

3. A randomly varying frequency about the mean F0 value generated by low pass filtering of Gaussian noise. The user selects the extent of variation (deviation) and the filter cutoff, which effectively determines the mean frequency of variation.

4. The same tremor as the original voice. The base value of fundamental frequency is obtained from interpolation on the low pass filtered fundamental frequency track (tremor)

of the original voice (for example, the top curve in Fig. 2.21). The instant of interpolation on the tremor track is selected using the time of the first sample of the currently being constructed LF pulse in the simulated time series; fundamental frequency is not varied within a single source pulse.

To calculate the specific samples for each pulse, the instantaneous frequency is used, along with the absolute finish time of the last sample of the previous pulse, to convert sample instants in real time to phase arguments specifying abscissa values on the LF waveshape. The final LF samples are then generated via linear interpolation at these abscissa values. In this manner, changes in fundamental frequency specified by the selected fundamental frequency generation method are smoothly produced, with no perceptually discernable jumps in frequency. By contrast, when fundamental frequency variation is implemented via simple truncation or addition of samples to the pulse, a quantization effect is generated, creating the impression of "steps" in fundamental frequency during linear changes in fundamental frequency.

5.1.3 Source Synthesis – High Frequency Fundamental Frequency Variation

High frequency fundamental frequency variations are simulated in the same manner as low frequency variations by effectively changing the instantaneous fundamental frequency with fundamental period modification. HFPV can be applied in the synthesizer independently of the low frequency fundamental frequency variations. As

each new fundamental frequency pulse is synthesized, the base fundamental period determined by any of the methods mentioned (Section 5.1.2) is perturbed by a random increment to lengthen or shorten it, thus modeling the measured HFPV (Sections 2.3.3-2.3.4). The random incremental change in fundamental period length is created by generating a random modification factor with Gaussian distribution, unity mean, and standard deviation determined by the desired level (usually the measured value) of HFPV. Setting synthesizer jitter to 100% implies the creation of a standard deviation in fundamental period length equal to the fundamental period. This modification factor is then applied to the base fundamental period to arrive at the final synthetic fundamental period

Setting the modification factor to get the desired level of jitter in the synthetic signal as measured by the fundamental frequency tracker and analysis software involves a complication. Unfortunately, setting the standard deviation of the modification factor exactly equal to the level implied by the desired HFPV does not produce this same level of HFPV in the resulting synthesized source time series. When the HFPV analysis is applied to the synthetic signal produced, a smaller level of HFPV is always measured. The cause of this discrepancy is illustrated in Fig. 5.1, which illustrates synthesis of two successive flow derivative waveforms. Note that although the length of each pulse is determined by a single random number, the peak to peak interval (Tpp), which is measured by the fundamental frequency tracker, is determined by the sum of fractions of two random subintervals, as shown in Fig. 5.1 and Eq 1.

$$Tpp = (1 - a)T1 + aT2 \quad [1]$$

And

$$T1 = T(1 + (PJ/100)R1),$$

$$T2 = T(1 + (PJ/100)R2),$$

Where:

Tpp = measured negative peak to peak interval,

T1,T2 = first and second fundamental periods,

PJ = percent HFPV set in synthesizer,

R1,R2 = Gaussian random numbers with zero mean and $\sigma = 1.0$,

a = fractional position of negative peak within the fundamental frequency pulse = Te/T,

T = unmodified fundamental period,

Te = time of negative peak in pulse.

The expected variance of Tpp is the sum of the variances of the two components:

$V = V1 + V2,$

where the variances are:

$V = (T\ PJf/100)^2,$

$V1 = (a\ T\ PJ/100)^2,$

$V2 = ((1-a)\ T\ PJ\ /100)^2,$

and PJf = resulting percent HFPV in Tpp. Solving for PJf as a function of PJ and peak position a yields the relationship in Eq. 2:

$$PJf = PJ\ (2a^2 - 2\ a + 1)^{0.5} \qquad [2].$$

The validity of this relation was confirmed with a Monte Carlo MATLAB simulation of fundamental pulse peak-to-peak interval measurement. The expected measured fundamental frequency period of the synthetic voice was calculated using averages of 100,000 randomly generated pulses for each of a range of a values. For each pair of simulated pulses, the predicted fundamental frequency period (as measured between adjacent minima as shown in Fig. 5.1) was calculated. This measurement was repeated 100,000 times and then averaged; the whole process was repeated for values of 0.1, 0.2, …1.0 corresponding to negative peak positions ranging from the beginning to the end of the fundamental pulse. Fig. 5.2 displays the result of the simulation. The circles show the result of the simulation, and the line is the standard deviation predicted by Equation 2. There is good agreement, which improves with more samples. Thus, a correction factor of $1/(2a^2 - 2a + 1)^{0.5}$ must be applied to the desired level of HFPV to obtain the value to use in the synthesizer simulation equations when simulating HFPV.

5.1.4 Source Synthesis – Low Frequency Power Variation

In a manner analogous to low frequency FM synthesis, provision is made for applying low frequency power modulation to the synthesized voice. The measured low frequency power variations (Section 2.4.3) of the original voice can be applied to the synthetic voice to generate the intensity variations perceived by the listener in the original voice. Signal power is proportional to the square of the signal voltage. In order to apply these variations, a gain correction time series is generated that is proportional to the square root of the low frequency power variation (upper dashed curve in Fig. 2.27). The gain correction is then applied to the synthesized signal to achieve a power variation approximating the original voice.

5.1.5 Source Synthesis – High Frequency Power Variation

Similar to the HFPV synthesis, high frequency power variations (shimmer) are available in the synthesizer. Shimmer is synthesized in a manner analogous to the way it is measured, as a perturbation of pulse power with a Gaussian distribution. To synthesize pulses with randomly varying power, a Gaussian random gain is generated and applied to the samples of each fundamental pulse (the same gain value is used over all the samples within a pulse). The applied gain has unity mean and standard deviation determined by the amount of desired shimmer.

As with HFPV, there are many methods of measuring shimmer [3]. Assuming shimmer is a small perturbation of fundamental period length with a Gaussian distribution, linearity allows conversion between several types of measures, including gain, power, and dB. The percentage power variation measured in the analysis of the original voice (Section 2.4) can be converted to shimmer in dB (used as input in the synthesizer) and a gain value for fundamental frequency pulses (used in the synthesis equations). The nonlinear relations between these quantities are linearized about the mean value of shimmer to yield simplified formulae. In general, probability distributions of a nonlinear function of a variable with Gaussian distribution are themselves not Gaussian. Small perturbations in the conversion equations used here, however, are Gaussian as a reasonable approximation, allowing the use of standard deviation as a measure of shimmer. Therefore, the quadratic relation between power and gain simplifies to the approximation:

$$PPS = 2*GPS$$

Where

GPS = percent gain variation (linear)

PPS = percent shimmer in power

= 100*standard deviation in power/mean power

The logarithmic relation between power and dB simplifies to the approximation:

$$PPS = 10*\ln(10)*DBS = 23.0*DBS$$

Where

DBS = shimmer in dB

= standard deviation of signal dB measure

5.1.6 Aspiration Noise Implementation

The final step in source synthesis is the addition of spectrally shaped Gaussian noise to simulate aspiration at the glottis. The current model assumes high frequency (>10 Hz) nonperiodic signal content other than HFPV and shimmer is modeled by aspiration noise. This assumption appears to be approximately true for a subset of pathological voices in which an excellent synthetic match to the original is obtained with aspiration noise. The Gaussian statistical distribution and the spectral shape of source aspiration noise are preset in the synthesizer to the measured values of the corresponding original voice. The energy level of aspiration noise relative to the periodic signal level can be fine-tuned by the user via the adjustments available in the synthesizer.

5.1.6.1 Source Noise Spectral Shaping

White noise with Gaussian distribution and unity variance is first generated. A 100-tap FIR filter is synthesized to match the spectral shape of the original source (25 point piecewise linear approximation determined from analysis); the noise is passed through the filter to match the original noise source shape.

5.1.6.2 Source Noise Energy Level

In order to complete the calculation for inclusion of aspiration noise, the relative gain of the aspiration noise signal relative to the glottal source signal must be found. The preset or user adjusted aspiration noise level in dB is used to find the correct gain value. It is calculated using the relative energies of the glottal source and aspiration noise time series before they are summed to obtain the final synthetic source time series. The nominal value of aspiration noise to apply in order to achieve the best match to the original voice is determined via the cepstral filtering method described in Section 2.5.2.

5.1.7 Vocal Tract Model

The final step in voice synthesis is applying the vocal tract filter to the glottal flow derivative time series, which at this point includes the adjusted LF waveform and the selected levels of nonperiodic features, such as AM, FM, and aspiration noise. Currently, the synthesizer uses fixed formants for the entire time series. The formants determined in the analysis (Section 2.1) are converted to all-pole resonator filters, and applied to the source time series to generate the final synthetic time series. The synthesizer automatically normalizes the amplitude of the maximum excursion of the final time series signal to the full range of the D/A used for sound generation, thus minimizing quantization effects while preventing clipping.

5.2 Synthesis Validation

With skillful adjustment of synthesizer parameters (including aspiration noise, HFPV, and shimmer) it is possible to achieve synthetic samples that are very close to the original; in some cases, synthetic voices are indistinguishable from the original. Since one of the initial motivations for this project was creation of synthetic vowels as perceptually close to the original as possible, considerable effort was made to objectively and perceptually compare the resulting synthetic vowels with the originals after which they were modeled. In this section, the success of several aspects of analysis/synthesis is evaluated with tests addressing the nonperiodic model parameters. In order to objectively evaluate the accuracy and consistency of the overall analysis/synthesis process, the processing loop is closed by re-analyzing the synthetic voices with the same software used to analyze the original pathological voices. The levels of nonperiodic components in the synthetic versions are then checked to guarantee values consistent with original values.

5.2.1 Aspiration Noise (AN) Verification

In the absence of AM and FM modulations, the cepstral NSR measurement of the synthetic voice should reflect the value of shaped source noise set in the synthesizer when the voice was created, since any nonperiodic energy should be entirely due to this aspiration noise. For each of the 31 voices, synthetic versions were created with the levels of AM and FM modulation set to zero, and the level of aspiration noise set to that measured in the original voice. Using the same noise analysis procedure used on the original voice, the synthetic NSR was measured. The result is shown in Fig. 5.3, in which

the measured synthetic NSR is plotted against the measured original NSR for all 31 cases. The original voices span a measured NSR range of about –25 dB to –5dB. Over this range, the agreement between natural and synthetic NSR is within about 1 dB, which is well within perceptible limits, as approximately determined by varying this parameter on the synthesizer and comparing the resulting vowels. Thus, the process of measurement and synthesis of aspiration noise appears consistent.

5.2.2 HFPV Verification

In a manner similar to the NSR verification, HFPV in the synthetic voice was checked against the value set in the synthesizer (which was the measured value in the original voice). The measured values of HFPV in the synthetic voices achieved agreement with that of the original voice to within 0.1%, which is well within perceptible limits. Thus, the process of measurement and synthesis of HFPV appears consistent.

5.2.3 Effect of AN on HFPV

Another relevant question is the degree of interaction between aspiration noise and HFPV. The addition of aspiration noise to the source time series would be expected to affect the measurement of HFPV due to perturbation of the position of time domain features (eg. peaks) detected by the fundamental frequency tracker. The relevant question is how significant is the effect for the levels of aspiration noise and HFPV measured in the set of original pathological voices. To asses the increment in measured HFPV due to

the inclusion of aspiration noise in the synthetic voices, a set of 31 voices was synthesized with the original levels of HFPV (Sections 2.3.3 and 5.1.3) plus the level of aspiration noise set to the NSR level measured in the original voice before any demodulation (this represents the worst case of additive noise). The FM analysis was then carried out on these synthetic voices with both aspiration noise and HFPV. The result is shown in Fig. 5.4, which plots measured HFPV in the synthetic voices with aspiration noise versus the level of HFPV in the synthetic voices without aspiration noise (Sections 2.5 and 5.1.6). As can be seen, there is an increment in HFPV of about 0.2%, which was near the limit of perception.

5.2.4 Effect of HFPV on AN

Similarly, the effect of HFPV on measured aspiration noise is addressed. The increment in measured NSR due to the addition of HFPV at the level measured in the original voice was evaluated. Starting with synthetic voices with aspiration noise only (Section 5.1.6), HFPV was added and the resulting NSR measured. The result is displayed in Fig. 5.5, which plots the cepstral NSR of synthetic voices with HFPV versus those without. The result appears to be about a 4 dB increment in NSR, which seems consistent with the result of Fig. 2.32.

5.2.5 SABS for Aspiration Noise

Pilot perceptual experiments were conducted comparing original voice samples with synthetic vowels. The effect of FM demodulation on the accuracy of NSR measurement was demonstrated. Listeners (who were demonstrated the effects of NSR parameter variation) attempted to match synthetic samples to the original ones by varying the synthetic aspiration noise level. The synthetic HFPV was turned off for this test. The results are displayed in Figs. 5.6, 5.7, and 5.8 which plot the mean level of aspiration noise listeners chose to match the perceptual effect of the original samples versus the original measured cepstral NSR. Fig. 5.6 displays the result for the original voice. Fig. 5.7 displays the result for the cepstral NSR measurement on the voices with tremor removed. Fig. 5.8 displays the result for the voices with both AM and all FM removed. There is a good indication of correlation with the original voice (Pearson = 0.51). However, the correlation increases when tremor is removed (Pearson = .71), and then increases again when all AM and FM is removed (Pearson = 0.87). In addition, the best-fit line moves from as much as 10 dB off (from perfect correlation) in the case of the original voice, to within 2 dB in the case with all AM and FM removed. Thus, the major disagreement between cepstral measured NSR and listener-set aspiration level is accounted for by FM modulation

5.2.6 SABS for HFPV

In a same manner as with aspiration noise, SABS pilot tests were conducted to vary HFPV. With the level of aspiration noise (which proved to be more perceptually distinguishable than HFPV for the 31 voices) first set for best match to the original,

listeners adjusted the level of HFPV to improve the match to the original. In most cases, it proved more difficult to set HFPV when compared to aspiration noise. The results are displayed in Fig. 5.9, which plots the mean of HFPV set on the synthesizer to match the original sample versus measured HFPV in the original voice. The level of correlation (Pearson coefficient = 0.403) is lower than that of aspiration noise.

5.3 Summary

This Chapter described the efforts for re-synthesis of pathological vowels. The algorithms for implementing synthesis of model parameters derived in analysis defined in Chapter 2 (LF source parameters, formants, aspiration noise, etc.) have been described. Validity of the overall analysis/synthesis process was tested by closing the loop with re-analysis of synthesizer outputs and with listener comparisons of original and synthetic vowels. Key findings include the fact that AM and FM demodulation improves the agreement between measured levels of aspiration noise and levels set by listeners in SABS (subjective analysis by synthesis) tests. The effect of AM demodulation was much less than FM demodulation. Tests showed less correlation between measured and listener-set HFPV levels in SABS tests than was observed for aspiration noise.

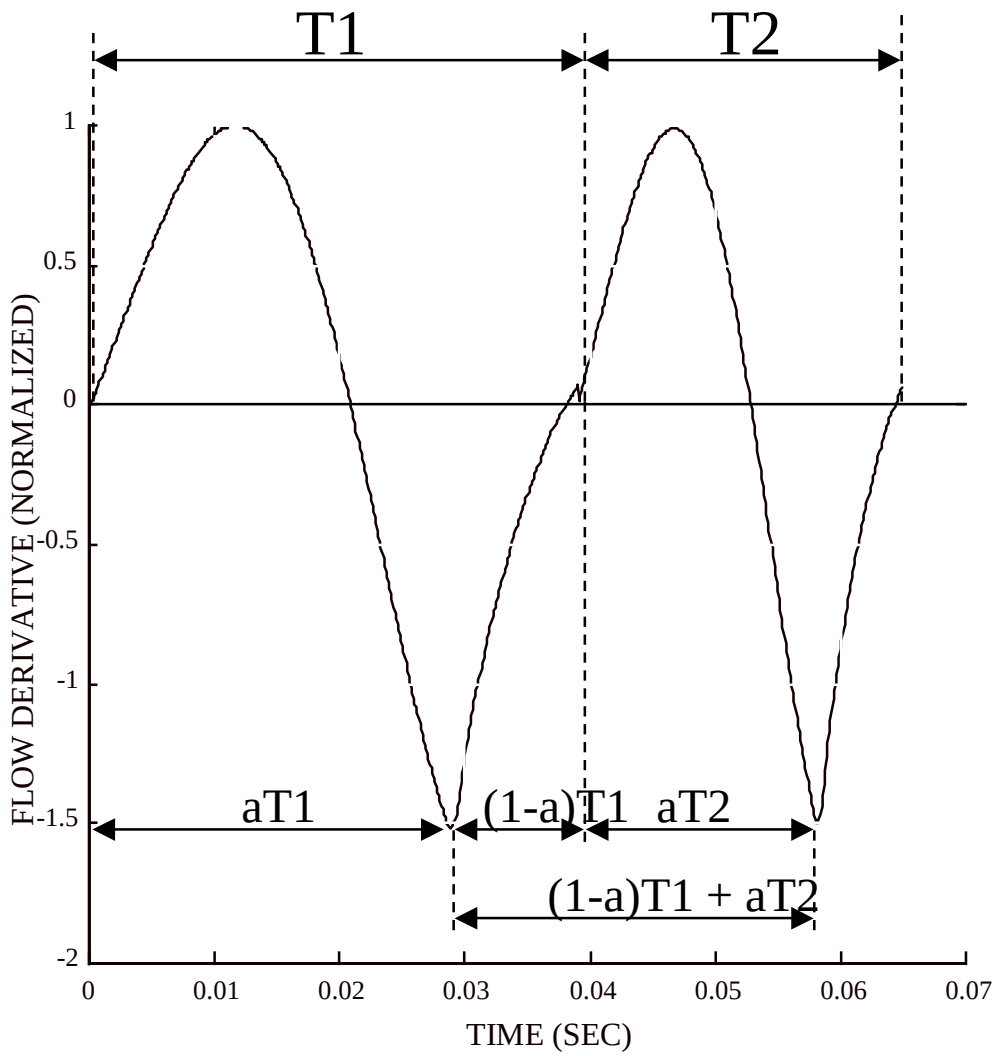


Figure 5.1. Measured synthetic fundamental period combines portions from two separately synthesized pulses; time between minima is measured: the last part of pulse T1, with length (1-a)T1, is combined with the first part of pulse T2, with length aT2. The resulting sum of two random variables is subject to basic rules of error propagation.

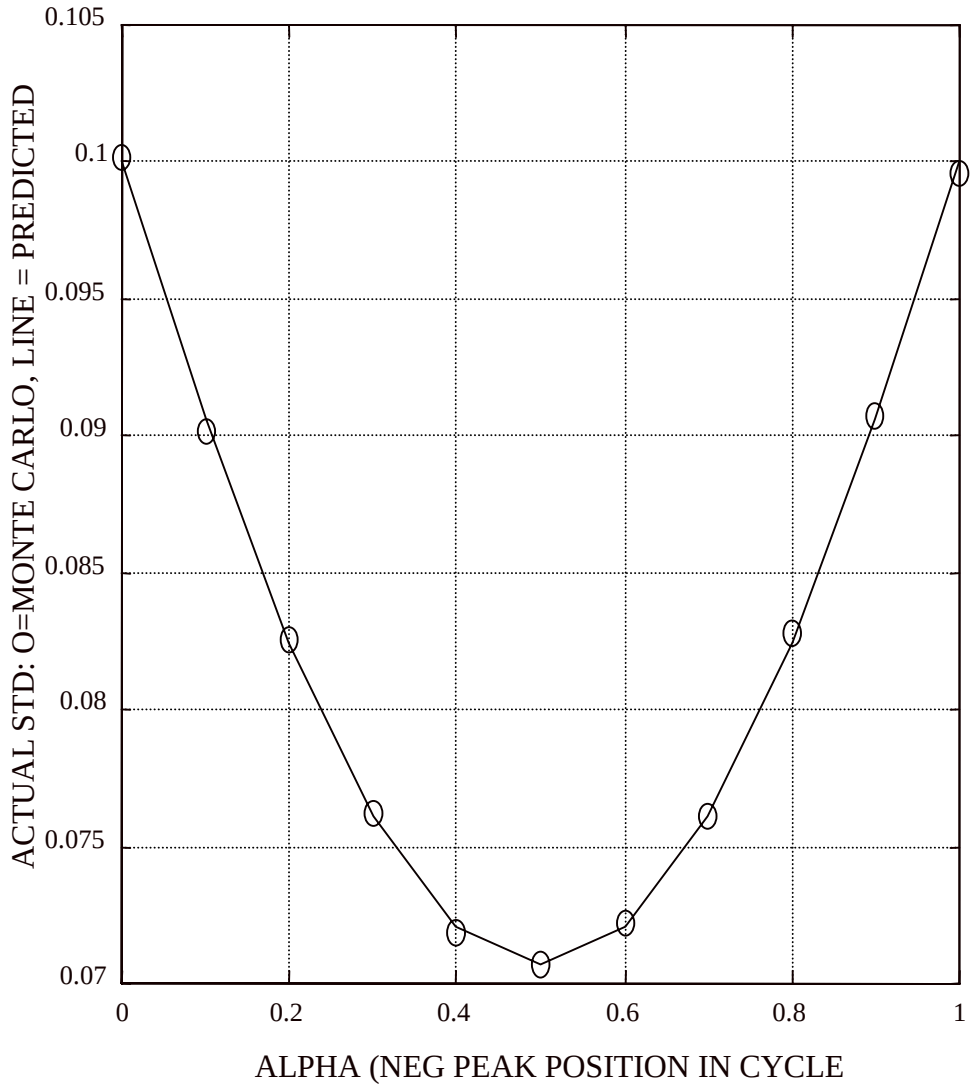


Figure 5.2 . Simulation demonstrating the effect of fundamental peak position on standard deviation of measured fundamental period length. Circles show average simulated result using 100,000 random pulses. Line shows predicted values.

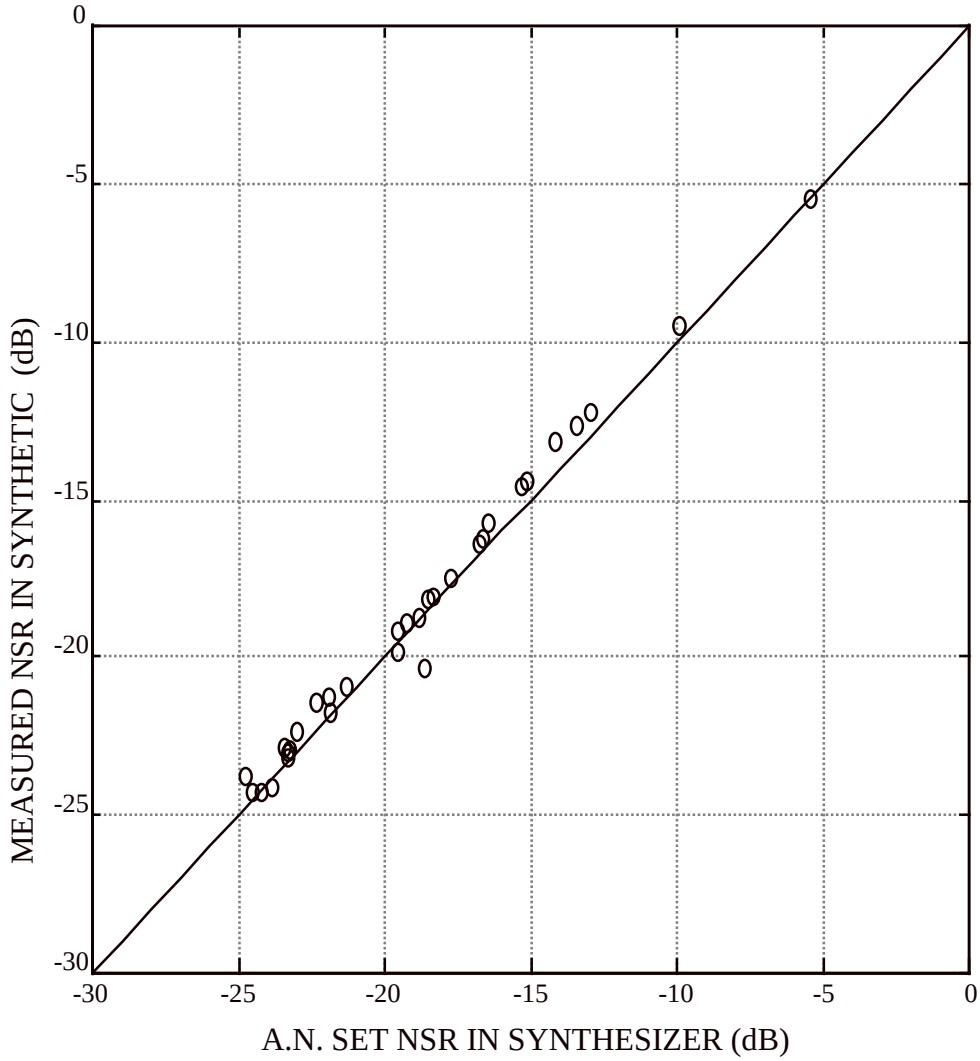


Figure 5.3. Verification of NSR level in synthesized voice. Aspiration noise (AN) levels programmed predict measured levels within about 1 dB.

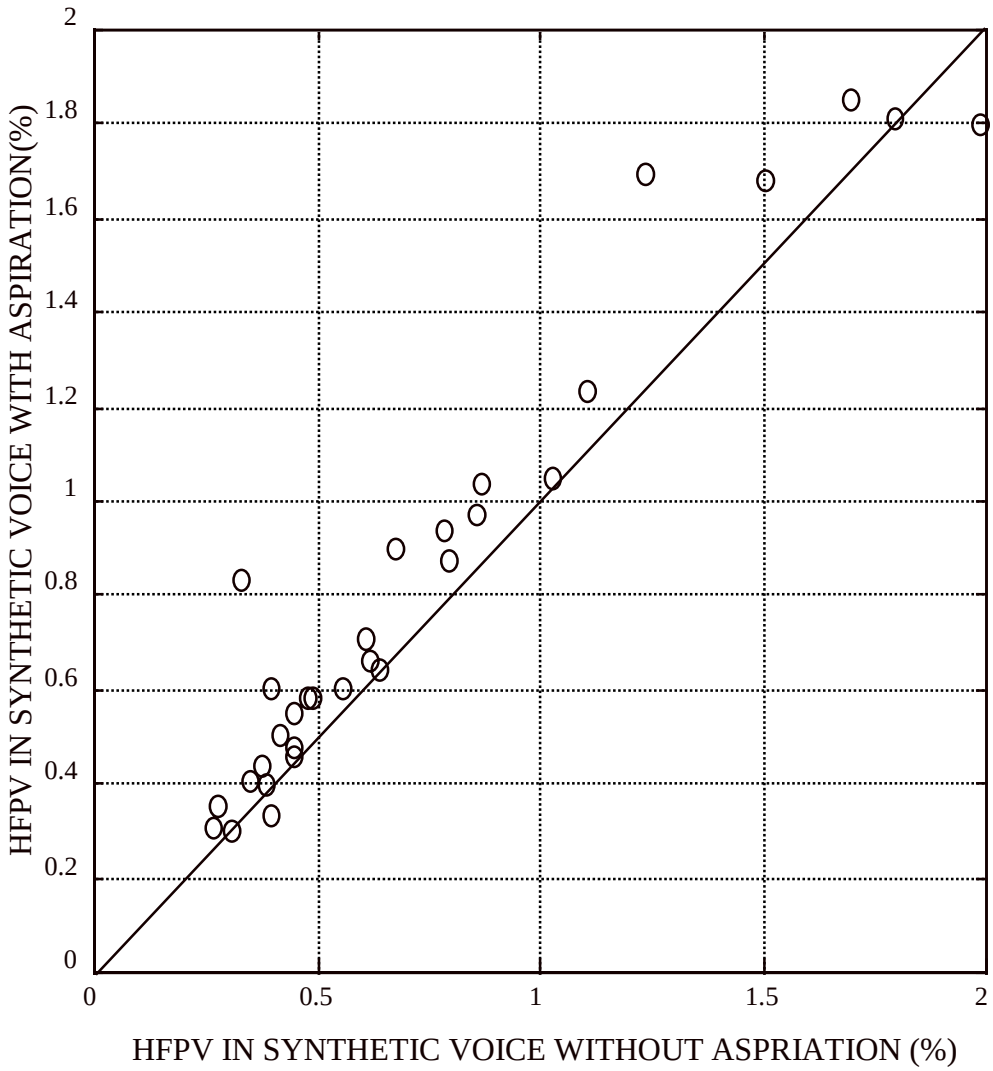


Figure 5.4. Effect of AN on measured HFPV in the synthetic voice. Horizontal axis is programmed HFPV level. Vertical axis is HFPV measured in synthetic voice when the original measured level of AN is added. AN seems to add about 0.2% to the HFPV measurement for the levels found in the voice set.

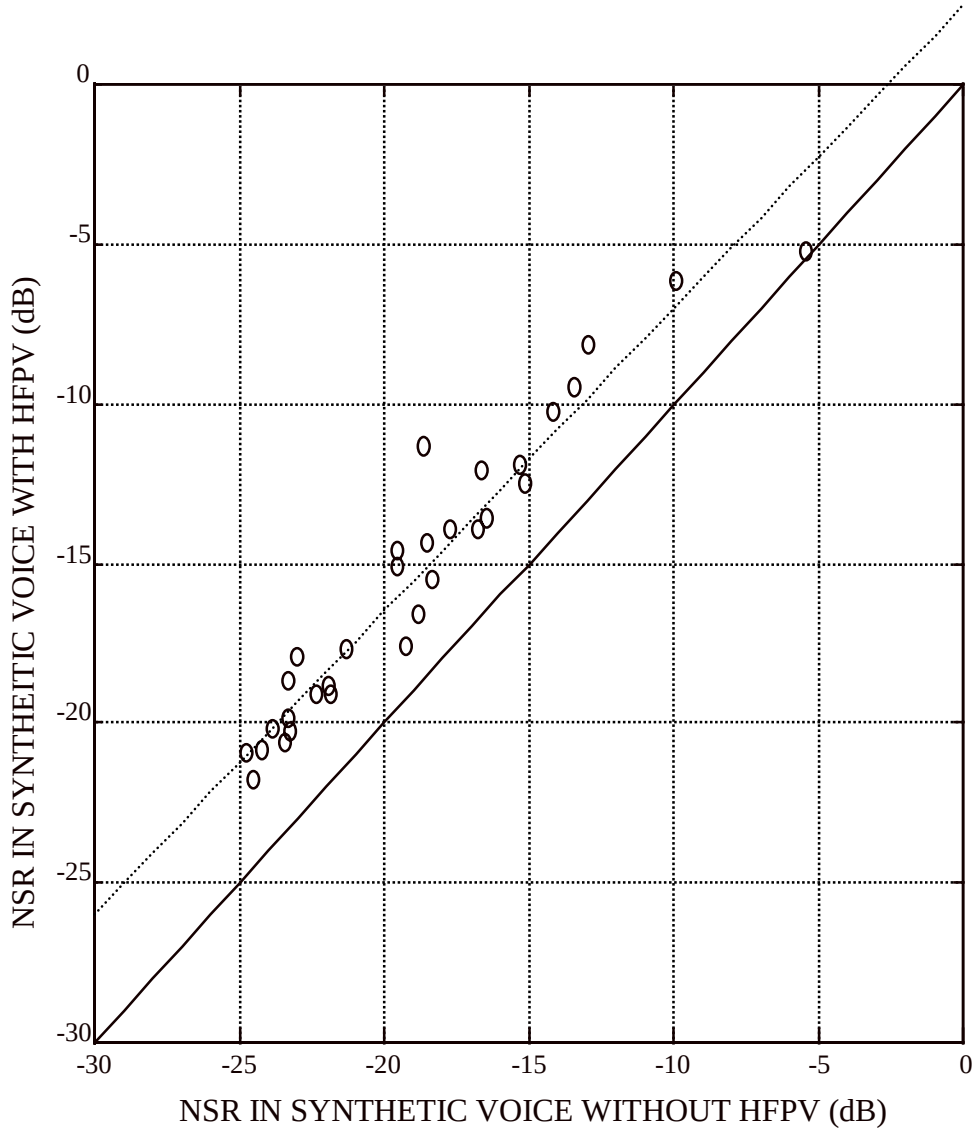


Figure 5.5 Effect of HFPV on measured NSR in the synthetic voice. Horizontal axis is programmed AN level. Vertical axis is measured NSR in the synthetic voice when the original measured level of HFPV is added. HFPV seems to add about a 4 dB increment in NSR for the levels found in the voice set. Dashed line shows best linear fit.

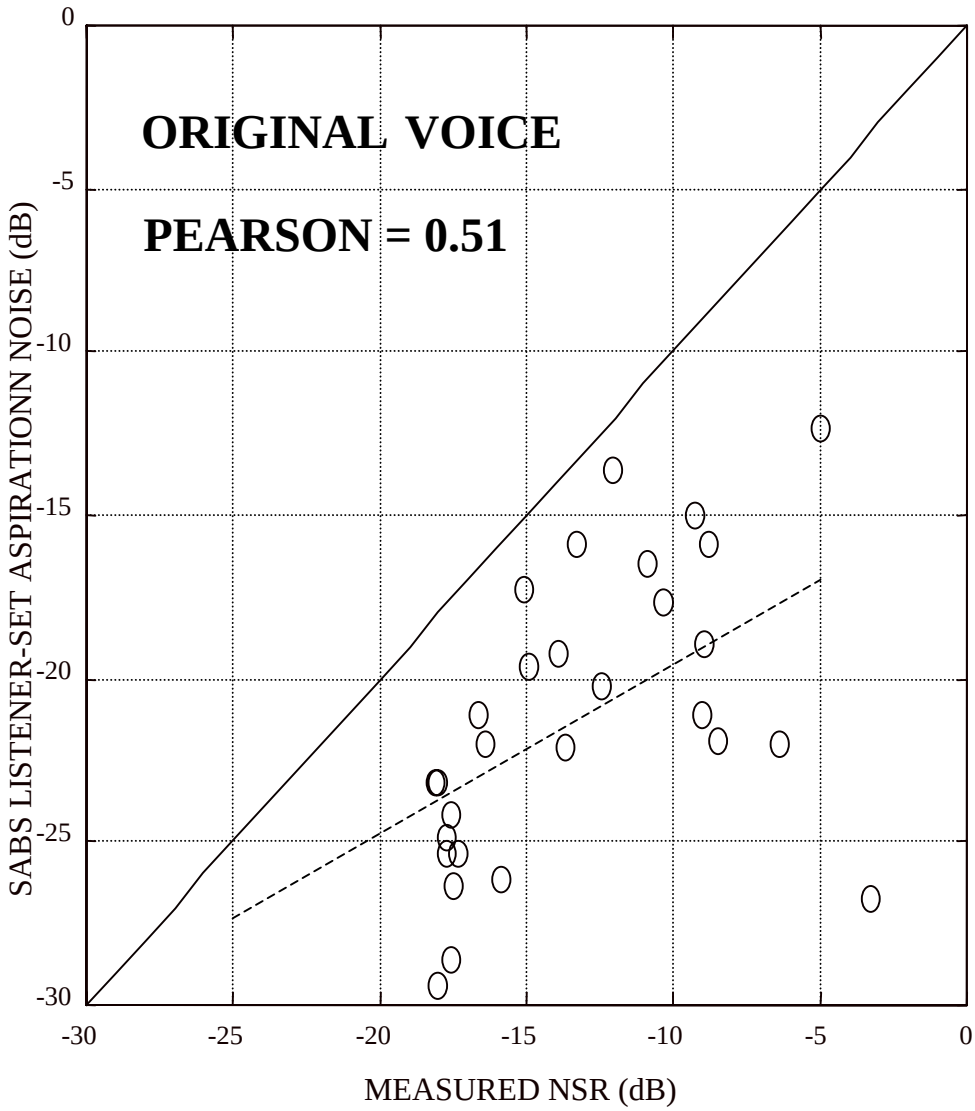


Figure 5.6. Mean of listener-set aspiration noise in SABS experiments versus the measured original NSR. The listeners set AN levels 5 - 10 dB below measured, indicating original NSR includes more than AN effects. Dashed line shows best linear fit.

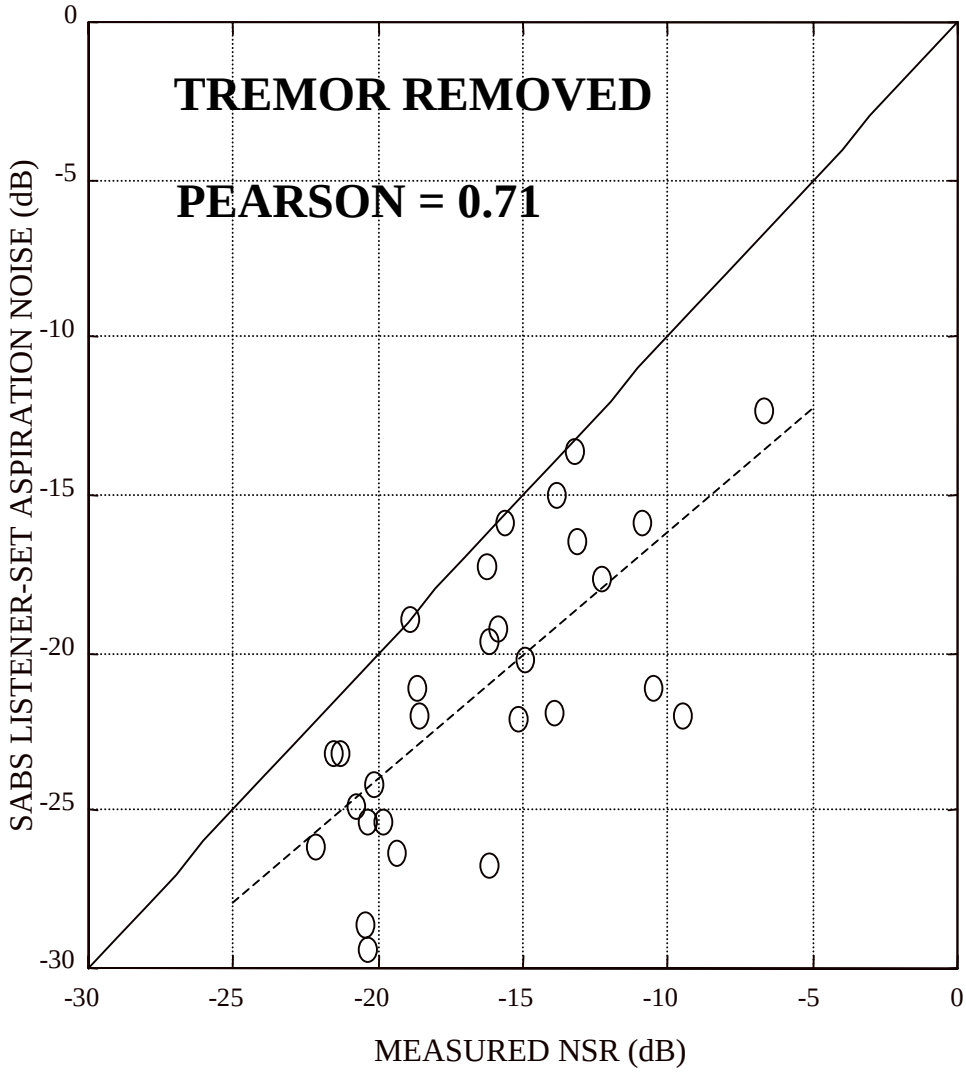


Figure 5.7. Mean of listener-set aspiration noise in SABS experiments versus NSR of voices with tremor removed. Agreement between SABS and measured NSR improves over Fig. 5.6. Dashed line shows best linear fit.

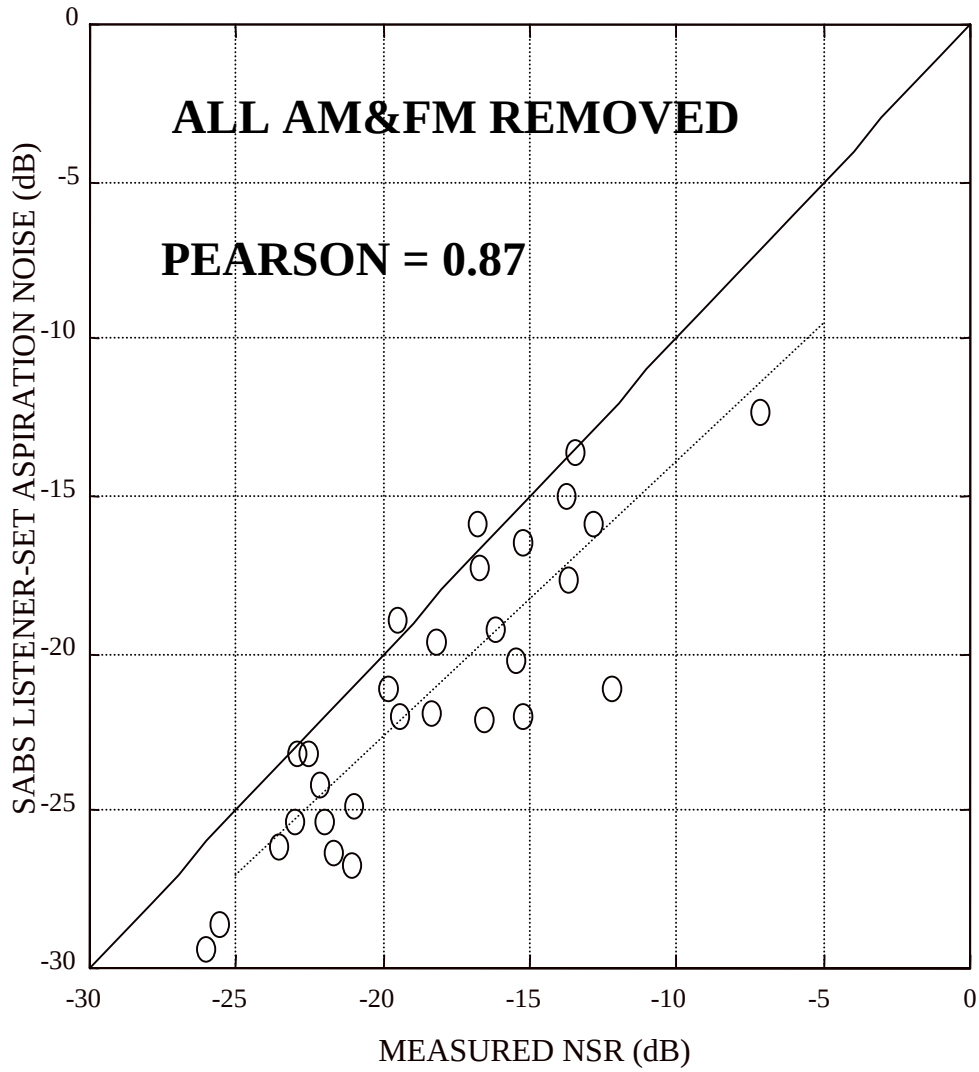


Figure 5.8. Mean of listener-set aspiration noise in SABS experiments versus NSR of voices with all AM and FM removed. Agreement between SABS and measured NSR improves further over Fig. 5.6. Dashed line shows best linear fit.

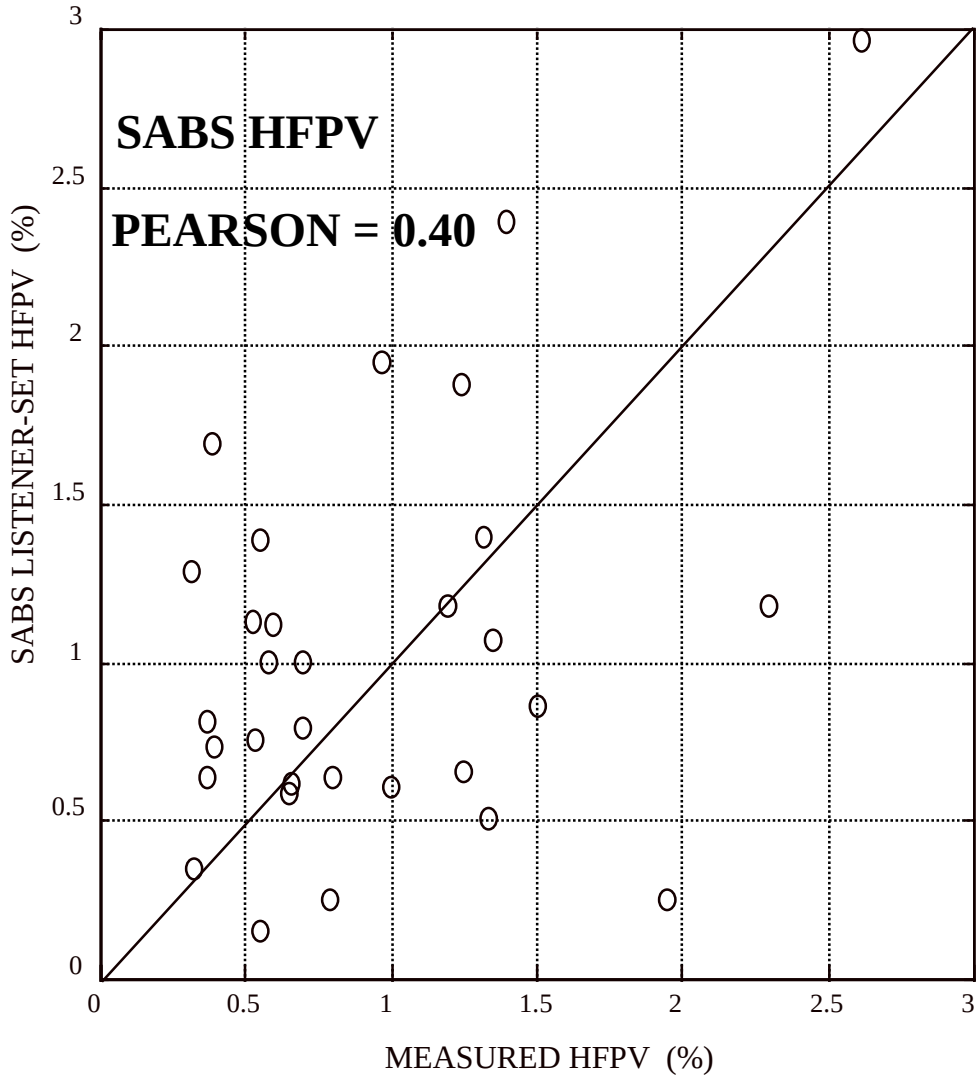


Figure 5.9. SABS comparison of user set HFPV versus measured HFPV. Weak correlation indicates listener difficulty perceiving HFPV at the levels present in the voice set.

Chapter 6

Summary and Conclusion

This dissertation has resulted in an advanced model for the analysis and synthesis of pathological vowels. The success of the process is demonstrated in the fact that many of the synthetic vowels are indistinguishable for the original vowels from which they were modeled. The overall development proceeded as illustrated in Fig. 6.1, and is now summarized.

In regard to voice analysis as described in Chapter 2, the steps of the analysis phase of voice processing have been detailed. Using the source-filter model of voice production as a basis, voices have been parameterized into formants, LF source waveform, fundamental frequency time history, amplitude time history, FM nonperiodic effects, AM nonperiodic effects, and aspiration noise. The limitations and practical aspects of LP analysis for inverse filtering and the determination of formants have been described as well as source waveform fitting using the LF model. The nonperiodic features of pathological voices have been expressed in terms of AM and FM variations

and aspiration noise. Gaussian distributions have been shown to model the small, high frequency effects in AM and FM variations. Aspiration noise was found to be well modeled with spectrally shaped Gaussian noise. The results of analysis form a set of parameters that may be used both for comparison of voices and for the generation of synthetic versions for perceptual testing.

In regards to external stimulation (ES) of the vocal tract, a major limitation of formant analysis and inverse filtering pathological voices is the reliance upon the (possibly spectrally deficient) source to reveal formant information. Segregation of the source waveform from the vocal tract using LP and inverse filtering may be ambiguous in the case of pathological voices, resulting in difficulty in achieving good vowel quality in synthesis. Without a good glottal source energy representation across the entire spectrum of interest (which is supplied mostly by the sharp return phase of a normal voice), resonances in pathological voices may not be detected with LP and other techniques. By externally stimulating the vocal tract with a spectrally rich source (chirp, impulse, white noise, etc), all resonances are clearly detected. Information obtained from ES analysis may then be combined with other approaches to yield more accurate formants and inverse filtered waveforms. The application of ES testing was successfully demonstrated via a progressive series of experiments described in Chapter 3. Formants of a known physical form (tube) were detected at the expected frequencies. Application of ES testing to the vocal tract revealed high-resolution detection of the expected formants. In addition, comparison of ES test results with traditional LP and FFT analysis of the same voices

reveals ES testing produces higher resolution, more detail, and additional resonances not detected at all by LP and FFT. Experiments with simulated pathological (breathy) voices demonstrate a particular case in which ES testing improves formant estimation. Problems of articulator movement during ES testing may be solved by any of several technical approaches. Thus, the value of ES testing for pathological voice analysis is illustrated.

In regard to voice synthesis, synthesis provides a valuable tool in the study of pathological voices. Synthetic versions of pathological vowels measure levels and changes in pathological qualities in terms of their model parameters of synthesis, such as aspiration noise level, tremor, shimmer, and HFPV. Synthetic vowels may also be used to test perceptual significance of model parameters. Chapters 4 and 5 have described the efforts for re-synthesis of pathological vowels. The function and implementation of two synthesizers has been explained in Chapter 4. A hardware-based real-time synthesizer was constructed with the capability of generating immediate responses to changes in voice model parameters. The real-time synthesizer was designed with an expansible architecture allowing easy addition new features and model parameters, deterministic real-time performance, and the capability to perform real-time closed loop control by performing all system computations within a single sample period. A software synthesizer based on MATLAB was implemented to provide the capability to quickly code and evaluate complex algorithms that would be more time-consuming to code in real-time. The algorithms for implementing synthesis model parameters derived in analysis defined in Chapter 2 (LF source parameters, formants, aspiration noise, etc.) have been described

in Chapter 5. Lastly, the validity of the overall analysis/synthesis process was tested by closing the loop with re-analysis of synthesizer outputs and with listener comparisons of original and synthetic vowels, as described in Chapter 5.

Contributions of this dissertation include:

1. Improved analysis techniques, especially those that measure NSR.

2. Improved AM and FM demodulation techniques which are effective for pathological voices.

3. Demonstration of the applicability of external source vocal tract transfer function testing for pathological voices.

4. Implementation of real-time and software-based voice synthesizers optimized for pathological vowels.

A key finding is that modeling of nonperiodic pathological voices with AM, FM, and aspiration noise does improve the accuracy of analysis and fidelity of synthesis, which was the original experimental question posed.

6.1 Future Work

The current approach to analysis/synthesis yields voices approach the originals in many cases, but some limitations were observed, which suggest possible future areas of work:

1. The current synthesizer uses a constant value for the level of aspiration noise (relative to voiced energy). Some pathological voices are particularly unsteady (even though the subjects were asked to maintain constant volume), and they exhibit considerable shift in the ratio of unvoiced and voiced energy during the one-second sample. Making the aspiration noise level a function of time will improve the synthetic versions of these voices.

2. The current synthesizer uses constants for the formants for the entire sample. FM demodulation should remove all apparent fundamental frequency variation. However, when listening to the FM demodulated natural voice time series, the impression of fundamental frequency variation can still be perceived in some cases despite the fact that reanalysis of this demodulated natural voice shows constant fundamental frequency. A possible explanation of this effect could be formant modulation. Inclusion of time varying formants in analysis and synthesis should eliminate this effect and improve the fidelity of synthesis. Efforts at implementation are currently under way.

3. Careful comparison of some of the original and synthetic voice samples in some cases reveals short period differences in quality difficult to assign to any parameter. The current model uses constant LF parameters for the entire segment. Analogously to formants, time varying LF parameters may improve fidelity and may also explain remaining small differences between measured and perceived aspiration noise levels.

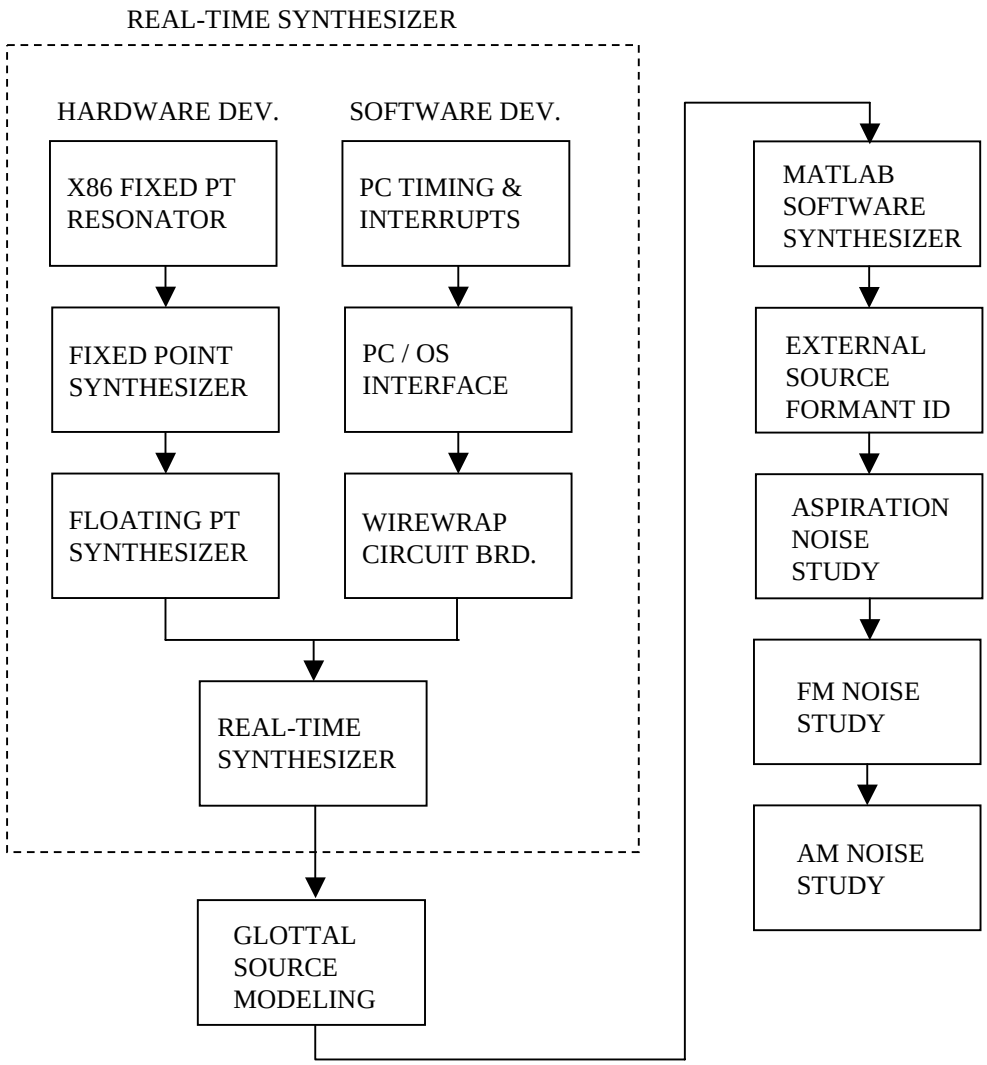


Figure 6.1 Overview of research development. Initial work resulted in a real-time synthesizer, which employed X86 native signal processing capability on a PC platform and a custom made wirewrap adapter card with timing, interrupt, and D/A circuitry. Work next focused on improvement of the synthesizer glottal source waveform. A MATLAB software synthesizer was then created to more quickly evaluate new synthesizer algorithms. Problems with modeling vowel quality then motivated research into external source identification of the vocal tract. Lastly, nonperiodic properties were analyzed and modeled as aspiration noise, FM (frequency modulation), and AM (amplitude modulation).

References

[1] Antonanzas, N. The inverse filter program developed by Norma Antonanzas can be investigated online at the following web site: www.surgery.medsch.ucla.edu/glottalaffairs/software_of_the_boga.htm

[2] Bangayan, P., Long, C., Alwan, A., Kreiman, J., and Gerratt, B. “Analysis by Synthesis of Pathological Voices Using the Klatt Synthesiser.” Speech Communication 22, pp. 343-368, 1997.

[3] Buder, E. H. “Acoustic Analysis of Voice Quality: A Tabulation of Algorithms 1902-1990.” Included as Chapter 9 in *Voice Quality Measurement*, Kent, R. D. and Ball, M. J., Singular Publishing Group, 2000.

[4] Chasaide, A. N. and Gobl, C. “Vocal Source Variation.” Included as pages 427-461 in *Handbook of Phonetic Sciences*, W. J. and Laver, J., Hardcastle, Oxford. 1997.

[5] Childers, D. G., and Lee, C. K. “Vocal Quality Factors: Analysis, Synthesis, and Perception.” JASA, Vol. 90, pp. 2394-2410, 1991.

[6] Deem, J. F., Manning, W. H., Knack, J. V., and Matesich, J. S. “The Automatic Extraction of Pitch Perturbation Using Microcomputers: Some Methodological Considerations.” JSHR, Vol.32, pp. 689-697, 1989.

[7] Deller, J.R. Jr. “On the Time Domain Properties of the Two-Pole Model for the Glottal Waveform and Implications for LPC.” Speech Communication, Vol. 2, pp. 57-63, 1983.

[8] Dennis, J. E. Jr, Woods, D. J. “New Computing Environments: Microcomputers in Large-Scale Computing.” Edited by Wouk, A. SIAM, pp. 116-122, 1987.

[9] Djeradi, A., Guerin, B., Badin, P., and Perrier, P. “Measurement of the Acoustic Transfer Function of the Vocal Tract: a Fast and Accurate Method.” Journal of Phonetics, Vol 19, pp. 387-395, 1991.

[10] Endo, Y. and Kasuya, H. “A Speech Analysis-Conversion-Synthesis System Taking Period-to-Period Fluctuations into Account.” Electronics and Communications in Japan, Part 3, Vol. 82, No. 12, 1999. (Translated)

[11] Epps, J., Smith, J. R., and Wolfe, J. “A Novel Instrument to Measure Acoustic Resonances of the Vocal Tract During Phonation.” Meas. Sci. Technol., Vol. 8 (10), pp. 1112-1121, October, 1997.

[12] Fant, G., Liljencrants, J, and Lin, Q. G. “A Four Parameter Model of Glottal Flow.” STL-QPSR 4, pp. 1-12, 1985.

[13] Fujimura, O. and Lindqvist, J. “Sweep-Tone Measurements of Vocal-Tract Characteristics.” JASA, Vol 49, No 2, pp. 541-558, 1971.

[14] Gabelman, B. and Alwan, A. “Analysis by Synthesis of FM Modulation and Aspiration Noise Components in Pathological Voices.” In ICASSP Conference Proceedings, pp. 449-452, Orlando, FL, May, 2002

[15] Gabelman, B. and Alwan, A. “Analysis and Synthesis of AM Components of Pathological Voices.” In IEEE Workshop on Speech Synthesis, Paper #20154. Santa Monica, CA., 9/11/2002. IEEE Catalog Number 02EX555. ISBN 0-7803-7396-0.

[16] Gabelman, B., Kreiman, J., Gerratt, B., Antonanzas-Barroso, N., and Alwan, A.

"LF Source Model Adequacy for Pathological Voices." Poster 5aSC17 presented at the 134th Meeting of the Acoustical Society of America, San Diego, CA., November, 1997.

[17] Gabelman, B., Kreiman, J., Gerratt, B., Antonanzas-Barroso, N. "Perceptually Motivated Modeling of Noise in Pathological Voices." Proceedings of the 16 International Congress on Acoustics and 135th Meeting of the Acoustical Society of America, pp. 1293, and unpublished Poster 2pSC30, Seattle, WA., June, 1998.

[18] Hillenbrand, J. "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.

[19] House, A. S. and Stevens, K. N. "Estimation of Formant Band Widths from Measurements of Transient Response of the Vocal Tract." Journal of Speech and Hearing Research, Vol. 1, No 4, pp. 309-315, Dec., 1958.

[20] IEEE Press. *Programs for Digital Signal Processing* (Section 8.1). John Wiley & Sons, 1979.

[21] Kent, R. D., and Read, C. *The Acoustic Analysis of Speech (*Chapt 7*)*. Singular Publishing Group, Inc., San Diego, CA., 1992.

[22] Klatt, D. H. and Klatt, L. C. "Voice Quality." Journal of the Acoustical Society of America, Vol. 87, No. 2, pp. 838, February, 1990.

[23] Kreiman, J., Gerratt, B.R., Precoda, K., and Berke, G. S. "Individual Differences in Voice Quality Perception." Journal of Speech and Hearing Research, Vol 35, pp. 512-520, April, 1995.

[24] Krom, Guus de. "A Cepstrum-Based Technique for Determining a Harmonics – to-Noise Ratio in Speech Signals." JSHR 93, Vol. 36, pp. 254-266, 1993.

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

[26] MATLAB for Windows. Version 4.2c. The Mathworks, Natick, MA. 01760. Copyright 1984-1994.

[27] Milenkovic, P. "Least Mean Square Measures of Voice Perturbation."

JSHR 30, pp. 529-538, 1987.

[28] Rabine, L. R., and Schafer, R. W . *Linear Prediction Coding of Speech in Digital Processing of Speech Signals* (Chapt. 8). Prentice – Hall, Englewood Cliffs, New Jersey, 1993.

[29] Tarnoczy, T. H. “Vowel Formant Bandwidths and Synthetic Vowels.” Letters to the editior, JASA, Vol. 34, pp. 859, 1962.

[30] Tarnoczy, T. H., “Uber Eigenfrequennz und Dekrement der Vokalresonatoren der menschlichen Stimme.” Arch. f Sprach und Stimmphysiol, 6, III/Iv 75-87, 1942.

[31] Qi, Y., and Bi, N. “A Simplified Approximation of the Four-parameter LF Model of the Voice Source.” JASA, Vol. 96, pp. 1182-1185, 1994.

[32] Sensyn 1.1 version of Klatt synthesizer. Sensimetrics, Cambridge, MA.

[33] Solari, E. *ISA & EISA Theory and Operation*. Annabooks, San Diego, CA., 1992.

[34] Yumoto, E., Gould, W. J., and Baer, T. “Harmonics to Noise Ratio as an Index of the Degree of Hoarseness.” JASA, Vol. 71, pp. 1544-1550, 1984.