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.